Dimensions mémorielles de l'interaction écrite humain-machine une approche cognitive par les modèles mnémoniques pour la détection et la correction des incohérences du système dans les dialogues orientés-tâche

Léon-Paul Schaub

université
PARIS-SACLAY

Dimensions mémorielles de l'interaction écrite humain-machine : une approche cognitive par les modèles mnémoniques pour la détection et la correction des incohérences du système dans les dialogues orientés-tâche

Memory dimensions of human-computer written interaction : a cognitive approach using mnemonic models for the detection and correction of system inconsistencies in task-oriented dialogues

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 : sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : **Informatique**
Graduate School : Informatique et Sciences du numérique, Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Laboratoire interdisciplinaire des sciences du numérique** (Université Paris-Saclay, CNRS), sous la direction de **Patrick PAROUBEK**, ingénieur de recherche HDR, et les co-encadrements de **Gil FRANCOPOULO**, docteur en informatique chez Akio ainsi que de **Samuel RUMEUR**, directeur produit chez Akio

Thèse soutenue à Paris-Saclay, le 22 mars 2022, par

Léon-Paul SCHAUB

Composition du jury

Frédéric LANDRAGIN Directeur de recherche, CNRS Laboratoire LATTICE, ENS Paris et Université Sorbonne Nouvelle	Président
Chloé CLAVEL Professeure, Télécom ParisTech	Rapportrice et examinatrice
Yves LEPAGE Professeur, Université de Waseda	Examinateur
Magalie OCHS Maîtresse de Conférence HDR, Université Aix-Marseille	Examinatrice
Frédéric BECHET Professeur des Universités, Université Aix-Marseille	Examinateur
Patrick PAROUBEK Ingénieur de Recherche HDR, CNRS Université Paris-Saclay	Directeur de thèse

THESE DE DOCTORAT

NNT : 2022UPASG023

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Dimensions mémorielles de l'interaction écrite humain-machine : une approche cognitive par les modèles mnémoniques pour la détection et la correction des incohérences du système dans les dialogues orientés-tâche

Mots-clés : Système de dialogue orienté tâche, modèle de mémoire, réseaux de neurones, détection d'incohérences

Résumé : Dans ce travail, nous nous intéressons à la place des systèmes de dialogue orientés-tâche à la fois dans le traitement automatique des langues, et dans l'interaction humain-machine. Nous nous concentrons plus particulièrement sur la différence de traitement de l'information et de l'utilisation de la mémoire, d'un tour de parole à l'autre, par l'humain et la machine, pendant une conversation écrite de type clavardage. Après avoir étudié les mécanismes de rétention et de rappel mémoriels chez l'humain durant un dialogue, en particulier dans l'accomplissement d'une tâche, nous émettons l'hypothèse qu'un des éléments susceptible d'expliquer que les performances des machines demeurent en deçà de celles des humains, est la capacité à posséder non seulement une image de l'utilisateur, mais également une image de soi, explicitement convoquée pendant les inférences liées à la poursuite du dialogue. Cela se traduit pour le système par les trois axes suivants. Tout d'abord, par l'anticipation, à un tour de parole donné, du tour suivant de l'utilisateur. Ensuite, par la détection d'une incohérence dans son propre énoncé, facilitée, comme nous le démontrons, par l'anticipation du tour suivant de l'utilisateur en tant qu'indice supplémentaire. Enfin, par la prévision du nombre de tours de paroles restants dans le dialogue afin d'avoir une meilleure vision de la progression du dialogue, en prenant en compte la potentielle présence d'une incohérence dans son propre énoncé, c'est que nous appelons le double modèle du système, qui représente à la fois l'utilisateur et l'image que le système renvoie à l'utilisateur. Pour mettre en place ces fonctionnalités, nous exploitons les réseaux de mémoire de bout-en-bout, un modèle de réseau de neurones récurrent qui possède la spécificité de créer des sauts de réflexion, permettant de filtrer l'information contenue à la fois dans l'énoncé de l'utilisateur et dans celui de l'historique de dialogue. De plus, ces trois sauts de réflexion servent de mécanisme d'attention "naturel" pour le réseau de mémoire, à la manière d'un décodeur de transformeur. Pour notre étude, nous améliorons, en y ajoutant nos trois fonctionnalités, un type de réseau de mémoire appelé WMM2Seq (réseau de mémoire de travail par séquence). Ce modèle s'inspire des modèles cognitifs de la mémoire, en présentant les concepts de mémoire épisodique, de mémoire sémantique et de mémoire de travail. Il obtient des résultats performants sur des tâches de génération de réponse de dialogue sur les corpus DSTC2 (humain-machine dans le domaine de restaurant) et MultiWOZ (multi-domaine créé avec Magicien d'Oz) ; ce sont les corpus que nous utilisons pour nos expériences. Les trois axes mentionnés précédemment apportent deux contributions principales à l'existant. En premier lieu, ceci complexifie l'intelligence du système de dialogue en le dotant d'un garde-fou (incohérences détectées). En second lieu, cela optimise à la fois le traitement des informations dans le dialogue (réponses plus précises ou plus riches) et la durée de celui-ci. Nous évaluons les performances de notre système avec premièrement la f-mesure pour les entités détectées à chaque tour de parole, deuxièmement le score BLEU pour la fluidité de l'énoncé du système et troisièmement le taux d'exactitude jointe pour la réussite du dialogue. Les résultats obtenus montrent l'intérêt d'orienter les recherches vers des modèles de gestion de la mémoire plus cognitifs afin de réduire l'écart de performance dans un dialogue entre l'humain et la machine.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title : Memory dimensions of human-computer written interaction : a cognitive approach using mnemonic models for the detection and correction of system inconsistencies in task-oriented dialogues

Keywords : Task-oriented dialogue system, memory model, neural networks, inconsistency detection

Abstract : In this work, we are interested in the place of task-oriented dialogue systems in both automatic language processing and human-machine interaction. In particular, we focus on the difference in information processing and memory use, from one turn to the next, by humans and machines, during a written chat conversation. After having studied the mechanisms of memory retention and recall in humans during a dialogue, in particular during the accomplishment of a task, we hypothesize that one of the elements that may explain why the performance of machines remains below that of humans, is the ability to possess not only an image of the user, but also an image of oneself, explicitly summoned during the inferences linked to the continuation of the dialogue. This translates into the following three axes for the system. First, by the anticipation, at a given turn of speech, of the next turn of the user. Secondly, by the detection of an inconsistency in one's own utterance, facilitated, as we demonstrate, by the anticipation of the user's next turn as an additional cue. Finally, by predicting the number of remaining turns in the dialogue in order to have a better vision of the dialogue progression, taking into account the potential presence of an incoherence in one's own utterance, this is what we call the dual model of the system, which represents both the user and the image that the system sends to the user. To implement these features, we exploit end-to-end memory networks, a recurrent neural network model that has the specificity not only to handle long dialogue histories (such as an RNN or an LSTM) but also to create reflection jumps, allowing to filter the information contained in both the user's utterance and the dialogue history. In addition, these three reflection jumps serve as a "natural" attention mechanism for the memory network, similar to a transformer decoder. For our study, we enhance a type of memory network called WMM2Seq (sequence-based working memory network) by adding our three features. This model is inspired by cognitive models of memory, presenting the concepts of episodic memory, semantic memory and working memory. It performs well on dialogue response generation tasks on the DSTC2 (human-machine in the restaurant domain) and MultiWOZ (multi-domain created with Wizard of Oz) corpora ; these are the corpora we use for our experiments. The three axes mentioned above bring two main contributions to the existing. Firstly, it adds complexity to the intelligence of the dialogue system by providing it with a safeguard (detected inconsistencies). Second, it optimizes both the processing of information in the dialogue (more accurate or richer answers) and the duration of the dialogue. We evaluate the performance of our system with firstly the F1 score for the entities detected in each speech turn, secondly the BLEU score for the fluency of the system utterance and thirdly the joint accuracy for the success of the dialogue. The results obtained show that it would be interesting to direct research towards more cognitive models of memory management in order to reduce the performance gap in a human-machine dialogue.

A mes parents et aux Asturies

Remerciements

Tout d'abord, je souhaite remercier tous les membres du jury pour leurs remarques, leurs questions et leurs conseils. En particulier, je voudrais remercier Monsieur Landragin et Madame Clavel pour ces rapports détaillés qui m'ont permis de préparer la soutenance dans les meilleures conditions.

Remontons un peu le temps : je désire remercier l'INALCO, six ans d'études ne s'oublient pas facilement. L'enseignement du japonais m'a ouvert les portes de la linguistique et de la didactique en licence, et le master TAL et toute son équipe pédagogique m'ont donné les armes nécessaires pour parvenir à mener à terme ce travail. J'ai une pensée particulière pour Damien, qui m'a initié à l'apprentissage automatique, pour Clément et son enseignement de la programmation ainsi que pour Jean-Michel et Mathieu, la gentillesse qu'ils m'ont montrée pendant deux ans est un souvenir indélébile.
Je souhaite aussi remercier Mathilde et Chloé, mes plus grands soutiens pendant ce master de TAL, sans qui je n'aurais pas fait cette thèse.
Je remercie aussi Akio et en particulier Patrick pour avoir parié sur moi et pour les quatre années incroyables passées dans l'entreprise. Dans ce contexte difficile, le fait d'apprécier aller au bureau au lieu de rester chez soi est d'une valeur inestimable. L'environnement chaleureux de Akio est rendu possible grâce à Carnetta, les Christines, Christophe, Marc, Gwen, Fabrice, anciennement Hien...
Lynda et Sam, je vous remercie particulièrement. Lynda, tu es une amie. J'ai pu compter sur toi dans les moments faibles comme j'ai pu rire avec toi pour rendre les moments forts. Je n'aurais pu espérer meilleure collègue. Sam, tu as été un guide, tu m'as encadré pendant plus de deux ans, avec la difficulté, j'en ai conscience, que représente la tâche. Pourtant, c'est grâce à toi que je suis parvenu à terminer ce travail. C'est toi aussi qui me donnais l'envie de travailler tous les jours.
Merci à Gil aussi, c'est toi qui m'as recruté, c'est toi qui m'as aidé, qui as été un soutien indéfectible pendant deux ans. Nous avons partagé nombre de conversations de TAL et de linguistique qui m'ont manquées lorsque tu es parti, sans toi le quotidien n'aurait pas été aussi intéressant.
Je n'oublie pas non plus mes collègues du LIMSI, les "anciens", Swen, Sanjay (merci Strava-man), Julien (Strava-man too), Arnaud, Zheng... Comme les "jeunes", Hicham (devenu ancien entre-temps), Corentin, Mathilde et Oralie. Vous rendiez mes rares visites au labo intellectuellement et gastronomiquement divertissantes. Je ne t'oublie pas non plus Bénédicte, toujours souriante, chaque matin où je suis venu, tu m'accueillais gentiment. Ni Blanche, pour les fois où je t'ai envoyé des ordres de mission la veille de celle-ci.
Corentin, encore, merci pour toutes les fois où, à 23h, je t'envoyais des messages commençant par "Tu as cinq minutes ? J'ai un problème sur une fonction de loss", auxquels non seulement tu répondais favorablement mais en plus tu savais que ces cinq minutes se transformeraient en heures. Merci pour ta gentillesse et ta sympathie.
Merci à Jennifer pour un quotidien plus facile, toujours à l'écoute, toujours prête à m'aider, tu as été un grand soutien notamment pendant les confinements.
Je remercie spécialement Timothée, Kévin, Aurélien et PL, merci les gars, sans vous ces quatre années auraient paru bien longues. Je vais enfin pouvoir vous expliquer de quoi parle ma thèse.
Je veux bien sûr remercier mes parents, mon frère et Félix, merci d'être dans ma vie.
Merci à tous ceux que j'oublie mais qui m'ont aidé d'une façon ou d'une autre à accomplir ce travail.
Nadège, tu as été avec moi depuis le début, amie, compagne, coloc et confidente, tu as tout assumé, tu ne m'as

jamais lâché, malgré les moments difficiles par le contexte, la distance et la thèse. Pour moi, tu faisais toutes les semaines l'intégralité de la ligne 13, et cela n'a pas de prix. Merci d'avoir été là, à mes côtés et dans mes pensées.

Enfin, je ne peux finir ces remerciements sans t'évoquer, Patrick. Je n'ai pas les mots pour décrire tout ce que tu as fait pour moi pendant cinq ans. Je n'oublie pas que c'est avec toi que tout a commencé, en 2016, après un de tes fameux cours sur le Lambda Calcul, où je t'ai demandé, timidement, si tu voulais bien accepter d'être mon directeur de mémoire de Master. Tu m'as emmené au LIMSI, et m'as montré comment était la vie de labo. C'est toi qui m'as encouragé à faire une thèse. Dès le moment où l'opportunité est apparue, je savais que ça ne pouvait être que toi le directeur. Merci pour toutes ces discussions, ces repas, débats, des hauts, des programmes avec plus de bugs que de lignes, avec des articles avec plus de références que de contenu... Merci pour m'avoir orienté vers le travail qui aboutit maintenant. Merci pour ta patience, pour ton côté zen, pour m'avoir fait relativiser, m'avoir rassuré quand il le fallait, m'avoir motivé quand il le fallait. Tu as été le meilleur directeur de thèse possible.

Table des matières

Chapitre 1

Introduction

> Nobody cares about how it works as long as it works
>
> Hamann

L'étude de la langue naturelle par le linguiste repose sur la formalisation des phénomènes linguistiques diachroniques et synchroniques , communs ou uniques à chaque langue, grâce à des approches génératives [Chomsky, 1965] ou fonctionnelles [Creissels, 1995, Stolz, 2010]. A l'inverse, l'étude de la langue par le didacticien se définit par l'apprentissage, l'assimilation et l'enseignement d'une langue cible à partir, d'une part, d'une maîtrise écrite et orale de celle-ci (souvent la langue maternelle du didacticien) et d'autre part d'une connaissance partielle d'une langue source (parfois seulement la langue maternelle de l'apprenant, car la tendance actuelle est de considérer que la langue source de tout un chacun est l'anglais) où le passage de l'une à l'autre s'effectue à travers différentes méthodes pédagogiques dont les applications varient selon l'âge, les objectifs et le nombre des apprenants [Rubin, 1975, Blanchet, 2015]. Notons que pour le didacticien, il s'agit toujours d'un couple composé d'une langue cible (aussi appelée L2) et d'une langue source, mais rien ne l'empêche d'être capable d'enseigner à partir de plusieurs langues sources ou dans plusieurs langues cibles.
Nous précisons que nous parlons ici de langue naturelle[1] et non pas de langage naturel[2]. C'est pourquoi l'on a pris l'habitude de parler de langage de programmation et non pas de langue de programmation, le choix des mots se révèle rarement innocent. Cette distinction, sans être propre au français, ne semble pas universelle selon les cultures. Par exemple, alors que l'anglais utilise l'expression *natural*

1. un langage vocal ou graphique, propre à des humains appartenant à un même groupe social

2. toute forme d'expression ou de communication (dont la langue) entre des êtres, vivants ou mécaniques

language pour désigner à la fois la langue et le langage naturel, ce qui amène à confondre les deux concepts, certes cousins mais néanmoins différents, le français permet d'exprimer cette différence, car dans la mesure où le langage représente un moyen de communication -sans besoin d'être graphique ou vocal- entre des êtres -qui ne sont pas nécessairement des humains appartenant à un même groupe social- il dépasse le cadre qui concerne notre travail, à savoir celui du **traitement automatique des langues**. Une dernière précision semble être bienvenue, nous utilisons pour notre travail un raccourci fréquent : nous considérons la Langue (et non pas le langage) comme un système abstrait sous-jacent à tout acte de parole de toute langue du monde[3].

Toutefois, d'après ces définitions, le linguiste serait alors susceptible de connaître la plupart des phénomènes de la Langue sans comprendre comment l'on traduit *week-end* en anglais. De son côté, le didacticien pourrait maîtriser plusieurs dizaines de langues sans pour autant être capable d'expliquer de façon satisfaisante la différence entre l'accusatif et le nominatif. Autrement dit, une approche linguistique se voudrait agnostique d'une ou de plusieurs langues en particulier car elle posséderait comme objectif une définition théorique de la langue naturelle, alors qu'une approche didactique serait rattachée à au moins deux langues spécifiques -langue source et langue cible- dans la mesure où son but serait d'étudier la meilleure façon de transférer les compétences langagières humaines de la première à la seconde, sans pour autant avoir besoin de posséder la moindre connaissance théorique sur le fonctionnement de l'une ou l'autre.

Cependant, nous pouvons considérer que pour qu'une approche linguistique aboutisse à la formalisation d'un certain nombre de phénomènes communs ou propres à une langue, cette dernière devrait utiliser des exemples concrets d'une langue A puis d'une langue B afin d'illustrer les règles -et les exceptions les confirmant- qu'elle cherche à établir, ce qui implique que le linguiste possède une compétence langagière minimale dans les deux langues (ou un ami bilingue systématiquement à portée de main)[4]. En d'autres termes, l'étude de la langue naturelle avec une approche linguistique utilise une certaine méthodologie propre à la didactique. En outre, pour rendre l'apprentissage d'une langue cible efficace, le didacticien tend à rendre abstraits et formels des phénomènes concrets dans une langue cible afin de fournir un ancrage théorique familier aux apprenants qui facilite l'assimilation [Halliday et al., 1964]. Lors de l'apprentissage d'une L2, il a été démontré que le retour à des concepts abstraits permettant de comprendre le fonctionnement de telle conjugaison dans la langue cible par rapport à la langue source accélère la production orale, que ce soit à un niveau débutant ou avancé [Liu et al., 2016b].

3. d'après F. de Saussure *Cours de Linguistique générale* page 25
4. DeepL, à défaut d'ami

Nous pouvons alors supposer que l'étude de la langue naturelle repose à la fois sur la formalisation de phénomènes communs ou uniques à certaines langues et la compétence langagière permettant d'illustrer ces phénomènes par des exemples réels.

Si nous considérons désormais le traitement automatique des langues -naturelles- (TALN ou TAL) comme l'utilisation de l'informatique pour l'étude de la langue naturelle, se traduisant à la fois par une modélisation mathématique et une réalisation logicielle de celle-ci, nous pouvons alors supposer que le TAL exploite à la fois les théories linguistiques et les méthodes didactiques. En effet, le taliste (ou info-linguiste) se doit de définir certains aspects de la langue naturelle grâce à des formalismes mathématiques. Il veille également au développement d'outils informatiques facilitant l'analyse de la langue naturelle, par exemple lors d'une étude de documents de très grande taille, qui serait trop longue à effectuer manuellement.

Partant de l'hypothèse que l'analyse de la langue naturelle repose sur les cinq grandes aires linguistiques que sont **la phonologie, la morphologie, la syntaxe, la sémantique et la pragmatique** [Gleason, 2005], nous trouvons des ambiguïtés ou des difficultés à tous ces niveaux, qui semblent à priori uniquement solubles par l'humain. Ces difficultés étant à résoudre tant pour le linguiste (formalisation de ces axes et de leurs difficultés à travers des exemples concrets) que pour le didacticien (appropriation de la nature de la difficulté dans la langue cible et la langue source à travers la formalisation), on en conclut qu'elles sont à résoudre également pour le taliste, avec la difficulté ajoutée de devoir les formaliser par des équations et des algorithmes[5]. Nous illustrons ces axes avec des exemples tirés de deux langues flexionnelles (le français et l'espagnol) ainsi que d'une langue agglutinante (japonais) pour montrer que les difficultés ne sont pas propres à une famille particulière de langue.

1. **La phonologie** est l'étude des sons de la langue, et de la façon dont ils sont interprétés pour devenir informatifs. La difficulté à ce niveau d'analyse réside dans le fait que le même son s'écrit de différentes façons et inversement.

   (a) ***cent-sang-sent-sans-s'en***... se prononcent exactement de la même façon */sã/*. Alors que *p**ub**(publicité, /pb/)* et p***ub**(bar, /pyb/)* qui pourtant s'écrivent de façon identique, ne se prononcent pas pareil.

   (b) En espagnol : ***C**orcho (Le 'C' se prononce /k/) **C**iudad (le 'C' se prononce /θ/)*[6].

   (c) En japonais, 人[7] peut se prononcer

5. Certains des exemples ci-dessous proviennent de Loic Suberville :https://www.youtube.com/channel/UCywGsTdh_qqZUYmA2Gro2CA/featured

6. bouchon de liège/Ville

7. l'individu

— */çito/* comme dans この人[8],
— */bito/* comme dans 恋人[9],
— */ri/* comme dans 一人[10],
— */ɲ/* dans 人間[11]
— ou encore */ɕĩũɳ/* dans 人生[12]

2. **La morphologie** est l'étude de la forme minimale sémantique autonome de la langue : le mot. Il se décompose en morphèmes (plus petite unité de sens), en lemmes (association générique de morphèmes qui définissent la racine grammaticale d'un mot) et en paradigmes (toutes les formes possibles de dérivation d'un lemme, par exemple une conjugaison ou une déclinaison). La difficulté à ce niveau-ci réside dans le fait qu'une même lettre peut posséder plusieurs fonctions.

   (a) En français comme en espagnol, le 's' peut servir pour décrire le pluriel des substantifs, comme il peut servir de terminaison pour la première personne du singulier des verbes.

   (b) En Japonais, le concept de lettre n'existant pas vraiment, il est remplacé par celui de syllabe. La syllabe ない (nai) peut tantôt signifier la négation, pour les verbes et les adjectifs verbaux (食べる・食べない[13]) (taberu/tabenai) que la terminaison pour des adjectifs (危ない・汚い[14]) (abounai/kitanai)

3. **La syntaxe** est l'étude du syntagme, c'est-à-dire un ensemble de mots ordonnés avec une hiérarchie dictée par les règles grammaticales de relations entre elles. La difficulté ici réside dans l'ambiguïté des dépendances entre les mots.

   (a) En français comme en espagnol, les langues étant dépourvues de particules (comme en japonais) ou de déclinaisons (comme en allemand) pour déterminer la fonction de chaque mot, c'est l'ordre de ceux-ci dans la phrase qui le détermine.
   Jean regarde Marie avec ses jumelles. Dans cette phrase, est-ce *Jean* qui regarde Marie grâce à ses jumelles, ou est-ce que Jean regarde Marie qui elle-même possède des jumelles.

   (b) Persigo al ladrón en bici[15] Est-ce moi qui suis en vélo, ou le voleur, ou les deux ?

8. cette personne
9. l'être aimé
10. une personne
11. être humain
12. vie humaine
13. je mange/ je ne mange pas
14. dangereux/sale
15. Je poursuis le voleur en vélo

(c) この女の子 (kono on'na no ko). Peut à la fois signifier*"cette jeune fille"* et *"la fille de cette femme"*

(d) 頭が赤い魚を食べる猫 (atama ga akai sakana wo taberu neko) Peut signifier jusqu'à quatre choses à la fois. Il est composé des mots : tête, rouge, poisson, manger chat [16]. Les particules が et を indiquent respectivement le sujet et l'objet direct du verbe.
La traduction naturelle serait :
(1)*"Le chat qui mange le poisson à tête rouge."*
Cependant, il peut aussi signifier :
(2) *"Le chat à tête rouge qui mange le poisson"*
(3) *"Le chat dont la tête mange le poisson rouge"*
(4) *"J'ai une tête de chat qui mange un poisson rouge"*
(5) *"J'ai une tête rouge de chat qui mange un poisson"*
Si l'humain peut hésiter entre cinq sens possibles pour une phrase entière, comment imaginer qu'une machine puisse faire mieux ?

4. **La sémantique** est l'étude du sens des mots. Elle a pour but de résoudre le problème posé par la syntaxe. La difficulté à ce niveau réside dans le fait que le même mot peut signifier des choses différentes au sein d'une même langue, ou bien des choses différentes entre plusieurs langues.

   (a) Jean regarde Marie avec ses jumelles. Dans cette phrase, les jumelles, s'agit-il de l'appareil pour mieux observer à longue distance, ou bien des filles de Marie ?

   (b) En français, *"vache"* peut signifier l'animal que tout le monde connaît, mais également *"quelqu'un de vache"* peut être synonyme de quelqu'un de méchant, "*la vache*" peut tout aussi bien être une expression péjorative envers le physique d'une personne, qu'un énoncé interjective synonyme de "Sapristi". Elle a même droit à son adverbe dérivé *"vachement"* qu'on peut traduire par "extrêmement".

   (c) En espagnol, il existe deux mots pour parler du poisson : *pez* [17] et *pescado* [18]. Une phrase comme *El pez pescado.* [19] peut prêter à confusion. A l'inverse, le mot *esposas* peut signifier à la fois épouses et menottes. Enfin, le mot *ternera*, qui, seul, signifie "veau", possède un autre sens utilisé en boucherie. En effet, si l'on dit *filete de ternera*, on parle d'un steak de veau. En revanche, *solomillo de ternera* signifie "faux-filet". Or, le problème intervient lorsque par exemple un client arrive et

16. Exemple tiré de https://www.reddit.com/r/linguisticshumor/comments/hhfs91/syntactic_ambiguity_in_japanese/
17. poisson vivant dans l'eau
18. poisson dans notre assiette
19. le poisson pêché.

réprimande le boucher ainsi :"*No me ha gustado la ternera !*"[20]. A ce moment, comment peut-on déterminer s'il parle de veau ou de bœuf ?

(d) En Japonais, les confusions sémantiques sont souvent désambiguïsées par l'utilisation de sinogrammes appelés *kanjis* et porteurs de sens. 生姜ない。しょうがない。[21] *(shyouga nai. shyou ga nai.)* sont deux phrases qui se prononcent exactement de la même façon, pourtant elles ne signifient pas du tout la même chose. 一日 . 一日[22] (tsuitatchi/itchi-nitchi) ces deux mots s'écrivent de la même façon et pourtant ils se prononcent et se définissent différemment.
Ces ambiguïtés ne peuvent en général pas être levées sans le contexte dans lequel elles apparaissent.

5. **La pragmatique** est justement l'étude des phrases dans leur contexte. La plus grande difficulté ici est de déduire le sens des mots utilisés dans un contexte qui induit un nouveau sens, lié au reste de ce contexte.

   (a) En français, les locuteurs ont tendance à utiliser des litotes ou des ellipses dans leurs phrases. " C'est *un peu* piquant". *un peu* remplace ici "trop". Ou bien "- Veux-tu que j'ajoute du sel ? - C'est *très bien*" . Ici, *très bien* sous-entend "non merci".

   (b) En espagnol, on trouve le même genre de phénomènes comme dans "- Quieres café ? - Lleva cafeina..." [23] qui sous-entend "Non car" dans la deuxième proposition.

   (c) En japonais, les ellipses sont d'autant plus fréquentes que les actants du verbe sont souvent sous-entendus dans le verbe lui-même. Par exemple 嘘をつかさせられる[24] intraduisible en français mais qui peut se rapprocher de " On m'oblige à faire dire à quelqu'un un mensonge à quelqu'un d'autre ". On trouve aussi d'autres exemples d'ellipse comme dans 今日、天気はどう？傘を持って行った方がいいと思う[25], la deuxième phrase sous-entendant qu'il se pourrait bien qu'il se mette à pleuvoir dans la journée.

Toutes ces difficultés présentes dans une langue peuvent être surmontées par les humains la parlant car ils ont mémorisé, tout au long de leur vie, des contextes dans lesquels tel ou tel sens de tel ou tel mot est utilisé. Cette compétence est appelée sociolinguistique [Labov, 1973]. C'est justement cette compétence qui explique également les limites des analyseurs automatiques, que ce soient des analyseurs de surface (au niveau lexical) ou de profondeur (au niveau syntaxique) car ils ne possèdent pas cette connaissance sociale et cette expérience de la langue. Un descriptif

20. Je n'ai pas aimé la viande de veau/bœuf
21. Il n'y a pas de gingembre ? Tant pis
22. 1er jour du mois. la journée
23. -Tu veux du café ? - Il y'a de la caféine...
24. litt. être faire dire un mensonge
25. Quel temps fait-il aujourd'hui ? Je te conseille de prendre un parapluie.

de certains de ces outils et de leurs limites [Sadoun et al., 2016] est disponible en ligne. [26].

A la différence du linguiste-didacticien qui formalise la langue par le prisme du micro-phénomène linguistique et des situations socio-historiques faisant évoluer la langue, le taliste travaille en binôme avec la machine, voire, de nos jours, lui confie une grande partie du travail. En effet, par définition, le taliste possède une compétence de plus que tout linguiste-didacticien, à savoir la connaissance d'au moins un langage informatique. En outre, le langage informatique, à la différence des langues naturelles, est un langage entièrement formel, fini, sans exceptions, avec une syntaxe exhaustive et des patrons définis, a priori un langage parfait.

L'enjeu alors du taliste est de formaliser la langue naturelle de façon à ce que la machine puisse l'interpréter.

Dans le cas de la langue naturelle, les humains possèdent une capacité d'interprétation de ce qu'ils entendent ou lisent liée à leur compétence langagière y compris lorsque les mots échangés leur sont inconnus [Swanborn and de Glopper, 1999]. Ainsi, la communication est possible entre deux humains malgré la présence de néologismes, de mots désuets, d'argot, de fautes d'orthographe... Les fautes d'orthographe sont un exemple de la capacité des humains à interpréter une langue naturelle. Dans une langue comme le français, où lire un texte ne contenant pas de faute est un événement, les erreurs d'orthographe ne sont pas un frein à la production et à la compréhension de la langue, car le récepteur du message se met à la place de l'émetteur, et comprend ce que ce dernier a voulu dire. Par exemple, en français, parmi des fautes courantes, nous trouvons :
(1) "Salut, sa va ?",
(2) "tu a fais quoi ce week-end ?"
(3) "je pas trouver le sorti."
(4) "Il a parti dans deux apres heure"

Le récepteur comprend le message de l'auteur. C'est le modèle du signifiant-signifié de Ferdinand de Saussure qui explique que lors d'une production de mots et de phrases, l'émetteur se met à la place du récepteur pour anticiper ce que celui-ci va interpréter [Bailly et al., 2008]. De la même façon, le récepteur possède un modèle de l'émetteur, c'est-à-dire qu'en même temps qu'il écoute et interprète la production de l'émetteur, il imagine ce que l'émetteur a voulu dire, d'une certaine façon il se met à la place de l'émetteur. D'autre part, nous pourrions affirmer que les phrases écrites ci-dessus sont illisibles ou incompréhensibles, mais cela est faux. Toutes, même la (4), sont compréhensibles par un locuteur francophone. A l'inverse, une phrase comme :

26. http ://multital.inalco.fr/project

(5) "food food yes in restaurant food north is is north"[27]

n'est pas compréhensible pour un humain.

Dans le cas du langage de programmation, l'interlocuteur est un ordinateur. Or, par définition, un ordinateur ne se met pas à la place de l'humain (ce serait de la science-fiction). Le locuteur, appelé alors programmeur, communique avec la machine à travers un langage informatique en écrivant des programmes. Ces programmes contiennent une suite d'instructions destinées à être suivies par la machine. Si le locuteur commet une faute dans son programme (par exemple une instruction non reconnue par la grammaire du langage), l'ordinateur ne va tout simplement pas être capable de l'interpréter et lui renverra un message aussi sympathique que "ERREUR FATALE", "COMMANDE NON TROUVÉE" ou "SYNTAXE INCORRECTE". Il n'existe pas de néologisme, de désuétude, d'argot, ou de faute d'orthographe pour les langages de programmation. Nous trouvons certes des mises à jour et des versions obsolètes, mais elles sont codifiées, formalisées, unifiées ; la machine ne comprend pas nos programmes, elle les lit, puis les exécute.

La question que l'on peut alors se poser c'est dans quelle mesure peut-on écrire des programmes qui permettront à la machine d'interpréter les langues naturelles ? Comment faire tout simplement de la linguistique informatique ?

Par exemple, dans le domaine de la traduction, le traducteur développe des techniques de traduction fondées sur son expérience et sa connaissance de la langue cible pour traduire un texte. Il peut également se servir de la traduction assistée par ordinateur (TAO) pour automatiser certains traitements, notamment dans le cas de la traduction de textes officiels[28]. Le taliste-traducteur fabrique une machine de traduction (MT), qui va apprendre à traduire des textes d'une langue source vers une langue cible par le biais de corpus parallèles (des textes identiques traduits dans deux langues différentes et alignés de façon à être lus par une machine), préalablement traduits. La machine crée des règles formelles ou statistiques, basées sur les observations des corpus parallèles, puis évalue ses règles sur un autre corpus parallèle afin de vérifier son efficacité. Le site de traduction automatique DeepL est basé sur ce principe grâce aux corpus parallèles disponibles sur Linguee [Rescigno et al., 2020].

Dans le domaine de l'analyse de sentiments, le principe est similaire : le taliste observe des milliers de textes puis les annote, parfois assisté par la machine qui apprend des schémas d'annotation au fur et à mesure que le taliste annote les

27. exemple tiré du corpus DSTC2 [Henderson et al., 2014a]

28. Depuis quelques années, https://ec.europa.eu/inea/sites/default/files/cefpub/2020-1_automated_translation_faqs_batch4_final.pdf cela est devenu obligatoire (E-translation) au niveau européen

textes, c'est de l'annotation semi-automatique [Grouin et al., 2014]. Cependant, c'est la machine qui, une fois ce travail d'annotation fini, lit les textes annotés et apprend un modèle de sentiments à partir du contenu textuel et de l'annotation correspondante. Ce modèle appris lui sert ensuite à prédire le sentiment dans un texte encore non observé. Autrement dit, le taliste doit non seulement se mettre à la place des auteurs de tous les textes pour inférer les sentiments présents dans chacun d'eux, mais aussi à la place de la machine qui va les analyser. Le taliste sert en quelque sorte de relais indirect entre l'humain et la machine. On peut considérer que le taliste *dialogue* avec la machine, à travers les langages de programmation, pour lui permettre d'apprendre à analyser la langue. En effet, pour analyser le sentiment dans un document, écrit ou parlé, une méthode utilisée par l'analyste est de se mettre à la place de l'auteur du document [Kuckartz, 2014]. Il engage donc une sorte de *dialogue* virtuel où il tente d'interpréter le sentiment de l'auteur. En somme, il engage un *dialogue* à la fois avec la machine à travers ses compétences informatiques, et avec lui-même en analysant des documents.

Que peut-on alors dire d'un taliste qui crée un système de dialogue, dont le but est de converser avec un humain ? Il incombe au taliste :

— d'analyser des documents contenant des dialogues.
— d'apprendre à la machine ce qu'est le dialogue.

Pour répondre à cette question, nous pouvons étudier l'étymologie des mots *conversation* et *dialogue*. *Le Dialogue* vient du grec et signifie la raison (logos) qui traverse, qui pénètre (dia). *La conversation* vient du latin et a pour signification l'association, la fréquentation. Nous pouvons interpréter le dialogue comme étant l'utilisation de raison de la part de chaque participant, soit pour produire, soit pour comprendre (aspect linguistique), et la conversation comme le lien social, concret qu'entretiennent les deux participants (aspect didactique). Lorsque nous dialoguons avec quelqu'un, nous apprenons tout autant que nous enseignons. Ce sont ces deux éléments, que l'on peut également appeler études, qui nous permettent de développer notre compétence sociolinguistique.
Autrement dit, chaque personne mémorise et améliore tout au long de sa vie le *mode d'emploi* conversationnel et dialogique. En effet, en fonction du contexte, de la personne qu'elle a en face, du sujet de la conversation, de son état de fatigue du message qu'elle veut transmettre... chaque personne dialogue de manière singulière.
Intuitivement, nous supposons que cette manière est enregistrée dans la mémoire de tout un chacun. En fait, ceci est faux, ce ne sont pas une mais plusieurs mémoires qui, chacune interagissant avec les autres, permettent le bon déroulement de la conversation que nous désirons engager, en enregistrant et restituant les informations que s'échangent les participants.
Cependant, comment peut-on faire le lien entre l'analyse de la langue et l'utilisation de nos mémoires ? Par exemple, si nous parlons à un ami et qu'il nous dit :
Je suis allé au cinéma hier.

Notre mémoire dite à très court terme enregistre les mots les plus saillants, à savoir une action *(allé)*, un lieu (*cinéma*) et un temps (*hier*). Elle les transforme en informations. Ces informations sont traitées par la mémoire dite de travail qui effectue le lien entre ces informations et les mémoires à long terme [D'Esposito, 2007] :

— d'abord par le filtre le plus général de la mémoire sémantique
-> *Et qu'as-tu vu ?*
Nous savons que notre ami a vu un film, à moins que nous ne possédions un ami qui n'irait au cinéma que pour acheter des pop-corns. Au cinéma, on regarde des films.

— puis par le filtre plus contextuel de la mémoire épisodique ou autobiographique.
-> *Tu as vu le dernier Tarantino ?*
Nous savons que notre ami apprécie les films de Quentin Tarantino. Il est probable que ce soit cela qu'il ait vu au cinéma.

En enregistrant puis en restituant des informations contextuelles et générales ancrées dans nos mémoires pour faire avancer la conversation, nous nous mettons d'une certaine manière à la place de notre ami.
Toutefois, dans le domaine du dialogue, une simple restitution d'informations liées à celles préalablement enregistrées ne suffit pas. En effet, l'interaction étant au cœur du domaine, il s'agit pour chaque participant d'un dialogue de non seulement comprendre ce que l'autre énonce mais également d'anticiper ce que l'autre comprendra lors de l'énoncé suivant [Mandel et al., 2016]. Autrement dit, chaque participant utilise sa mémoire non seulement pour écouter et parler mais aussi pour se représenter, pendant qu'il parle, une image de ce que l'autre interprète et donc ce qu'il pourra répondre.

La question qui se pose ici est, lors d'un dialogue entre deux personnes dans quelle mesure chaque participant parvient-il à se mettre à la place de l'autre, et quels sont les mécanismes sollicités pour la mise en œuvre de ce phénomène ?

Dans nos travaux, nous nous intéressons à un sous-domaine du dialogue que l'on appelle dans la littérature le dialogue orienté-tâche. Également appelé dialogue orienté-but, ce type de dialogue se différencie du dialogue dit conversationnel en cela qu'il possède comme son nom le laisse deviner, un but, une tâche, une fonction... Cette notion reste assez vague car cela implique que certains dialogues peuvent ne pas posséder de but. Si l'on se réfère à la définition donnée du dialogue, à savoir la raison qui traverse, on s'aperçoit qu'il existe un paradoxe. En effet, si la raison (par le biais de la parole et donc de la langue) traverse, ici chacun des participants du dialogue, cela signifie qu'elle possède une direction, et un sens, autrement dit, elle peut être représentée par un vecteur. Or la définition de la raison implique l'utilisation d'une pensée par celui la possédant, donc un vecteur de raison est pourvu d'une pensée, et est chargé, dans le cas du dialogue, de transporter cette pensée depuis son origine, celui qui l'exprime, à sa destination,

celui qui l'écoute. Ce transport n'est-il pas en soi une fonction, un but, une tâche ? En ce cas, quelle est la frontière entre dialogue à but et sans but ? Parler de la météo ou de la dernière mesure gouvernementale contre un certain virus a-t-il un but, une fonction ? Si l'on prend une définition formellement acceptée de ce qu'est une tâche, à savoir un *travail défini et limité, imposé par autrui ou par soi-même, à exécuter dans certaines conditions.* [29], on s'aperçoit de l'ambiguïté des termes "travail" "imposé" et "conditions". En effet, un travail imposé sous conditions s'appelle une corvée. Or à notre connaissance, il n'existe pas de dialogue orienté-corvée. Il faut interpréter ici "travail" comme "but", "imposé" comme "décrit" et "conditions" comme "contexte". Un dialogue orienté-tâche est donc un dialogue dont le but est clairement explicité par le contexte de sa réalisation. On peut alors considérer qu'une conversation de type quotidienne n'entre pas dans cette catégorie car son but n'est pas clairement explicité par son contexte mais par le dialogue lui-même. A l'inverse, un dialogue comme celui d'un client avec un vendeur, ou d'un patient avec un médecin sous-entend un contexte (des conditions dialogiques) précédant le début du dialogue et **orientant** celui-ci. Il existe une tâche sous-jacente, présente avant même le début du dialogue qui oriente ce dernier. Cela n'empêche pas la présence au sein d'un même dialogue de sous-dialogues faisant partie de la catégorie du dialogue conversationnel [Larsson, 2017].

Toutefois, c'est la relation entre la tâche du dialogue et le dialogue comme domaine linguistique qui nous intéresse ici. Prenons comme exemple le domaine de la relation-client.
Pour l'utilisateur comme pour l'employé, la tâche est identique avant même le début de la conversation : résoudre le problème de l'utilisateur. On parle alors de tâche co-construite [Bertin and Masson, 2020]. Nous illustrons la relation entre tâche et dialogue à partir d'une conversation de clavardage sur un site de vente de parfum en ligne entre un utilisateur et un employé [30]. La conversation a été anonymisée (le vrai nom d'un individu devient `_Name1_`) . Les tours de parole sont numérotés de 1 à n. Chaque tour de parole contient un énoncé de l'utilisateur et un de l'employé. Si pendant un tour de parole l'un ou l'autre ne s'est pas exprimé, son énoncé est matérialisé par <**SILENCE**>. Dans ce dialogue, après une introduction de la part de l'employé :

```
(bonjour, je suis _Name2_. en quoi puis je vous aider ? afin d'
    accélérer le traitement de votre dossier, merci de nous
    indiquer votre adresse email et / ou votre numéro de
    commande.)
```

L'utilisateur expose ses intentions - à savoir sa contrariété quant à l'utilisation d'un code de promotion - à travers un acte de dialogue que [Landragin, 2008]

29. source : CNRTL
30. tous les exemples en français de ce chapitre proviennent de conversations de clavardage réelles

```
1 Utilisateur : <SILENCE>
1 Employé : bonjour, je suis _Name2_. en quoi puis je vous aider
    ? afin d'accélérer le traitement de votre dossier, merci de
    nous indiquer votre adresse email et / ou votre numéro de
   commande.
2 Utilisateur : bonjour, mon code promo marche pas jdf2019
2 Employé : ce code était valable jusqu'au 12/03
3 Utilisateur : ah tant pis alors, de toute façon les prix ne
   sont pas très intéressant quand on ajoute la livraison
3 Employé : avez vous d'autres questions ?
4 Utilisateur oui
4 Employé : je vous écoute
5 Utilisateur : vous êtes un robot ? genre vous avez de la
   chance de pas être allé à l'Ecole pour apprendre, c'est dur
   et chiant
5 Employé : non je ne suis pas un robot
6 Utilisateur : <SILENCE>
6 Employé : quelle est votre question ?
```

TABLE 1.1 – dialogue 1

définit comme "*l'unité minimale d'échange entre la machine et son utilisateur. C'est l'unité sur laquelle se base le gestionnaire de dialogue pour comprendre l'action de l'utilisateur et décider du comportement à adopter, de l'action à exécuter, de l'information à fournir et du message à envoyer au générateur de langue naturelle*"

bonjour, ***(acte=)mon code promo marche pas.*** jdf2019

et offre des informations à travers des entités utiles pour l'employé.

bonjour, mon code promo marche pas.***(entité=)jdf2019***

La réponse du vendeur

ce code était valable jusqu'au 12/03

sous-entend que le code en question n'est plus valable. Ici, le vendeur sait qu'il risque de créer de l'insatisfaction chez l'utilisateur dans la mesure où il répond à sa contrariété par une contrariété encore plus importante. Sachant que l'employé est la face visible de l'entreprise, il répond en son nom. Souvent, dans une telle situation, un employé a tendance à présenter ses excuses et offrir une alternative de façon à limiter l'impact de la réponse négative. Il n'en est rien dans ce cas-ci. Cela pourrait avoir des conséquences pour l'utilisateur.

ah tant pis alors

Ici il montre comprendre la réponse du vendeur et exprime sa déception. Mais il ne s'arrête pas là et continue :

`de toute façon les prix ne sont pas très intéressant quand on ajoute la livraison`

Exprimant sa frustration et en émettant une opinion de valeur sur l'entreprise, que l'on pourrait interpréter comme de la provocation. Loin de s'en émouvoir, l'employé lui demande :

`avez vous une autre question ?`

Ici, l'employé a ignoré la frustration de l'utilisateur et refuse de s'écarter de son but initial, à savoir répondre au besoin du client. Ce faisant, il refuse de réagir à la provocation et donc ne démarre pas un nouveau sous-dialogue avec comme tâche hypothétique la résolution de la frustration de l'utilisateur. Sauf qu'en ignorant le sentiment de l'utilisateur, il génère le doute dans la tête de ce dernier, qui se demande qui il peut bien avoir en face de son écran de clavardage. Le client demande alors :

`vous êtes un robot ?`

Cette phrase exprime une question que tout un chacun serait susceptible de se poser s'il conversait en ligne avec une personne qu'il ne connaissait, ne voyait et n'entendait pas. Le doute apparaîtrait, mêlé à la frustration accumulée et à la non-réaction de l'interlocuteur. C'est ce qui arrive ici.
Ce à quoi l'employé rétorque :

`non je ne suis pas un robot`

Nous nous trouvons ici dans un nouveau dialogue, un dialogue pour déterminer l'humanité d'un des deux interlocuteurs. Qu'il est loin le code promo du début de la conversation ! Ce nouveau dialogue, dont le but n'a pas été créé a priori, mais pendant la conversation elle-même n'est donc pas à proprement parlé une tâche. Il fait néanmoins partie de la conversation, et peut être considéré comme une sous-tâche nécessaire à l'accomplissement de la tâche principale. La tâche principale n'est dont plus uniquement de résoudre cette histoire de code promo, mais également de lever le doute sur l'humanité de l'employé.
Par conséquent, est-il encore juste d'affirmer que c'est la tâche qui oriente le dialogue ? Ne serait-ce pas l'inverse ? Cet exemple nous invite à penser que l'influence est mutuelle, la tâche influence - oriente - l'échange de paroles qui lui-même altère, déforme la tâche initiale pour créer une nouvelle tâche, co-construite, partagée, faite de raison et qui traverse les participants par l'alternance des tours de parole, autrement dit un **dialogue**. Deux questions émergent alors :

1. Dans quelle mesure cette nouvelle tâche co-construite est identifiable, prévisible ?
   Intuitivement, et d'après ce que nous avons abordé précédemment, cette question sous-entend un "pourquoi" et un "comment".

(a) Pourquoi : Nous avons précédemment évoqué le fait que lors d'un dialogue, le locuteur possédait au moment de son énoncé une image de son interlocuteur, où il projette à la fois la façon dont son énoncé pourrait être interprété par l'autre puis comment cet autre lui répondrait. Ainsi, il possède également une image de lui-même se préparant à la réponse qu'il aurait imaginée de l'autre et ainsi de suite jusqu'à saturation de la mémoire de travail que nous abordons dans le chapitre 3. De ce fait, chaque participant possède un modèle de l'autre l'écoutant et lui parlant. Lors d'une conversation orientée-tâche comme celle du dialogue 1.1, l'utilisateur modélise l'image qu'il se fait de l'employé dans sa mémoire grâce à la fois à ce que lui projette l'employé à travers ses réponses (mémoire à court terme), sa propre expérience passée dans un contexte similaire (mémoire épisodique) et ses connaissances du monde (mémoire sémantique). Ce modèle, comme nous le verrons dans la section 3.1, est la raison qui conduit l'utilisateur à analyser, à partir de l'historique de la conversation, la tâche à résoudre, son expérience et sa connaissance du monde, la réponse de l'employé en se demandant si celle-ci ressemble à la réponse qu'il aurait considérée acceptable par la représentation qu'il se fait de l'employé.

(b) Comment : Tout simplement c'est la réalisation linguistique de l'énoncé de chaque participant qui permet à chacun d'ancrer et situer le tour de parole courant dans un schéma qui leur est familier dans leurs mémoires. Tel mot, ou son absence provoque irrémédiablement des rappels, des chemins à travers les mémoires permettant de décoder ce dernier. Cette réalisation linguistique, dans une définition plus formelle, correspondrait à l'activation d'une ou plusieurs des fonctions langagières de Jakobson [Jakobson, 1960]. Par exemple dans 1.1, l'employé, par son apparente froideur lors de ses réponses, active à la fois la fonction expressive (transmission des sentiments du locuteur) et conative (déclenchement des émotions chez le récepteur). L'employé reçoit comme information cette absence d'empathie de la part de l'employé et, étonné, se met à douter de l'humanité de ce dernier. Cela signifie que la réalisation linguistique de l'employé en réponse à celle de l'utilisateur a été à même de faire croire à celui-ci qu'il dialoguait avec un *robot*.

2. Dans quelle mesure peut-on évaluer la résolution de la tâche ? Il faut différencier ici le principe de résolution de tâche et de résolution - ou succès - du dialogue. De même, il faut différencier ici le principe de satisfaction de l'utilisateur et de succès du dialogue. D'après la littérature, [Testoni and Bernardi, 2021] la différence entre réussite de la tâche et réussite du dialogue se situe en cela que la réussite de l'un n'implique pas celle de l'autre. Une tâche est considérée résolue lorsque l'agent (ici l'employé)

1	Utilisateur : <SILENCE>
1	Employé : bonjour, je suis _Name2_. en quoi puis je vous aider ? afin d'accélérer le traitement de votre dossier, merci de nous indiquer votre adresse email et / ou votre numéro de commande.
2	Utilisateur : bonjour faites vous des livraisons au Canada ?
2	Employé : oui bien sûr, il vous suffit de sélectionner le Canada dans la liste déroulante lors de l'ajout de votre adresse
3	Utilisateur: 2ème question, avez vous Quatre en Rose de Boucheron pour femme ?
3	Employé : oui tout à fait, vous pourrez retrouver ce produit à l'adresse suivante : _URL1_
4	Utilisateur ce que je cherche c'est Quatre en Rose
4	Employé : excusez moi, j'avais mal compris. malheureusement non, ce produit n'est pas disponible sur notre boutique à l'heure actuelle
5	Utilisateur : merci de vos réponses. bonne soirée.
5	Employé : de rien, je suis à votre service

TABLE 1.2 – dialogue 2

répond correctement aux questions ou demandes de l'utilisateur. Un dialogue est considéré réussi s'il est 'finalisé'. On entend par dialogue finalisé [Salmon Alt, 2001] un dialogue possédant des marqueurs linguistiques de la part de chaque participant montrant qu'ils considèrent que le dialogue est finalisé. Ici, l'interruption brutale du dialogue (l'utilisateur ne répond plus aux messages de l'employé) nous conduit à penser que ce dialogue n'est pas réussi. Toutefois, la tâche principale de ce dialogue (le code promo et l'humanité de l'employé) est possiblement résolue.
En d'autres termes, on peut considérer que le dialogue a échoué alors même que l'employé a correctement répondu aux questions de l'utilisateur. On peut alors se demander si un dialogue pourrait réussir alors que l'employé ne répond pas correctement aux questions de l'utilisateur.

Dans ce dialogue 1.2, nous observons une erreur de la part de l'employé, entre le tour 3 et 4. Cette erreur est relevée par l'utilisateur. Cela déclenche le processus de réparation : *"il s'agit de la manière dont les interactants résolvent les problèmes liés à la production, à la réception et à la compréhension. Il s'agit là d'une séquence qui interrompt le cours d'action, le mettant en suspens le temps de la résolution du problème"* [Chernyshova, 2018, Schegloff et al., 1977]. Ici, la réparation est dite hétéro-réparée et hétéro-initiée (pas d'amorce de réparation dans l'énoncé de celui qui provoque le besoin de réparation ; la réparation est entièrement complétée par

l'autre participant). Cette erreur visiblement de compréhension

oui tout à
fait, vous pourrez retrouver ce produit à l'adresse suivante
ce *que je cherche* c'est Quatre en Rose

et sa postérieure réparation confirmée par les excuses de l'employé

excusez moi, j'avais mal compris.

ne répondrait-elle pas à un schéma bien plus classique d'une conversation entre un client et un employé que les échanges de 1.1 ? Ne serions-nous pas tentés d'affirmer que l'erreur est humaine et par conséquent, et bien que cela puisse paraître contradictoire, plus acceptable pour un client que les réponses presque robotiques de 1.1 ? La suite de 1.2 et la finalisation du dialogue poussent à répondre positivement à ces questions. En définitive, doit-on évaluer la résolution indépendamment de la finalisation du dialogue, ne peut-on pas dès lors considérer que naît une troisième mesure d'évaluation, capable à la fois d'estimer la résolution de la tâche co-construite et la façon dont celle-ci est résolue [André, 2010] ?

Dans le cadre d'un système automatique de dialogue orienté-tâche, nous nous trouvons face à une interaction humain-machine. Cela signifie que le système automatique reçoit en entrée un énoncé de l'utilisateur et il doit l'interpréter, l'analyser puis décider de la meilleure réponse à fournir. Pour fournir cette réponse, le système se base sur un modèle qui a appris (ou bien s'est entraîné) à lire ces locutions de l'utilisateur et à choisir quelle est la meilleure action à effectuer. Traditionnellement, les systèmes s'appuient sur le tour de parole courant ainsi que sur l'historique de dialogue afin de prédire la réponse suivante à donner. On sait que lors d'une conversation entre deux humains, celui qui écoute n'attend pas la fin de l'énoncé de celui qui parle pour prévoir la réponse à fournir, il anticipe sa fin [Allwood et al., 1992]. En outre, celui qui parle anticipe également la fin de sa propre locution et se prépare déjà à écouter la réponse de son interlocuteur. De cette façon, il y a un chevauchement mutuel dans l'enchaînement des tours de parole, aussi appelé interprétation incrémentale du dialogue [Schlangen and Skantze, 2009].
D'après cette explication, nous pouvons supposer que l'humain raisonne, pendant le dialogue, avec une dimension du passé (l'historique de dialogue), le présent (le tour de parole courant, sa taille, donc sa position relative dans le dialogue), mais aussi l'avenir (anticipation de la fin du tour de parole courant et de la possible réponse suivante). Cela pourrait expliquer la faculté des participants à un dialogue à s'auto-corriger ou à se corriger mutuellement, à ne pas rompre la fluidité du dialogue et à ne pas fournir des tours de parole incohérents.
Dans quelle mesure est-il alors nécessaire d'apprendre au modèle du système de dialogue à créer une représentation de son interlocuteur durant une conversation, et de se souvenir d'une réalisation linguistique qui permette à la fois le succès de

la tâche et du dialogue ?

Dans les systèmes de dialogue actuels les plus performants, les modèles les plus utilisés sont des encodeurs-décodeurs basés sur des réseaux de neurones, que ce soient des transformeurs [Vlasov et al., 2019], des modèles à plusieurs étapes de raisonnement [Wang et al., 2020c] ou des réseaux de mémoire s'inspirant des travaux en recherche cognitive et psychologique sur la mémoire humaine. Pendant nos travaux, nous nous sommes particulièrement intéressés à ces derniers car leur architecture semblait correspondre le plus à un schéma du fonctionnement de la mémoire humaine durant un dialogue [Chen et al., 2019b]. D'autre part, les réseaux de neurones de type tranformeurs ne permettant pas une grande explicabilité, nous ne savons pas si, durant son apprentissage, le modèle fait émerger les dimensions latentes de positionnement relatif du tour courant de l'utilisateur non seulement par rapport à l'historique de dialogue mais aussi à la durée restante du dialogue. Nous ne savons pas non plus si le système effectue une interprétation incrémentale en faisant émerger une représentation de l'utilisateur humain [Tseng et al., 2021] , représentation qui contient l'image de l'utilisateur et donc aussi l'image que l'utilisateur se construit du système lui-même.
En d'autres termes, comment créer un système de dialogue capable d'apprendre à identifier ses propres incohérences et à se corriger en temps réel, à l'image d'un humain ?

Pour fournir des éléments de réponse à cette question , nous proposons :

1. Chapitre 2, une revue de l'état de l'art des travaux sur les systèmes de dialogue orientés-tâche et plus particulièrement de ceux qui s'intéressent à l'émergence d'un modèle de l'utilisateur et d'un moyen de prévention des incohérences du système.

2. Chapitre 3, une description du fonctionnement de la mémoire humaine et son utilisation lors d'un dialogue orienté-tâche pour faire le lien entre les modèles actuels de système de dialogue et les différences de traitement de l'information lors d'une conversation entre l'humain et la machine.

3. Chapitre 4, un nouveau modèle de système de dialogue orienté-tâche basé sur les modèles de réseaux de mémoire de travail intégrant explicitement, pendant l'apprentissage, la position relative d'un tour de parole dans le dialogue ainsi que l'anticipation du tour de parole suivant.
4. Chapitre 5, notre procédure expérimentale permettant d'évaluer notre modèle à la fois sur la détection des incohérences du système et sur l'utilité de cette détection pour améliorer la qualité de la réponse finale du système. Nous évaluons notre modèle avec plusieurs métriques classiques en dialogue telles que décrites dans la littérature de référence [Budzianowski

et al., 2018b, Rosset, 2018].

5. Chapitre 6, une discussion sur les résultats des évaluations mais également sur les considérations éthiques et légales d'un tel modèle, ainsi que sur le transfert de compétence vers l'industrie afin de pouvoir mettre en production et en commercialisation des systèmes de dialogue basés sur ce type de modèles en montrant les difficultés actuelles pour effectuer ce transfert et sur le fossé séparant les méthodes d'évaluations actuelles des systèmes de dialogue avec la mesure de leur performance réelle lors de leur déploiement.

Pour nos travaux, nous utilisons notamment le jeu de données DSTC2[31] et ses améliorations[32], un corpus de dialogue humain-machine dans le domaine de la recherche d'un restaurant, afin de détecter les incohérences du système. Nous mettons au point également une fusion de plusieurs corpus et de leur ontologie respective, humain-machine comme humain-humain dans plusieurs domaines différents (dix-neuf au total), afin de trouver des points communs dans les erreurs du système en fonction du domaine, mais aussi en fonction du corpus d'origine du dialogue. Cette fusion nous permet d'établir un modèle du positionnement temporel d'un tour de parole donné dans un dialogue tout domaine confondu.

31. https ://github.com/matthen/dstc
32. https ://github.com/Divye02/baby-jarvis/tree/master/data/dstc2

Chapitre 2

État de l'art

Choice. The problem is choice.

Néo

Dans ce chapitre, nous définissons un système de dialogue, notamment en listant les différents synonymes que l'on trouve dans la littérature, afin de clarifier certains concepts. Dans l'imaginaire collectif, les systèmes de dialogue sont nés en 2011 avec SIRI d'Apple, suivis d'ALEXA, ... Certains connaissent également IBM Watson, notamment grâce à sa victoire contre un humain à un jeu télévisé appelé Jeopardy aussi en 2011. Pourtant, les systèmes de dialogue sont un sujet de recherche et de développement depuis 1966 avec ELIZA.
Nous résumons l'évolution de ce domaine de 1966 à nos jours. Nous proposons ensuite une revue de l'état de l'art des systèmes de dialogue orientés-tâche les plus récents. Enfin, nous nous penchons sur les systèmes industriels et commercialisés ainsi que sur les limites des technologies et méthodes actuelles.

2.1 Historique et définitions

2.1.1 Qu'est-ce qu'un système de dialogue ?

Une des définitions parfois admises du système de dialogue est qu'il s'agit d'un agent conversationnel [Jurafsky and Martin, 2017], c'est-à-dire un logiciel chargé de dialoguer avec l'utilisateur dans le cadre de l'interaction humain-machine (IHM). Cependant, une autre définition acceptée est qu'un agent conversationnel est une intelligence artificielle utilisant des techniques de traitement automatique des langues pour interagir, alors que le système de dialogue est un programme où l'utilisation d'IA n'est pas un pré-requis [Io and Lee, 2017]. [Radziwill and Benton, 2017] donnent la définition inverse. Une autre définition encore est que le système de dialogue représente le terme générique pour désigner toute interaction humain-

machine impliquant du langage , et que l'agent conversationnel est une catégorie de système de dialogue exclusivement oral [Cassell et al., 1994]. Toutefois, une autre définition énonce qu'un système de dialogue est un diminutif de système de dialogue parlé, par opposition à un système de dialogue écrit aussi appelé *chatbot* [Fryer and Carpenter, 2006]. Dans la littérature, nous trouvons aussi des exemples où le système de dialogue est synonyme de *chatbot*, contraction de l'anglais *chat* [1] et *bot*, lui-même la contraction de robot, autrement dit un robot de clavardage [2] [Hussain et al., 2019, Hancock et al., 2019]. Cependant, nous pouvons lire aussi des définitions où le système de dialogue s'oppose à *chatbot* en cela que le premier est le diminutif de système de dialogue orienté-tâche et le second de *chitchat* [3] *bot* [Jokinen and McTear, 2009, Jurafsky and Martin, 2017]. Il existe aussi une opposition entre système conversationnel et système orienté-tâche. Dans d'autres cas, notamment dans les médias, le *chatbot*, contraction de *chatter-bot* devient synonyme de système de dialogue orienté-tâche lui-même synonyme de système orienté-but opposé à système conversationnel [Liu and Mazumder, 2021]. Toutefois, d'autres définitions affirment que l'agent conversationnel est nécessairement incarné, contrairement au système de dialogue, en d'autres termes, l'expérience utilisateur ultime [Wik and Hjalmarsson, 2009].

Nous sommes d'accord avec [Bibauw et al., 2015] quand ils affirment qu'il n'y a pas de cohésion typologique à l'heure de définir le domaine du dialogue automatisé. Cela est résumé par [McTear, 2020] :"Le site *chatbots.org* [4] dénombre pas moins de 161 synonymes pour décrire les systèmes conversationnels. Il n'y a peu ou pas de cohérence dans l'utilisation des nombreux termes en recherche ou dans les médias [5]". Nous proposons alors une version mise à jour de notre nomenclature, [Schaub and Vaudapiviz, 2019b]. Pour précéder à cette nomenclature, nous précisons une définition, cette fois couramment acceptée, de l'IHM : entité, physique ou virtuelle, permettant le bon déroulement de l'interaction entre un humain et une machine. Bien que l'objectif demeure de proposer une nomenclature universelle, nous précisons que nous considérons celle-ci valable uniquement pour le français. Enfin, dernière précision, nous ne discutons pas des interactions multimodales incluant la vidéo (avec des agents conversationnels animés ou des robots par exemple) car nous estimons que cela dépasse le cadre de nos travaux, contrairement aux retranscriptions écrites de conversations orales. Cette nomenclature propose une définition des différents types et sous-types de dialogues automatisés, ainsi qu'une liste non exhaustive de synonymes pour ces titres :

1. **Système de dialogue**
   (a) **Définition** : C'est un ensemble de programmes informatiques possé-

1. *conversation* en anglais
2. contraction du terme *clavier* et *bavardage*
3. *conversation casuelle* en anglais
4. https ://www.chatbots.org/synonyms/
5. traduit de l'anglais

dant comme fonction première d'établir une conversation avec un humain. Cela ne signifie pas qu'un système de dialogue ne peut converser exclusivement avec des humains, mais comme nous considérons qu'elle fait partie de la famille de l'IHM, nous supposons que tout concepteur de système de dialogue pense d'une façon ou d'une autre à son utilisation par un humain.

(b) **Synonymes** :

— **Agent dialogique**. Nous considérons que le terme d'agent dans le cadre du dialogue désigne à la fois l'interface interagissant avec l'utilisateur et les programmes permettant cette interaction. L'interaction entre ces différents éléments se révèle être un système de dialogue.

— **Agent conversationnel**. Dans la mesure où, d'après la définition que nous avons établie dans l'introduction (1), la notion de dialogue se trouve corrélée à celle de conversation, nous les considérons synonymes.

— **Dialogueur**. C'est un terme à l'origine exprimant celui qui crée le dialogue, il a été adopté officiellement[6] en France comme un synonyme de système de dialogue. Sa non-utilisation nous contraint à ne pas le définir comme étant le terme générique.

— **Bot** Dans l'imaginaire collectif et les médias, le mot *bot*[7], à la différence de robot, n'évoque pas nécessairement la matérialisation visuelle (un avatar par exemple) du système de dialogue. Il ne précise pas non plus la voix prise par le système (orale ou écrite) ni son objectif (ouvert ou orienté-tâche). D'autre part, certains blogs ou médias distinguent le *bot* de l'agent, en cela que ce dernier, à la différence du *bot*, remplacerait l'agent humain à tous les égards -y compris dans son empathie et sa conscience autonoétique [Tulving, 1985][8], .Cependant, nous estimons que la différence est plus liée à des problématiques de marketing et de buzz qu'à une réflexion sérieuse sur la sémantique de l'un ou l'autre terme. C'est pourquoi nous considérons que **bot** reste le synonyme non-universitaire le plus acceptable pour le domaine du dialogue.

2. **Orientation :** Nous appelons *orientation* la fonction première du système. Nous distinguons trois orientations possibles pour un système de dialogue.

   (a) **Système de dialogue orienté-tâche** :

      i. **Définition :** C'est un système de dialogue conçu pour la résolution d'une ou plusieurs tâches pré-établies avant le début de tout dialogue et dont la résolution conditionne fortement la réussite de ce

6. https ://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037783813
7. prononcé *botte* ou encore *bôte*
8. conscience de soi, de qui il est

dernier (réservation d'un billet, d'une chambre d'hôtel, suivi d'une commande...).

ii. **Exemple** : Utilisateur : - "Je veux un avion pour Genève demain soir".
Système : - "Très bien. Au départ de quelle ville ?"
Utilisateur : - "Paris"
Système : - "Il y a un vol demain à 20h42 au départ de Paris-Orly. Quelle classe voulez-vous ?"
Utilisateur : - "Economique et côté fenêtre"
Système : - "Quelle classe voulez-vous ?"
Utilisateur : - "Economique j'ai dit"
Système : - "Très bien. Fenêtre ou couloir ?"
Utilisateur : - "A votre avis.."
Système : - "Je n'ai pas saisi votre réponse, pouvez-vous répéter ?"
Utilisateur : "Fenêtre"
Système : - "Très bien. Le vol de Genève a Paris-Orly a été réservé demain à 21h27 sur le numéro de réservation EF879. Votre compte a été directement débité d'une somme de 800$. Ce billet n'est ni échangeable ni remboursable. Au nom de toute l'équipe je vous souhaite une bonne journée.

iii. **Synonymes** :
   — **Bot orienté-tâche**
   — **Système de dialogue orienté-but**. Comme expliqué en introduction (chapitre 1), la tâche et le but peuvent être considérés comme synonymes pour le domaine du dialogue. Cependant, un système de dialogue possède toujours un **but**, ne serait-ce que le fait de parvenir à interagir avec un utilisateur. C'est pourquoi nous préférons le terme de **tâche**.

(b) **Système de dialogue ouvert** :

i. **Définition :** Système de dialogue conçu pour maintenir une conversation avec un humain. Le but n'est pas tellement de résoudre une tâche pré-établie, mais plutôt d'établir durant le dialogue des mini-tâches dynamiques à co-résoudre avec l'utilisateur (son nom, sa musique préférée, le temps qu'il fait...) afin de maintenir l'utilisateur en conversation. L'objectif pour le système est de finir par faire croire à l'utilisateur qu'il parle à un humain, c'est le fameux test de Turing.

ii. **Exemple** : Utilisateur : - "Alors comme ça tu es un chatbot"
Système : - "Plus ou moins oui, je préfère le terme système de dialogue si vous voulez bien."

Utilisateur : - "Bon, c'est bien"
Système : - "Ca peut aller"
Utilisateur : - "C'est une bonne situation, ça, chatbot ?"
Système : - "Mais, vous savez, moi je ne crois pas qu'il y ait de bonne ou de mauvaise situation."

iii. **Synonymes :**
- ***chitchat bot***
- ***social bot***
- **système de dialogue non-orienté-tâche**

(c) **Assistant virtuel** :

i. **Définition** : Un assistant virtuel est un hybride entre le système orienté-tâche et le système ouvert. C'est un système conçu pour bavarder mais dans un nombre fini de domaines, et avec certaines tâches pré-câblées (gestion d'agenda, prévisions météo...). C'est ce genre de systèmes que l'on trouve dans les mobiles intelligents (SIRI, Cortana...).

ii. **Synonymes** :
- **Assistant conversationnel**
- **Assistant personnel intelligent**
- **Assistant numérique personnel**

3. **Moyen de communication** :

(a) **Système de dialogue écrit** :

i. **Définition :** l'interaction est uniquement écrite.

ii. **Synonymes** :
- ***Chatbot***
- ***ChatterBot***
- **Système de messagerie instantanée**

(b) **Système de dialogue parlé** :

i. **Définition** : l'interaction peut être écrite ou parlée

ii. **Synonymes** :
- ***VoiceBot***
- **Assistant vocal**
- ***CallBot*** [9]

iii. **Autres**
- **Système de dialogue orienté-tâche pour l'interaction écrite** (aussi ***Chatbot*** **orienté-tâche**, ***bot conversationnel*** ou encore **agent conversationnel pour l'écrit**)

9. Ce terme est généralement réservé aux systèmes parlés utilisés par les services téléphoniques automatisés

Cette nomenclature permettra, peut-être, d'harmoniser l'utilisation des termes appartenant au traitement automatique du dialogue dans la littérature francophone.

2.1.2 Les premiers systèmes

Dans cette section, après avoir effectué une présentation chronologique des systèmes de dialogue, il sera question des systèmes plus récents.
Le premier système de dialogue, tout du moins celui qui correspond à sa description générique en 2.1.1 s'appelle ELIZA [Weizenbaum, 1966]. C'est un système de dialogue simulant un psychanalyste, basé sur un système symbolique qui reconnaît des formes et des mots-clés dans les énoncés de l'utilisateur, et les reformule en questions avec l'utilisation de patrons pré-construits. Plus que le premier système de dialogue, c'est le premier système à essayer de remporter le Jeu de l'Imitation [TURING, 1950], appelé plus communément le Test de Turing. Ce test, défini par Alan Turing comme la réponse à la question : "Est-ce que les machines peuvent penser ce que nous, humains, pouvons penser ?" Autrement dit, peut-on faire croire à un humain qu'une machine peut penser comme lui ? Dans ce cas, comme l'humain serait-il capable de savoir s'il parle à une machine ou à un autre humain ? Se pose donc la question de la possibilité d'une pensée, d'une conscience, d'une intelligence artificielle. Cette problématique possède des origines bien plus anciennes car nous pouvons remonter au XVIIe siècle et à René Descartes qui pose la question, non pas de la machine pensante mais du corps sans pensée qui serait capable d'imiter l'humain. A l'époque des Lumières, Denis Diderot écrivait dans *Pensées Philosophiques* : "Si un perroquet pouvait répondre à tout, j'affirmerais sans hésiter qu'il est intelligent", en d'autres termes, si une machine était capable de répondre à toute question d'un humain, nous ne serions plus capables d'affirmer que c'est une machine, donnant là un critère du Test de Turing. Pour clore cette parenthèse nous pouvons évoquer plusieurs œuvres culturelles faisant référence à ce fantasme de robots étant dotés de conscience comme dans le roman *Les androïdes rêvent-ils de moutons électriques ?* de Philip K. Dick, ou encore les films *Matrix* des sœurs Wachowski.
Pour en revenir à ELIZA, sa popularité est venue du fait que certains utilisateurs, notamment ceux qui étaient véritablement atteints de dépression, d'anxiété et autres troubles et maladies psychiques, finissaient par penser qu'ELIZA était un humain se faisant passer par une machine, voire pour les cas les plus aliénés, devenaient dépendants d'ELIZA, que ce soit parce que naissait en eux un sentiment amoureux ou parce qu'ELIZA remplaçait pour eux un vrai psychanalyste. C'est l'effet ELIZA [Hofstadter, 1996].
Le second système de dialogue est appelé PARRY [Colby, 1974], un système simulant un patient atteint de schizophrénie. Lors d'une évaluation pendant laquelle plusieurs psychanalystes devaient déterminer s'ils avaient conversé avec un humain ou avec PARRY, ils parvenaient à fournir la bonne réponse seulement 48% des

fois [Pinar Saygin et al., 2000]. Nous pouvons trouver également un projet appelé *PARRY encounters the DOCTOR* où l'on a fait dialoguer PARRY et ELIZA ensemble [10]. Jusqu'à la fin du XXe siècle, les systèmes de dialogue les plus répandus restent ceux basés sur le filtrage par motif, aussi appelé *pattern matching*. Un des plus célèbres s'appelle A.L.I.C.E [Wallace, 2009], ayant remporté plusieurs fois le prix Lœbner [11] qui récompense chaque année le système de dialogue étant le plus proche de passer le test de Turing. A.L.I.C.E est programmée sous formes de règles de filtrage par motif dans un langage inventé par le créateur de celle-ci, le AIML (artificial intelligence markup language), un dérivé du XML. La popularité de A.L.I.C.E à la fin des années 1990 est si importante que beaucoup d'imitations voient le jour à tel point qu'est apparu le terme "AliceBot" pour désigner les systèmes de dialogues basés sur le même principe comme par exemple le système appelé Mitsuku, aujourd'hui plus connu sous le nom de Kuki [12], créé par la société Pandorabots [13] et quintuple vainqueur du prix Lœbner. Alice a même inspiré le film de Spike Jonze, *Her*, racontant l'histoire d'un homme tombant amoureux d'un système de dialogue. Un historique plus complet des premiers systèmes de dialogue symboliques est disponible en ligne [14] ainsi que dans la littérature [Shum et al., 2018, Gao et al., 2018]. A partir des années 2000 et avec la puissance décuplée des processeurs, les systèmes à base de règles sont peu à peu abandonnés pour être remplacés par des systèmes contenant des modèles probabilistes basés sur l'apprentissage automatique.

2.1.3 Les systèmes statistiques

L'avantage des systèmes basés sur les règles, et ce qui représente aussi leur faiblesse, c'est leur précision "parfaite". En effet, comme les systèmes de dialogue déterministes se basent sur des patrons pré-conçus où, selon un énoncé donné de l'utilisateur à un tour t et un historique de dialogue h, un annotateur humain crée une règle ordonnant au système ce qu'il doit répondre, il n'y a pas la place pour l'erreur, pour la probabilité, pour l'incertitude. C'est pourquoi, au début du XXIè siècle, l'on observe le développement de systèmes de dialogues probabilistes, basés sur cette incertitude [Roy et al., 2000]. Une méthode efficace pour modéliser un système de dialogue pendant les années 2000 est nommée Processus Décisionnel de Markov Partiellement Observable (POMDP pour *partially observable Markov decision process*). Le POMDP comporte des éléments du modèle de Markov caché (HMM pour *Hidden Markov Model*) qui, à la manière d'une chaîne de Markov,

10. https ://datatracker.ietf.org/doc/html/rfc439
11. https ://fr.wikipedia.org/wiki/Prix_Lœbner
12. https ://www.kuki.ai/
13. https ://home.pandorabots.com/home.html
14. https ://medium.com/x8-the-ai-community/a-reading-list-and-mini-survey-of-conversational-ai-32fceea97180

tente de modéliser un automate à états finis, mais en cachant ces-dits états et en ne permettant que de visualiser les observations liées à chaque état, autrement dit, l'automate ne sait dans quel état il se trouve, il ne peut qu'en déduire des probabilités en fonction des observations générées par la transition de l'état précédent avec l'état courant. Le POMDP comprend également des composants du Processus Décisionnel Markovien (MDP pour *Markov Decision Process*) en cela qu'il incorpore le principe de récompense et de pénalité en fonction de l'état dans lequel il transite, ainsi que le principe d'action, qui fait qu'à chaque état, une action permettant de contrôler le système est déclenchée. Autrement dit, le POMDP non seulement établit des probabilités sur l'état courant dans lequel il se trouve sans pouvoir l'observer car il est caché, mais en plus calcule la probabilité de la récompense qu'il va obtenir s'il se déplace dans un certain état sachant l'état courant, l'état précédent et l'action qui a permis de passer de l'état précédent à l'état courant. Le POMDP est présenté par l'ensemble $\{ S, A, T, R, \Omega, O \}$ où S est l'ensemble fini des états possibles de l'automate, A l'ensemble des actions possibles sachant que l'on se trouve dans état si, T l'ensemble des transitions permettant de passer à l'état $si + 1$ sachant si et ai, Ω l'ensemble des observables, et O les fonctions d'observation établissant la probabilité d'observer un événement sachant un état s
Si l'on prend comme exemple le dialogue 1.1, on s'aperçoit que les tours 3, 4 et 5 peuvent être représentés comme un POMDP. En effet, lorsque l'utilisateur répond à la question de l'employé :

`avez-vous d'autres questions ?`

par

`oui`

l'employé ne sait ni quelle sera la question, ni la pensée de l'utilisateur, en d'autres termes, il ne sait ni où va le dialogue, ni où il se trouve. La seule information qu'il possède est que l'utilisateur a effectivement une autre question à lui poser. Cependant, rien n'indique l'intention réelle de l'utilisateur ni l'état suivant du dialogue. C'est ce qui amène l'employé à répondre :

`je vous écoute`

Des systèmes hybrides voient également le jour dans cette décennie, qui combinent à la fois des règles symboliques et une optimisation avec un POMDP [Paek, 2006, Lemon and Pietquin, 2012, Young et al., 2013] pour contourner l'aspect déterministe des règles et apporter une plus grande robustesse [Williams, 2008] à l'heure d'inférer l'action à entreprendre pour le système à un tour de parole donné. En effet, comme le POMDP cherche à maximiser la récompense liée à telle ou telle réponse fournie à l'utilisateur par le système de dialogue, il apprend par renforcement à faire de la gestion de dialogue [Gasic et al., 2010].
Que ce soient les règles, les POMDP ou d'autres méthodes de gestion de dialogue

[Hahn et al., 2011], ce sont des méthodes où les systèmes utilisent des réponses pré-câblées pour dialoguer. C'est ce qu'on appelle les méthodes basées sur l'extraction. Le système accède à une base de connaissances où sont stockées l'ensemble des réponses possibles que le système peut fournir en fonction de l'énoncé de l'utilisateur et de l'état de croyance du dialogue [Lee et al., 2009].
Les limites des POMDP résident, entre autres, dans leur gestion de l'historique de dialogue. En effet, les automates de Markov ne calculent la transition vers un état qu'en fonction de l'état précédent, autrement dit, un système de dialogue modélisé en POMDP ne prend en compte que le tour de parole précédent pour prédire le tour de parole courant, ignorant tout l'historique de la conversation. La seule façon alors pour pouvoir exploiter l'historique de dialogue est de manuellement ajouter l'information pertinente de l'historique $H = \{ h_0...h_t - 2 \}$ au tour de parole $t-1$ pour prédire t[Bennacef et al., 1996].

2.1.4 Des modèles neuronaux

C'est ainsi qu'à partir des années 2010 et la remise au goût du jour des réseaux de neurones artificiels [Rosenblatt, 1958, Lecun, 1985], en partie due aux progrès technologiques [Aslot et al., 2001], la communauté a commencé à développer des systèmes de dialogue basés sur les algorithmes d'apprentissage profond (*deeplearning*) appelés réseaux de neurones récurrents (RNN pour *recurrent neural network*) [Rumelhart et al., 1986] et plus particulièrement les réseaux à mémoire à court et long terme (LSTM pour *long-short term memory*) [Hochreiter and Schmidhuber, 1997] qui possèdent la particularité d'apprendre des séquences temporelles, et donc des historiques de dialogue [Su et al., 2017, Zilka and Jurcicek, 2015]. Cependant, par son architecture, le LSTM doit retenir des séquences entières avant de pouvoir effectuer une prédiction, c'est-à-dire que le réseau doit mémoriser mot à mot le tour de parole de l'utilisateur avant de pouvoir lui donner une réponse. C'est un mécanisme assez éloigné des modèles cognitifs, qui eux retiennent des concepts et des affordances mais pas l'intégralité d'une séquence [Mecklinger et al., 2004, Bailey et al., 2008, Macken et al., 2014]. Ainsi, pour permettre au LSTM de raisonner par filtrage des informations dans une séquence sans avoir à la mémoriser en entier, celui-ci se voit affublé d'un mécanisme d'attention qui change dynamiquement la valeur des poids dans une de ses couches pour obtenir une valeur dynamique et non pas statiques des vecteurs contextuels [Bahdanau et al., 2015]. Les modèles basés sur les LSTM bi-directionnels (Bi-LSTM) [Kim et al., 2019] se montrent encore plus efficaces car les informations retenues en mémoire appartiennent non seulement au passé mais aussi à l'avenir. Par exemple si la tâche consiste en la prédiction d'un mot à une position donnée dans la phrase, le Bi-LSTM observe non seulement les mots précédents, mais aussi les mots suivant celui à prédire, ce qui lui fournit plus d'informations et une plus grande probabilité

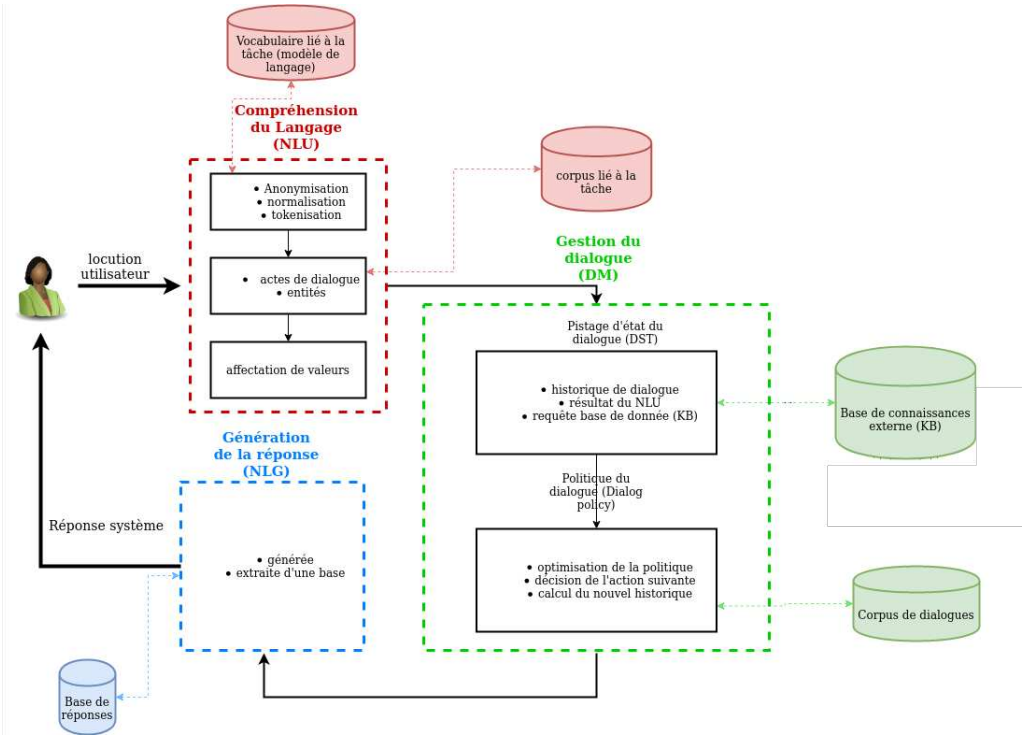

FIGURE 2.1 – Modèle classique de système de dialogue

d'inférer le bon mot.

Dans la seconde moitié de la décennie et jusqu'à nos jours, un terme est apparu concernant entre autres les systèmes de dialogue, mais pas uniquement, c'est celui de modèle de bout-en-bout (E2E pour end-to-end). Comme son nom l'indique, un système de dialogue basé sur un modèle de bout-en-bout contient un seul programme, qui permet d'effectuer à la fois la compréhension de l'énoncé de l'utilisateur, la gestion des états du dialogue, ainsi que la génération de la réponse. Les systèmes que nous avons évoqués en 2.1.2 et 2.1.3 ne concernent en général que la gestion du dialogue, car la compréhension de l'utilisateur et la génération de la réponse sont traitées par des méthodes symboliques pour la première et un index de réponses pré-conçues pour la seconde. Avant d'expliquer ce que cela implique, il nous semble judicieux d'ouvrir une parenthèse concernant l'architecture des systèmes de dialogue. Traditionnellement, nous distinguons les trois sous-tâches (aussi appelées modules) évoquées ci-dessus, qui, ensemble, forment un système de dialogue [Young et al., 2013] (voir 2.1).

2.1.5 Les modules d'un système de dialogue

La compréhension du langage

La compréhension du langage (NLU pour *Natural Language Understanding*) est le module qui interprète le tour de parole de l'utilisateur. De façon pratique, un énoncé se présente sous la forme d'une ou plusieurs phrases [Lopez, 1999], de taille variable. En fonction du thème du dialogue, mais également du numéro du tour de parole, certaines phrases seront plus ou moins courtes [DeVault et al., 2011]. Par exemple, lorsque l'énoncé du tour de parole précédent du système est une demande de confirmation : *"Est-ce bien le train de Paris à Lorient de 17h15 que vous voulez que je réserve ?"*, la réponse de l'utilisateur, et donc son énoncé, a de fortes chances d'être assez courte. L'énoncé est toujours lié à son contexte, donc à l'historique du dialogue. En d'autres termes, pour que le module de NLU devienne performant, il a besoin de posséder en mémoire suffisamment d'informations de l'historique du dialogue [Shawar and Atwell, 2003, Favre et al., 2017], sauf si le dialogue ne fait que commencer.
La question que l'on peut se poser est la suivante : comment sait-on si le module de NLU a vraiment compris l'énoncé de l'utilisateur, comment évaluer le module de NLU ? Naïvement, nous sommes tentés de répondre que la compréhension demeurant une tâche exclusivement humaine, seul l'humain peut évaluer la compréhension [Braun et al., 2017]. Cela signifierait qu'à chaque tour de parole de l'utilisateur, le système devrait être en mesure de reformuler l'énoncé de l'utilisateur, sous forme de résumé ou de paraphrase, pour convaincre ce dernier que sa demande a bien été comprise. Toutefois, lors d'une utilisation du système en situation réelle, il n'est pas imaginable d'implémenter une telle méthode.

Acte de dialogue	description
abandon	abandoning a unit, either choosing not to complete it or due to interruption
accept	responding in an active positive way
acknowledge	signalling decoding, understanding
acknowledgeThanks	responding to a thank you
agree	signalling explicit agreement
answer	answering a question
apologise	apologising
approve	expressing appreciation or approval
attribute	expressing attribution to someone
bye	saying farewell ; closing a dialogue
clarify	stating something in response to a request for clarification
complete	completing the interlocutor's move
conclude	indicating a (logical) conclusion
confirm	confirming a request for confirmation
contradict	contradicting the interlocutor
correct	correcting what the interlocutor has said
correctSelf	correcting one's own utterance
direct	eliciting the interlocutor's non-verbal response

Acte de dialogue	description
disagree	expressing disagreement
disapprove	expressing disapproval
disConfirm	negative response to a request for confirmation
echo	repeating the interlocutor's words for verification
elab	elaborating the answer to a question
emphatic	repeating something for emphasis, usu. yes, no, or a DM
enumerate	enumerating
exclaim	expressing emotion or surprise
explain	providing an explanation
expressAssumption	expressing an assumption
expressAwareness	expressing awareness, possibly knowledge of something
expressConviction	expressing conviction, e.g. through use of no doubt
expressDisappointment	expressing disappointment
expressDisgust	expressing disgust
expressDislike	expressing dislike
expressDoubt	expressing doubt
expressHope	expressing hope that something can or will be done, will happen
expressImPossibility	negative counterpart to the above
expressImProbability	negative counterpart to the above
expressIndifference	expressing indifference towards something
expressLiking	expressing a liking for something
expressNonAwareness	negative counterpart to the above
expressOpinion	expressing an opinion/evaluation
expressPossibility	expressing a possibility
expressPreference	expressing a preference
expressProbability	expressing the probability of something occurring
expressRegret	expressing regret
expressStance	expressing one's (neutral) attitude, e.g. through frankly (speaking), never mind, etc.
expressSurprise	expressing surprise
expressUncertainty	expressing uncertainty regarding something
expressWish	expressing a wish or desire
greet	greeting the interlocutor
hesitate	hesitating before the beginning of a turn/unit
hold	signalling to the interlocutor to hold the line, usually to look up information or to think
identifySelf	identifying the speaker's name/institution
init	initiating a new phase of the dialog
initConclusion	preface to a conclusion
initContrast	preface to a stating a contrast
initCounterExp	preface to stating something counter to expectation
initFollowUp	initiating one or more additional explanations
initGeneralisation	initiating a generalising statement, i.e. by using in general
initQ	preface to a query
initReason	preface to giving a reason
initSummary	preface to a summary
insult	insulting the interlocutor
listSequence	listing items using ordinals
muse	deliberating as to how to respond
negate	responding negatively
negateOpt	negative response to a request for an option
negatePreference	negative response to a request for information about a preference

Acte de dialogue	description
nominate	indicating that/when someone should speak in a sequence
offer	offering a service to benefit the interlocutor
pardon	signalling misunderstanding/the need for the interlocutor to repeat
phatic	semantically empty discourse-marking expression, such as initial you know
predict	predicting some future event
reaffirm	indicating that something is (assumed to be) part of the common ground
refer	indicating a deictic reference (neutral option)
referAmount	referring to an amount/number
referCondition	referring to one or more conditions
referConstraint	referring to one or more constraints
referDate	referring to a date
referDay	referring to a day
referDirection	referring to a direction
referDuration	referring to a duration
referHow	referring to how something is done
referLiking	referring to a liking
referOpt	referring to an option
referPerson	referring to a person/people (excluding vocatives)
referPlace	referring to a place/places
referPossibility	referring to a possibility
referProbability	referring to a probability
referProcess	referring to an ongoing action
referReason	referring to a reason
referThing	referring to a concrete or abstract object
referTime	referring to a specific (point in) time
referType	referring to a particular type of thing/entity
refuse	responding negatively to an offer, etc
reject	rejecting a proposal
rejectSelf	rejecting one's own ideas/suggestions
report	reporting what others, including the interlocutor, have said
reqApproval	requesting approval for a suggestion, plan, etc.
reqClarification	requesting a clarification
reqConfirm	requesting a confirmation
reqConfirmDay	requesting confirmation for a day mentioned
reqDirect	requesting a directive
reqHold	requesting a hold, see above
reqIdentification	requesting an identification from the interlocutor
reqInfo	requesting verbal information
reqInfoAge	requesting information about the interlocutor's or someone else's age
reqInfoAmount	requesting information regarding an amount
reqInfoAwareness	requesting if someone knows about/is aware of something
reqInfoConstraint	requesting information about potential constraints
reqInfoDate	requesting information about a particular date
reqInfoDay	requesting information about a particular day
reqInfoDirection	requesting information about a direction
reqInfoDislike	requesting information on whether the interlocutor dislikes something
reqInfoDuration	requesting information about the duration of an event, such as a journey

Acte de dialogue	description
reqInfoEntity	requesting information about a person or other entity
reqInfoExclaim	registering surprise or incredulity in the form of a question
reqInfoExists	requesting information about whether something exists
reqInfoHabit	requesting information about a habit
reqInfoHow	requesting information as to how to do something
reqInfoIntent	requesting information as to whether the interlocutor intends to do something
reqInfoLiking	requesting information about whether someone likes (doing) something
reqInfoLocation	requesting information about a location
reqInfoNeed	resquesting information about the necessity of something
reqInfoNumber	requesting information about a number or amount
reqInfoObjection	requesting information about whether the interlocutor does/would object to something
reqInfoOpinion	requesting information about the interlocutor's opinion
requInfoPreference	requesting information about whether a particular preference exists
reqInfoReason	requesting information about a reason
reqInfoTime	requesting information about a time
reqInfoTitle	requesting information about a title/term of address
reqInfoType	requesting information about the type of something
reqModal	requesting permission, advice, etc.
reqOpt	requesting an option
reqPrediction	requesting the interlocutor make a prediction
retract	retracting a statement
selfTalk	speaking to oneself (the speaker)
spell	spelling out something
state	conveying information/awareness
stateAmount	stating an amount/number, usu. in response to a request for information about an amount
stateAttempt	stating that an attempt was/is being made
stateCondition	stating a condition
stateConstraint	stating a potential constraint
stateDate	stating a date
stateDistance	stating a distance
stateDuration	stating a duration
stateHabit	stating a habit
stateIntent	indicating the speaker's intention
stateIntent-hold	indicating intention & potential delay in continuing the verbal interaction
stateNonConstraint	stating the absence of a constraint
stateNonIntent	stating the intention not to do something
stateNonOpt	stating the absence of a potential option
stateOpt	stating a potential option
stateReason	stating a reason
stateTime	stating a time
stateType	stating the type of something
suggest	proposing joint or interlocutor's potential action
swear	expressing an expletive
thank	thanking
thirdParty	speaking to someone who is not involved in the dialogue

Acte de dialogue	description
unclassifiable	a speech-act not classifiable according to the present scheme
uninterpretable	uninterpretable, due to missing or incoherent information

TABLE 2.2 – Tableau des actes de dialogue d'après [Weisser, 2016]

En pratique, la méthode encore aujourd'hui la plus répandue est d'évaluer les actes de dialogue, qui demeurent des valeurs discrètes, indexées dans des ontologies. Si les désirs de l'utilisateur sont le "pourquoi" de ses intentions, ses actes de dialogues en sont le "comment". [Bocklisch et al., 2017a] considèrent les actes de dialogue comme le couple intention-entité, où la première est un concept lié à l'action que désire effectuer le locuteur (poser une question, se plaindre, remercier, saluer, affirmer, nier...) et la seconde les objets caractérisant l'action (un nom de restaurant, un numéro de téléphone, une date, ou rien tout simplement). Les intentions sont dans ce cas-ci fortement dépendantes du domaine. D'autres conçoivent les actes de dialogue comme des caractérisations de trois grands sous-ensembles : **informer**, **demander**, et tout le reste [Mezza et al., 2018, Yu and Yu, 2021]. **Informer** étant l'action à travers laquelle l'utilisateur fournit les informations nécessaires à l'action du système, **demander** étant, elle, l'action à travers laquelle l'utilisateur interroge le système pour obtenir une information, les actes de dialogue sont alors indépendants du domaine, contrairement aux intentions qui caractérisent l'action entreprise par l'utilisateur en fonction du domaine. Les actes de dialogue sont accompagnés d'une entité et de sa valeur. Les autres actes de dialogue, moins fréquents et moins riches en information, ne sont pas considérés comme une catégorie à part entière. Nous nous basons sur la définition de [Kim et al., 2017b] qui décrit l'acte de dialogue comme le triplet domaine (thème du dialogue) - intention - entités. Dans ce modèle, chaque énoncé peut posséder plusieurs actes de dialogue. Cependant, un acte de dialogue ne contient qu'un et un seul domaine, une et une seule intention. Il peut par contre contenir plusieurs entités associées à la fois au domaine et à l'intention.
Plusieurs typologies des actes de dialogue ont été décrites [Stolcke et al., 1998, Weisser, 2016]. Nous pouvons également citer [Bunt et al., 2010] qui a proposé un ISO standard pour l'annotation des actes de dialogue. Celle de [Stolcke et al., 1998] au 2.1 est composée d'une quarantaine d'actes de dialogues considérés comme terminaux, c'est-à-dire qu'ils ne possèdent pas de sous-types. [Weisser, 2016] au tableau 2.2 affine encore plus ce modèle en en proposant plus de cent soixante, dont certains très détaillés au point que nous ne sommes pas capables de comprendre clairement la différence entre par exemple "negate", "refuse" et "reject".
Pour évaluer le NLU, on essaye de prédire les actes de dialogue sous-jacents à un énoncé. Jusqu'aux années 2000, on utilisait des techniques de machine à vecteur de support et des modèles de Markov cachés [Surendran and Levow, 2006]. Plus récemment, deux sous-tâches de NLU ont émergé : la détection des intentions, et le remplissage de cases (*slot filling*) qui correspond aux entités que doit trouver

NOM	Traduction	Exemple
STATEMENT	déclaration	Me, I'm in the legal department.
BACKCHANNEL [Ruede et al., 2017]	* voie rétrograde/canal arrière [15]	Uh-huh.
OPINION	opinion	I think it's great
ABANDONED	abandon	So, -/
AGREEMENT	acquiescement	That's exactly it.
APPRECIATION	adhésion	I can imagine.
YES-NO-QUESTION	question absolue	Do you have to have any special training ?
NON-VERBAL	son non langagier	Laughter Throat_clearing
YES ANSWER	oui	Yes.
CONVENTIONAL CLOSING	accord de fin de dialogue	Well, it's been nice talking to you.
WH-QUESTION	question relative	What did you wear to work today ?
NO ANSWERS	non	No.
ACKNOWLEDGMENT	compréhenion	Oh, okay.
HEDGE	vérification	I don't know if I'm making any sense or not.
DECLARATIVE YNQ	question absolue déclarative	So you can afford to get a house ?
OTHER	non défini	Well give me a break, you know.
BACKCHANNEL QUESTION	question par voie rétrograde	Is that right ?
QUOTATION	*citation	You can't be pregnant and have cats .
SUMMARIZE	résumé	Oh, you mean you switched schools for the kids.
AFFIRMATIVE	affirmation	It is.
ACTION DIRECTIVE	suggestion directive	Why don't you go first .
COMPLETION	complétude	Who aren't contributing.
REPEAT	répétition	Oh, fajitas .
OPEN QUESTION	question ouverte	How about you ? .
RHETORICAL	question rhétorique	Who would steal a newspaper ? .
HOLD ON	rétention	I'm drawing a blank.
REJECT	rejet	Well, no .
NEGATIVE NON-NO QUESTION	négation sémantique	Uh, not a whole lot.
SIGNAL NON UNDERSTANDING	non-compréhension	Excuse me ? .
OTHER ANSWERS	réponse non définie	I don't know .
CONVENTIONAL OPENING	accord de début de dialogue	How are you ? .
OR-CLAUSE	union	or is it more of a company ? .
DISPREFERRED ANSWERS	préference négative	Well, not so much that.
3RD PARTY TALK	Inclusion d'un troisième interlocuteur	M y goodness, Diane, get down from there.
OFFERS	offre	I'I1 have to check that out .
SELF-TALK	monologue	What's the word I'm looking for .
DOWNPLAYER	minimisation	That's all right.
MAYBE	peut-être	Something like that .
TAG-QUESTION	question-tag	Right ? .
DECLARATIVE WH-QUESTION	question relative déclarative	You are what kind of buff ? .
APOLOGY	présentation d'excuse	I'm sorry.
THANKING	remerciement	Hey thanks a lot .

TABLE 2.1 – Tableau des actes de dialogue d'après [Stolcke et al., 1998]

le système. Les modèles obtenant les meilleurs résultats pour le remplissage de cases sont construits avec des réseaux de neurones récurrents et des mécanismes d'attention [Raheja and Tetreault, 2019]. Ceux pour la détection des intentions sont construits sur des Bi-LSTM [Sreelakshmi et al., 2018] avec soit le modèle de plongement de mots ELMO [Peters et al., 2018] soit des transformeurs comme BERT [Devlin et al., 2019] pour mieux exploiter l'historique de dialogue lors de la prédiction. Les évaluations de cette tâche sont en général réalisées à partir du corpus SwitchBoard [Godfrey et al., 1992] et ICSI Meeting Recorder Dialog Act [Shriberg et al., 2004], Atis [Hemphill et al., 1990] et Snips [Coucke et al., 2018]. Pour les systèmes de compréhension orale, d'autres éléments entrent en compte pour évaluer le NLU tels que la prosodie [Shriberg et al., 1998, Mast et al., 1996] ou la reconnaissance de parole. Toutefois, nous nous intéressons pour nos travaux à l'interaction écrite.
Une liste des dernières méthodes et de leur évaluation pour toutes les tâches de NLU, notamment la classification jointe de l'intention et du remplissage de cases est disponible grâce à [Weld et al., 2021].

La gestion du dialogue

La gestion du dialogue (DM pour Dialogue Management) permet à un système de déterminer dans quel état de dialogue les locuteurs se trouvent, et quel est le prochain état à atteindre. Cela se décompose en deux sous-tâches, le pistage d'état du dialogue (DST pour *dialog state tracking*) et l'optimisation de la stratégie du dialogue (*dialog policy optimization*). Le pistage d'état du dialogue se définit comme la prédiction par un modèle de l'état courant du dialogue. Le système s'appuie sur les mêmes informations que le NLU, à savoir le tour de parole courant d'utilisateur et l'historique d'échanges entre l'utilisateur et le système. Il prend également en compte les résultats de prédiction d'actes de dialogue du dernier tour de parole utilisateur par le NLU. La particularité du pistage d'état du dialogue vient du fait qu'il prend également en compte son état de croyance (*belief state*). L'état de croyance d'un système correspond aux informations qu'il doit retenir tout le dialogue durant. Cela peut être des entités, les erreurs que le système a commises ainsi que le but final du dialogue si celui-ci est explicité comme dans certains corpus. Généralement, l'état de croyance se traduit par le triplet $domaine - entité - valeur$ [Ultes et al., 2017a]. L'état de croyance s'enrichit et grandit à chaque tour de parole, puisque les informations contenues dans chaque nouvel énoncé de l'utilisateur ajoutent des cases dans son état de croyance, ou mettent les existantes à jour. La tâche de pistage d'état du dialogue consiste en la prédiction du nouvel état de croyance.
Le problème du pistage d'état du dialogue vient du fait que le système doit avoir accès à l'intégralité du vocabulaire. En effet, dans la mesure où les actes de dia-

logue prédits contiennent un domaine et une entité, en plus d'une intention, le pistage d'état consiste, en fonction de ce domaine et de cette entité, à trouver la valeur correspondante. Cela signifie que le système doit enregistrer en mémoire l'intégralité des valeurs possibles pour un couple $(domaine - entité)$. En pratique, les ontologies ne sont jamais assez complètes pour y parvenir, il existe toujours des valeurs non indexées. Cependant, même dans le cas hypothétique où l'ontologie serait parfaite, il est très coûteux pour le système de devoir traiter toutes ces valeurs et de choisir laquelle il faut prendre.
Pour répondre à cette problématique, [Rastogi et al., 2017, Goel et al., 2018] ont mis point une méthode consistant à généraliser les valeurs possibles des entités de l'état de croyance afin de pouvoir générer la bonne valeur sans avoir à apprendre toute la liste de valeurs pour une certaine entité. Cette méthode a été combinée avec des RNN hiérarchiques, choisissant en fonction de l'entité, laquelle des deux méthodes est la plus adaptée, dans un système hybride appelé HyST [Goel et al., 2019]. Récemment, [Wu et al., 2020b] ont présenté une méthode de pistage d'état avec une représentation en graphe des couples $(entit - valeur)$, tandis que [Wu et al., 2020a] ont utilisé un transformeur avec des résultats similaires mais une plus grande adaptabilité à un domaine inconnu. [Zhang et al., 2020a, Zhao et al., 2021] ont combiné l'approche de HyST avec un transformeur pour améliorer encore les performances de ses prédécesseurs, mais en alourdissant et en complexifiant le modèle. On peut également citer le modèle TRADE [Wu et al., 2019] qui utilise un transformeur simplifié appelé DistilBERT [Sanh et al., 2019] ainsi que StateNet [Ren et al., 2018], basé sur des LSTM. [Kim et al., 2020] obtiennent des performances intéressantes en allégeant le modèle qui ne choisit que les informations les plus pertinentes à conserver d'un tour à l'autre pour recréer l'état du dialogue sans devoir tout copier à chaque itération. Enfin, [Xu et al., 2020] utilisent les réseaux de mémoire avec une couche d'attention pour obtenir à la fois un modèle allégé par rapport à un système basé sur un transformeur et avec des performances similaires. Une liste plus complète des différentes méthodes pour le pistage d'état et son évaluation a été établie par [Balaraman et al., 2021]. Les corpus de dialogue populaires pour cette tâche sont DSTC2 [Henderson et al., 2014a] et MultiWOZ2.0 [Budzianowski et al., 2018b].
L'optimisation de la stratégie de dialogue consiste en la recherche de l'action la plus adéquate à réaliser par le système pour se rapprocher le plus possible de la fin du dialogue, et de la satisfaction de l'utilisateur. Elle devient la suite logique du pistage d'état, qui lui-même se trouve dans la continuité des deux tâches de NLU. En effet, une fois que le système croit savoir (nous insistons bien sur ces deux verbes : là où dans la plupart des contextes, croire et penser peuvent être synonymes, nous considérons que la croyance et la pensée, dans le cas d'un système de dialogue, restent différents. La croyance exprime le fait qu'on est convaincu de posséder un savoir, une connaissance, souvent sans suffisamment de preuves pour que l'on puisse affirmer sans hésiter que l'on sait, on ne choisit pas de croire, on accepte de croire.

On se rapproche plus de la notion de foi que de savoir. La pensée, elle, exprime une réflexion, une remise en question, un échange cognitif entre plusieurs fonctions cérébrales mais surtout un choix [L'Haridon and Paraschiv, 2009], on choisit de croire, car on pense, on réfléchit. Le premier cas est possible pour une machine, le second relève (encore) de la science-fiction. Une machine ne connaît pas la notion de choix [Wachowski et al., 1999], elle accepte de considérer sa croyance comme un savoir. De la même façon, une machine ne sait pas. La notion de savoir est une considération socio-culturelle humaine voire animale, en aucun cas pour une machine) dans quel état se situe le dialogue, il décide de la meilleure action à effectuer. Un concours est organisé tous les ans depuis 2013 avec comme sujet la détection du pistage d'états dans le dialogue sous diverses formes et avec diverses contraintes [Henderson et al., 2014b, Li et al., 2020b, Kim et al., 2016, Kim et al., 2017a, Gunasekara et al., 2020, Henderson et al., 2014c]
Pour évaluer la stratégie de dialogue, on considère que l'action décidée par le système correspond à un ou plusieurs actes de dialogue, chacun composé d'un domaine, d'une intention et d'entités. Le silence est également considéré comme une action. Il suffit ensuite de comparer l'action effectuée par le système avec l'action que l'on attend qu'il accomplisse. Jusqu'aux années 2000, cette stratégie était encore réalisée avec des méthodes symboliques [Xu and Rudnicky, 2000] ou des apprentissages par renforcement [Litman et al., 2000, Singh et al., 2002]. Cependant, la rigidité de ces types de modèles ne permettait pas de créer un lien entre le pistage d'état du dialogue et la stratégie de dialogue. Les POMDP avec un processus gaussien [Gasic et al., 2010] ont permis de générer un état d'incertitude quant au résultat du pistage d'état du dialogue, ce qui a donné la possibilité aux systèmes de mieux corréler les différents états du dialogue et donc d'obtenir une stratégie de dialogue plus efficace. De nos jours, les modèles montrant les meilleures performances dans l'optimisation de la stratégie de dialogue se basent sur des méthodes neuronales comme le Bi-GRU [Su et al., 2019] ou des couches de transformeurs [Peng et al., 2021]. Ces méthodes possèdent la particularité d'encoder à la fois la séquence d'actes de dialogue et la séquence de mots composant les énoncés, ce qui améliore à la fois la stratégie de dialogue et la génération de la réponse du système.

La génération de la réponse

La génération de la réponse (NLG pour *natural language generation*) est un module légèrement différent des deux premiers, car il ne s'agit ni de classification d'intentions, ni de gestion de dialogue, mais bien de transformation des informations en langue naturelle [Gatt and Krahmer, 2018, Hovy, 1990]. C'est également le module qui est moins fréquemment traité séparément, contrairement aux prédictions d'actes de dialogue et à la stratégie du dialogue. Paradoxalement, c'est le module le plus important, en particulier lorsque le système est déployé en situation

réelle, car c'est lui qui produit la réponse que l'utilisateur lit et à laquelle il répond à son tour. Traditionnellement, les réponses n'étaient pas générées mais plutôt extraites d'une base de données ou d'un moteur de recherche, dont les requêtes étaient les résultats du gestionnaire de dialogue [Manzoor and Jannach, 2021]. Une autre technique classique était la création de patrons, créés automatiquement à partir des mots-clés obtenus par le gestionnaire de dialogue. Bien que certains systèmes statistiques aient existé [Hildebrand et al., 2005], leur efficacité par rapport à des méthodes symboliques demeurait discutable. Ce n'est que récemment que le module de NLG génère véritablement, mot à mot, caractère par caractère, voire syntagme par syntagme [Jagfeld et al., 2018], la réponse finale du système. Cependant, les modèles générant la réponse du système restent souvent une sous-partie d'un modèle effectuant également la gestion du dialogue ou la compréhension du langage [Liu et al., 2016a]. De nos jours, la plupart des modules NLG font partie intégrante de systèmes de bout-en-bout (voir 2.1.4).

2.2 Etat de l'art des systèmes de dialogue les plus récents

Dans cette section, nous nous penchons sur les systèmes de dialogue de bout-en-bout puis sur ceux possédant une mémoire inspirée des modèles cognitifs. Puis, nous nous intéressons aux systèmes de dialogue intégrant d'une façon ou d'une autre une réflexion sur un modèle de l'utilisateur. Un récapitulatif des progrès en matière de méthodes d'apprentissage pour les sytèmes de dialogue est disponible ici [16].

2.2.1 Les systèmes de bout-en-bout

La principale motivation de ces systèmes est d'éviter du travail manuel aux développeurs. Dans le schéma classique, chaque module est entraîné séparément, il faut donc préparer les données d'apprentissage ou les patrons de règles symboliques pour chacun d'eux, puis les assembler. De plus, dans la mesure où les modèles sont entraînés séparément, mais néanmoins dépendants, lorsqu'une erreur se produit dans l'assemblage, il est très difficile et chronophage de retrouver la source de l'erreur [Liu and Lane, 2018b]. Les systèmes de bout-en-bout ne sont une réalité que grâce à l'avènement des réseaux de neurones dans le domaine du traitement automatique des langues. En effet, un système de dialogue de bout-en-bout ne possède qu'un module qui encode à la fois le tour de parole de l'utilisateur, mais aussi les actes de dialogue qui lui sont liés, les actions du système et l'historique

16. https ://www.alibabacloud.com/blog/progress-in-dialog-management-model-research_596140

de dialogue. En sortie, le système doit uniquement fournir la réponse du système. L'avantage de ces systèmes est donc sa simplicité : nous passons de trois à un seul module. Ces systèmes sont plus faciles à prendre en main, plus faciles à entraîner, plus faciles à tester et plus faciles à évaluer. Par contre, un des inconvénients est la grande quantité de données à fournir au système pour qu'il obtienne des résultats satisfaisants [Serban et al., 2016]. L'autre inconvénient réside dans la gestion de la base de connaissances externe. En effet, dans les modules traditionnels, la base de connaissances n'intervient qu'au moment d'optimiser la stratégie de dialogue, pour que le système fournisse une information à aller chercher dans cette base. Dans les systèmes de bout-en-bout, la base de connaissances doit également être encodée et, en fonction de sa taille, peut allonger considérablement le temps d'apprentissage du modèle, voire empêcher le modèle de converger.
Nous avons déjà évoqué les modèles basés sur les Bi-LSTM (2.1.4), mais nous pouvons mentionner [Qun et al., 2020] qui combinent les Bi-LSTM avec un système d'intra-attention, permettant de calculer à chaque séquence (en l'occurrence des tours de parole) un vecteur d'attention pour chaque objet de la séquence. La récurrence du Bi-LSTM permet de calculer des interdépendances entre les objets d'une séquence vers l'avant mais aussi vers l'arrière. Pour des langues où la place des mots dans la phrase ne détermine pas leur fonction, il est indispensable de pouvoir établir des règles d'interdépendance, de subordination, et de complémentarité entre les mots qui ne soient pas uniquement liées à l'ordre dans lequel ils apparaissent. De même, pour la pronominalité, il est très important de pouvoir déterminer le substantif que remplace un pronom. Prenons les deux phrases suivantes :

(1) "*Le chat s'assoit sur le canapé car* ***il*** *est fatigué*"
(2) "*Le chat s'assoit sur le canapé car* ***il*** *est confortable*"

Dans chaque phrase, le pronom personnel ***"il"*** remplace un substantif, mais c'est la sémantique de son attribut qui nous permet de savoir lequel il remplace. Le vecteur d'attention permet de calculer la probabilité que le pronom soit lié à tel ou tel substantif, mais également de calculer la probabilité que ce soit tel ou tel attribut, ou complément circonstanciel qui le détermine.
Une autre approche de bout-en-bout qui a été développée pour les systèmes de dialogue est celle des réseaux génératifs antagonistes (GAN pour *generative adversarial networks* toujours au pluriels car c'est un système bicéphale) [Goodfellow et al., 2014] comme dans [Liu and Lane, 2018a, Olabiyi et al., 2019]. L'avantage de cette méthode est de simuler le retour que pourrait faire l'utilisateur de la réponse et donc d'avoir une meilleure précision dans la qualité de celle-ci. De plus, c'est une méthode qui est efficace sans nécessiter un gros volume de données.
[Cattan, 2020] se sont intéressé à l'adaptabilité dans les systèmes de bout-en-bout comme une nouvelle compétence, pour mieux apprendre à interagir dans des domaines sur lesquels ils sont peu ou pas entraînés.

Plus récemment, les systèmes les plus performants sont basés sur les transformeurs [17] [Yang et al., 2020, Su et al., 2021, Jeon and Lee, 2021], un type de réseau de neurones non récurrent, basé sur des mécanismes d'attention et des couches de neurones superposées (douze couches), devenus presque incontournables pour la modélisation de langue. Certains sont bien connus comme BERT ou GPT-2 et sont ceux les plus utilisés pour la tâche de séquence à séquence : une phrase est donnée en entrée (l'énoncé utilisateur), une phrase est attendue en sortie (la réponse du système). La différence entre les deux ainsi que le fonctionnement du transformeur dépassent nos compétences. Nous pouvons maladroitement résumer que BERT permet de générer des modèles de langue multilingues et est meilleur pour les tâches de classification et d'extraction d'information, alors que GPT-2 est plus performant pour les tâches de génération de langue. Il existe bien sûr des exceptions, mais cela reste une tendance, pour l'instant. L'avantage des transformeurs réside dans leur capacité à paralléliser l'encodage des données, et donc d'en traiter une grande quantité. L'inconvénient, mis à part le manque d'explicabilité des modèles [van Aken et al., 2019] (bien que certains travaux récents aillent dans le sens de transformeurs plus explicables [Rao et al., 2021, Li et al., 2021b]), est justement la taille des modèles pré-entraînés par un transformeur, qui alourdissent le module de bout-en-bout du système de dialogue. Comme nous l'avons évoqué, le système de bout-en-bout doit encoder toute la base de connaissances annexe à un système de dialogue afin de pouvoir fonctionner en déploiement. Cependant, certaines bases de connaissances étant volumineuses, si l'on additionne cela avec la taille des modèles de transformeurs pré-entraînés, la capacité mémorielle nécessaire des systèmes de dialogue tend à exploser. Lorsque nous évoquons la capacité mémorielle nécessaire, nous parlons en fait de la puissance de stockage totale nécessaire pour une machine afin de faire fonctionner un tel système, mais nous omettons de parler de la façon dont le système organise les informations au sein de cette puissance de stockage dans le but d'optimiser leur extraction (pour générer une réponse) ou leur indexation (pour enregistrer un historique de dialogue). En somme, comment fonctionne l'organisation de la mémoire d'un système de dialogue et comment l'optimiser.
Et puisque nous évoquons la mémoire, nous pouvons alors nous demander comment nous-mêmes traitons les informations durant un dialogue ; par conséquent, comment notre mémoire fonctionne-t-elle ?

2.2.2 Les systèmes avec réseaux de mémoire

Cette question, nous ne sommes pas les seuls à nous l'être posée. Nous expliquerons dans le chapitre 3 comment fonctionnent non pas notre mais nos mémoires

17. un classement officiel des meilleurs systèmes est disponible à https ://github.com/budzianowski/multiwoz

lors d'un dialogue. Inspirés par des modèles cognitifs de mémoire humaine, [Weston et al., 2015] ont mis au point un réseau de neurone appelé les réseaux de mémoire (MemNN pour *Memory Neural Network*). C'est une architecture de réseaux récurrents dans laquelle la mémoire du système est mise à jour à chaque nouvelle entrée d'information. Contrairement aux LSTM ou aux GRU, dans lesquels les informations sont entrées une et une seule fois, les MemNN s'appuient sur des sauts (ou *hops*) de réflexion, où le système va comparer plusieurs fois (trois dans la plupart des cas) l'information en entrée avec celles qu'il possède en mémoire afin d'émuler le système cognitif basé sur des aller-retour d'informations entre différentes mémoires avant de choisir la bonne réponse à fournir (voir 3) . Les MemNN, à l'origine conçus pour la tâche de question-réponse [Xiong et al., 2016], concurrencent voire dépassent en performance des architectures de système de dialogue basés sur les LSTM [Chen et al., 2018] voire sur des transformeurs [Chen et al., 2020a] sur la tâche de génération de réponse. La popularité des MemNN pour la génération de réponse dans les systèmes de dialogue orientés-tâche se doit d'une part à la façon dont il surpasse les LSTM dès qu'il est nécessaire d'aller chercher des informations éloignées du tour de parole courant dans l'historique de dialogue, mais aussi au fait qu'il puisse intégrer des bases de connaissances externes sans faire exploser ni la taille totale des modèles générés ni les cartes graphiques les calculant. D'une manière générale, la tendance de la littérature sur les MemNN montre des recherches pour obtenir une architecture de plus en plus inspirée des modèles cognitifs d'interaction dialogique et de mémoire. [Pei et al., 2021] mettent au point un MemNN coopératif dans le but d'établir ou de renforcer un profil-type d'utilisateur afin de mieux cerner sa demande et donc la meilleure réponse possible à celle-ci. [Chen et al., 2018] combinent les auto-encodeurs variationnels (VAE) [Kingma and Welling, 2014] et les MemNN pour obtenir une meilleure précision et un enrichissement des réponses du système aux énoncés de l'utilisateur.
Cependant, deux problèmes se posent :

1. Peut-on séparer dans l'architecture du système l'historique de dialogue et la base de connaissance ?
2. Dans quelle mesure peut-on faire en sorte que les dépendances linguistiques à long terme soient correctement prises en compte par le modèle tout au long du dialogue ?

Pour la première question, [Madotto et al., 2018] ont montré qu'en construisant une architecture basée non pas sur un encodage mais deux décodages, un pour l'historique et l'autre pour la base de connaissances, le système était capable de mieux comprendre les informations de l'utilisateur et donc de mieux gérer le dialogue. Pour la seconde question, [Chen et al., 2019c] ont développé un système fortement inspiré des études en psychologie sur le fonctionnement des mémoires humaines, le WMM2Seq, un système de MemNN dans lequel se trouve au centre une mémoire de travail (*working memory*) qui décide, à chaque mot de la réponse

du système à générer, si celui-ci provient de l'historique de dialogue, de la base de connaissances, ou de la question de l'utilisateur. Ce système permet à la mémoire de travail d'apprendre non seulement à choisir la source de sa réponse mais aussi à oublier au fur et à mesure (en baissant les poids) les éléments non pertinents de l'historique de dialogue. C'est un des systèmes que nous utilisons comme référence dans nos travaux et nos expérimentations (voir 5). [Wang et al., 2020a] ont tenté de combiner les deux approches afin de résoudre les deux problèmes, cependant le système n'est pas capable de retenir un état du dialogue ce qui affaiblit les performances lorsque la quantité d'information à retenir pour un tour donné croît.
L'autre question importante lorsque l'on étudie la mémoire humaine est la nature des informations enregistrées ainsi que la manière dont l'enregistrement s'effectue. En d'autres termes, il est nécessaire de comprendre, lors d'un dialogue entre deux personnes, comment chacun stocke ce qu'il dit et entend afin de pouvoir mener à bien (ou pas) l'interaction. Nous avons évoqué en introduction le phénomène de conscience de l'autre, où le locuteur possède une image de celui qui l'écoute, et inversement, ce qui permet la fluidité et la cohérence de l'échange. Cependant, quelle est concrètement cette image de l'autre et comment se forme-t-elle ? [Fuster, 2004] ont montré que l'enregistrement des informations en mémoire s'effectue grâce à la paire perception-action : la perception est la donnée qui arrive de nos sens (ouïe, vue...) à notre cognition (transformation de la donnée en information), alors que l'action est ce que notre corps et notre cerveau réalisent au moment de la perception. C'est le nombre élevé de paires perception-action similaires qui vont ensuite nous permettre de nous souvenir de la meilleure façon d'agir dans une situation donnée. Par exemple, lorsque nous avons soif et que nous prenons en main une bouteille d'eau, nous ne réfléchissons plus à comment l'ouvrir, nous savons qu'il faut tourner le bouchon dans le sens inverse des aiguilles d'une montre. Cela s'explique par le nombre de fois que nous avons vu effectuer ou effectué nous-mêmes l'opération.
Dans le domaine de la traduction automatique, [Ma et al., 2019] a montré qu'un modèle de traduction simultanée peut devenir plus rapide en anticipant les séquences, avec des résultats proches de la traduction de phrases complètes.
Pendant le dialogue, le locuteur spécule sur l'interprétation que fait l'écoutant de son énoncé, mais imagine aussi comment l'écoutant va comprendre et répondre à son énoncé, en se mettant à la place du locuteur. Autrement dit, le locuteur s'imagine l'écoutant en train de s'imaginer le locuteur. Il en va de même pour l'écoutant [Whiteside, 1993], ce phénomène est appelé la chaîne de discours. Chaque participant possède donc une image de l'autre mais également la représentation de l'image que l'autre a de lui. Si l'on veut schématiser, on pourrait dire que la conversation ressemble à une partie d'échec dont chaque tour peut être traduit par "S'il déplace la pièce X ici, alors je devrais placer la pièce Y là, mais alors il déplacera ça là... Mais s'il anticipe que je vais placer Y, alors il changera et placera Z à cet autre endroit alors je...". Cependant, à la différence d'une partie d'échec où il n'y

a pas d'interaction autre que par le déplacement des pièces dans le but de gagner la partie, un dialogue est une co-construction, aussi appelée *common ground* [Udagawa and Aizawa, 2020] où chaque participant, en se mettant à la place de l'autre, collabore au bon déroulement de la conversation. On peut alors se demander si dans un système de dialogue, il faudrait intégrer non seulement un modèle de l'utilisateur mais aussi un modèle du système. Certains travaux ont été dans ce sens pour la reconnaissance automatique de la parole [Becerra et al., 2018, Tjandra et al., 2017] pendant le dialogue, mais à notre connaissance il y a pas encore de modèle apprenant explicitement cette fonction cognitive de réflexivité.

2.2.3 Les systèmes avec modèle de l'utilisateur

Un modèle utilisateur est, selon [Wahlster and Kobsa, 1989], "Un modèle utilisateur est une source de connaissance dans un système de dialogue en langue naturelle qui contient des hypothèses explicites sur tous les aspects de l'utilisateur qui peuvent être pertinents pour le comportement de dialogue du système. Ces hypothèses doivent pouvoir être séparées par le système du reste de la connaissance du système." Cet intérêt pour le modèle de l'utilisateur n'est pas nouveau. Dès les années 1980, des études sur des modèles utilisateur dans les systèmes de dialogue orientés-tâches ont vu le jour [Wahlster and Kobsa, 1989, Morik, 1985]. Ils ont montré que pour pouvoir apprendre la sémantique des tours de parole et les intentions sous-jacentes de l'utilisateur, il était essentiel d'entraîner le système à modéliser les désirs et les actes de dialogue de l'utilisateur à travers ses énoncés. [Shang et al., 2020] a obtenu des résultats au niveau de l'état de l'art pour la classification des actes de dialogue en étiquetant le changement de locuteur dans les tours de dialogue pendant l'apprentissage, ce qui montre l'importance du rôle dans la conversation. [Auguste et al., 2019] vont plus loin en apprenant à classifier l'acte de dialogue du tour de dialogue actuel et du suivant, avec des résultats comparables aux références. Enfin, [Lin et al., 2020] créent un système appelé "Imagine then Arbitrate" (ITA) pour apprendre quand répondre et quand écouter, en imaginant ce que l'utilisateur va dire pour anticiper les erreurs possibles du système. [Whitney et al., 2017] modélisent avec un POMDP [Sammut and Webb, 2010] l'incertitude d'un agent de dialogue lorsqu'il répond à une question de l'utilisateur afin d'améliorer la précision de sa réponse.
Comme nous l'avons expliqué en 2.1.5, une des problématiques de l'interaction humain-machine est celle du choix. La machine, en tant qu'automate, ne connaît pas la notion de choix. Ainsi, pour un énoncé de l'utilisateur donné, le système donne une et une seule bonne réponse possible.
Une autre caractéristique du modèle utilisateur est de considérer que la concentration et la rétention d'information durant le dialogue fluctuent en fonction du temps qui s'écoule. [Xu and Reitter, 2018] ont montré que durant une conversation, la

concentration des participants était la plus haute au début et à la fin du dialogue et que le pic d'informations échangées se trouve environ au milieu du dialogue . [Dong and Wyer, 2014] ont montré également que la perception de la longueur de la conversation par les participants influaient sur la capacité de concentration de chacun d'eux et pouvait provoquer un déséquilibre dans la collaboration durant le dialogue. Il est fréquent également que les informations présentes en début de conversation sont peu à peu oubliées ou pas utilisées au fur et à mesure que le dialogue progresse [Boland, 2019] car la mémoire de travail se surcharge, sans pour autant perdre la fluidité du dialogue. Cela expliquerait les difficultés des modèles à gérer l'historique de conversation dès lors que celui-ci atteint une certaine taille. D'autre part, les systèmes actuels peuvent savoir combien de temps s'est écoulé depuis le début du dialogue. Par contre, ils ne prédisent pas encore le temps qu'il reste pour la fin du dialogue. [Quan and Xiong, 2020] ont expliqué que plus le dialogue se prolonge dans le temps, et plus la qualité des réponses du système diminue. Toutefois, [Sankar et al., 2019] ont montré que la longueur du dialogue n'affectait pas toujours les performances du système, à l'inverse de [Li et al., 2021a]. Enfin, [Papangelis et al., 2017a], en cherchant (et trouvant) des indices acoustiques pour prédire entre autres la taille totale d'un dialogue à un tour de parole donné, ont montré que le calcul de la croissance de l'entropie dans un dialogue permet de prédire sa durée de façon précise. L'utilisation de cette prédiction, considérée comme un indice supplémentaire pour la génération de la réponse, améliore la précision de celle-ci. [Walker et al., 2000] ont établi un lien entre la taille du dialogue et les erreurs qui apparaissent dans les réponses du système pour l'oral.

- Toutefois, à notre connaissance, il n'existe pas encore de méthode qui prenne en compte à la fois des modèles d'utilisateur et de soi ainsi que les considérations du choix inhérentes à l'humain et la prédiction du nombre de tours de parole restants au dialogue à un tour de parole donné. Toutes ces caractéristiques peuvent être définies comme certaines des dimensions mémorielles des systèmes de dialogue orientés-tâche.
Décrivons maintenant certains des jeux de données utilisés dans la littérature des systèmes de dialogue orientés-tâche.

2.3 Les différents corpus

Ces jeux de données de dialogue, à savoir des retranscriptions orales ou écrites de conversations réelles entre humains ou entre un humain et une machine sont appelés corpus. Ils constituent un ensemble de dialogues liés entre eux par une thématique, une source... Chaque dialogue peut également posséder des annotations qui viennent enrichir le texte.
La plupart des corpus de dialogue disponibles sont en langue anglaise. Il existe

quelques corpus de dialogue en français comme le corpus MEDIA [Devillers et al., 2004b], un corpus d'un peu plus d'un millier de dialogues dans le domaine de l'information touristique. Il a été constitué avec la technique du compère. Les instructions données au compère ainsi qu'à l'utilisateur peuvent servir de métadonnées équivalentes à des actes de dialogue, des intentions et des entités. Dans des travaux futurs, il serait intéressant d'explorer la possibilité d'exploiter à nouveau ce corpus. Nous pouvons également citer le corpus RITEL [Galibert et al., 2005] mais sa faible taille (trois cent soixante dialogues) nous a semblé un frein pour son exploitation dans une démarche de système doté d'un modèle à apprentissage profond. Toutefois, les corpus permettant d'évaluer à la fois le NLU, le pistage d'état du dialogue et le NLG ne sont pas nombreux. [Serban et al., 2015] dresse une liste de nombreux corpus de dialogue disponibles pour la plupart en licences libres, servant à la fois pour l'apprentissage de systèmes et pour leur évaluation [18].
Dans nos travaux, nous utilisons quatre corpus.

1. Tout d'abord nous étudions DSTC2 (dialogue state tracking challenge 2) [Henderson et al., 2014a] qui est notre seul corpus de dialogue humain-machine orienté-tâche dans le domaine de la restauration. Pour créer ce corpus, des utilisateurs possédaient des buts clairement définis en amont du dialogue, et devaient poser un certain nombre de questions au système. Le système était un POMDP pré-entraîné de l'université de Cambridge.
2. Nous utilisons aussi le CamRest676 [Wen et al., 2016c, Wen et al., 2016b], un corpus semblable au DSTC2 car ils partagent le même domaine, la même base de connaissances et la même ontologie. Cependant, le CamRest676 est construit par la méthode du compère, donc ce n'est pas de l'interaction humain-machine, mais du dialogue entre deux humains.
3. Nous exploitons aussi le célèbre MultiWOZ, un corpus construit aussi grâce à la méthode avec compère. C'est un corpus multi-domaine [19] qui a eu le droit à plusieurs améliorations [Eric et al., 2019, Ye et al., 2021a, Han et al., 2020a, Zang et al., 2020].
4. Enfin, nous utilisons le corpus SGD, un corpus basé sur des schémas de dialogue [Rastogi et al., 2020]. C'est un corpus composé de plus de vingt-mille dialogues multi-domaines (seize domaines). Contrairement à MultiWOZ, il n'est pas construit avec un compère, mais ce sont deux agents, chacun jouant soit le rôle de l'utilisateur soit celui du système, qui conversent selon des schémas préétablis et dont les agents jouant le système simulent des appels API à des bases de données externes.

Le point commun entre ces quatre corpus vient de leurs annotations (en actes de dialogue, état de dialogue, but du dialogue) ainsi que de leurs domaines. En effet,

18. http ://nlpprogress.com/english/dialogue.html
19. attraction touristique,réservation d'hôtel, commande d'un taxi, réservation d'une table de restaurant, appel de la police, contact d'un hôpital, billet de train, achat d'un ticket de bus

ils ont tous en commun de traiter au moins le domaine de la réservation ou de la recherche d'un restaurant.

2.4 Plusieurs méthodes d'évaluation

L'évaluation d'un système de dialogue orienté-tâche, par sa structure même (trois modules différents plus la satisfaction de l'utilisateur), est compliquée à réaliser dans sa globalité. Dès lors, comme l'évoque [Chen et al., 2017], on utilise différentes métriques pour chaque module du système. Pour la compréhension de l'entrée, on utilise les métriques classiques de fouille de texte [Yong Nahm and J. Mooney, 2002] pour déterminer la capacité d'analyse du système. Notons que par le passé [Devillers et al., 2003] ont proposé d'évaluer la compréhension en contexte en résumant l'historique du dialogue au moyen d'une paraphrase.
L'atelier défi sur le pistage d'état du dialogue a été créé (DSTC *dialog state tracking challenge*) pour évaluer les aspects gestion et décision du dialogue, pour lesquels plus de dix métriques différentes ont été développées. Il en est aujourd'hui à sa dixième version [20]. Pour la génération de réponses, ce sont des métriques de traduction automatique et de résumé automatique : BLEU, ROUGE et METEOR [yew Lin, 2004, Papineni et al., 2002a, Lavie and Agarwal, 2007] qui sont utilisées pour évaluer les systèmes (à l'exclusion des systèmes *retrieval-based* qui ne génèrent pas de réponse). Cependant, ces métriques ne reflètent pas la véritable performance du système et peinent à évaluer la satisfaction de l'utilisateur comme souligné par [Liu et al., 2016a] qui proposent une version améliorée de ces trois mesures. En effet, bien qu'il existe quelques campagnes d'évaluation de la satisfaction utilisateur sur les systèmes de dialogue [Forbes-Riley and Litman, 2006, Dzikovska et al., 2011] celles-ci sont surtout destinées aux systèmes de dialogue orientés enseignement, et sont difficilement adaptables et généralisables à tout système de dialogue orienté tâche. Dès lors, les évaluations les plus fiables restent les campagnes d'évaluation des systèmes de dialogue [Devillers et al., 2004a] ou les méthodes Test de Turing ou de système avec compère (Magicien d'Oz) [Kelley, 1983], cette dernière permettant en outre d'augmenter artificiellement la taille des corpus d'interaction. On peut également citer le protocole PARADISE [Walker et al., 1997a] développé dans le cadre du projet COMMUNICATOR [Walker et al., 2001] du DARPA et qui repose sur la corrélation entre la satisfaction de l'utilisateur et des mesures de performance objectives, pour proposer un protocole qui soit indépendant de la tâche.
Récemment, [Su et al., 2018b] présentent un agent conversationnel capable de s'évaluer à chaque tour de parole et de se corriger "en temps réel" en calculant à chaque interaction le taux de satisfaction de l'utilisateur, ainsi qu'en adaptant ses

20. https ://dstc10.dstc.community/

réponses selon ce taux.

2.5 Boîtes à outils et applications industrielles

2.5.1 Explorations universitaires orientées applications

Mis à part les entreprises, de nombreux laboratoires se sont penchés sur la question des agents conversationnels. Nous pouvons notamment citer en France le LIMSI qui a été l'auteur de plusieurs projets, comme par exemple RAILTEL [Bennacef et al., 1996], ou encore ARISE [Lamel et al., 1998] qui ont fait suite à la réalisation de la borne MASK [Gauvain et al., 1997] de réservation ferroviaire réalisée en partenariat avec la SNCF, un projet novateur pour son époque. D'autres travaux, toujours en collaboration avec la SNCF, ont également été menés par le LORIA en 2002 et 2003 pour le projet européen OZONE [Issarny et al., 2005] afin de développer sur tablette, une IHM pour la réservation de billets. De même, le projet AMITIES [Hardy et al., 2004], issu d'une collaboration entre l'Union Européenne (FP6) et le DARPA, a servi de cadre au développement d'un Serveur Vocal Intéractif prototype pour les banques.
Auparavant, le système GUS [Bobrow et al., 1977] proposait un système prototype pour la gestion du trafic de fret avec des résultats décevants. Les recherches universitaires ont parfois donné lieu à de nouveaux langages dédiés à la réalisation d'agents conversationnels comme AIML, devenu populaire grâce à A.L.I.C.E [S. Wallace, 2009] ou comme VoiceXML [Larson, 2003], conçu pour les serveurs vocaux interactifs. Les réalisations présentées dans cette section sont emblématiques d'une époque où les applications phares, indicatrices de ruptures technologiques en cours, provenaient principalement du monde universitaire à partir de financements publics.
Récemment, les travaux de [Campillos Llanos et al., 2017] ont abouti à un système de dialogue appelé PatientGenesys [21] qui simule un patient pour entraîner les jeunes médecins à la consultation.

Solutions et boîtes à outils grand public

Suite à l'amélioration des réseaux de neurones et leur popularisation [Lecun, 1985], les GAFAM [22] ont été les premiers acteurs du marché à s'emparer de la technologie. Outre leurs premiers produits finis (Google Assistant pour Google, Alexa pour Amazon, Siri pour Apple, Cortana pour Microsoft), ils ont également racheté des sociétés créant des environnements de développement (*framework*) afin

21. https ://www.theconnectedmag.fr/patient-genesys-virtuel/
22. Google, Amazon, Facebook, Apple, Microsoft

de les proposer au grand public. Google rachète en 2016 api.ai [23], appelé désormais DialogFlow et basé sur des RNN. Facebook acquiert wit.ai, Amazon construit son propre environnement de développement Lex (technologie utilisée par Alexa, l'assistant personnel d'Amazon disponible sur les enceintes connectées de la gamme Echo) [Janarthanam, 2017].
D'autres plateformes plus indépendantes donnent accès à des environnements de développement comme recast.ai ou encore Destygo, une start-up spécialisée dans la mobilité. Toutes ces plateformes proposent généralement deux accès à leur environnement de développement : un accès web et un accès sous forme de boîte à outils, téléchargeable et intégrable dans différents langages de programmation (Python, JS, etc.). Toutefois, les GAFAM ne sont pas les seuls à proposer des environnements de développement d'agents conversationnels. En effet, OpenDial [Lison and Kennington, 2016] propose une boite à outils en Java pour la création de systèmes de dialogue, mais aussi PyDial, [Ultes et al., 2017b], une boite à outils en Python créée par l'Université de Cambridge. Nous pouvons aussi mentionner ChatterBot [24] une boite à outils en Python.
Dernièrement, deux environnements de développement de système de dialogue en libre accès paraissent se dégager du lot. D'abord, nous trouvons *RASA* [Bocklisch et al., 2017a] qui est remarquable surtout pour son module de compréhension de la langue naturelle, car il utilise FastText [Joulin et al., 2016] et les plongements de mots [Mikolov et al., 2013] pour obtenir un modèle de langue spécifique au domaine d'application du système de dialogue. Cela rend la classification textuelle très efficace [Braun et al., 2017]. Ensuite, il existe *DeepPavlov* [Burtsev et al., 2018], une boîte à outil conçue pour deux cas d'usage : l'implémentation d'un *chatbot* prêt à l'emploi, et la recherche scientifique en systèmes de dialogue ou TAL.

2.5.2 Limites de l'industrialisation

Malgré l'engouement pour les systèmes de dialogue, il n'existe pas à ce jour un système commercialisé qui mette toute la communauté d'accord sur sa qualité. En effet, nous pouvons citer cet article de *chatbotslife.com* [25] qui explique qu'un *chatbot* demande beaucoup de ressources linguistiques pour être efficace. Ces ressources manquent également à IBM qui annonce licencier entre 50 et 70% des effectifs de sa branche Watson [26]. De plus, il faut préciser que ces *chatbots* ne possèdent pas de méthode d'évaluation officielle ou admise. [Landragin, 2013] explique par

23. https://stackshare.io/stackups/api-ai-vs-dialogflow
24. https ://github.com/gunthercox/ChatterBot
25. https ://chatbotslife.com/what-chatbots-can-learn-from-the-death-of-facebooks-failed-ai-acc05ad6d822
26. https ://www.usine-digitale.fr/article/licenciements-tensions-internes-ibm-watson-health-serait-dans-la-tourmente.N705304

ailleurs qu' "[i]l risque d'exister bientôt autant de méthodologies d'évaluation que de systèmes proprement dits. On peut s'interroger sur le bien-fondé d'une méthode d'évaluation proposée par les concepteurs d'un système dans le but d'évaluer ce seul système, la méthode étant elle-même évaluée par son application au système en question. Il est alors très difficile de savoir où va la technologie et comment l'amener plus loin" . Par ailleurs, les *chatbots* industriels demeurent des boîtes noires ne permettant pas de prendre facilement connaissance des méthodes avec lesquelles ils ont été constitués. Les corpus d'apprentissage par exemple, sont généralement inconnus lorsque le *chatbot* ne constitue pas une expérience scientifique. D'autre part, il est important de souligner que malgré le grand nombre de travaux universitaires, l'implémentation des systèmes de dialogue et leur commercialisation se soldent souvent par un échec. En effet, les utilisateurs se sentent rapidement frustrés de ne pas avoir des résultats satisfaisants et délaissent vite les agents conversationnels au profit d'êtres humains. Enfin, avec la nouvelle réglementation RGPD ainsi que les considérations environnementales, la marge de manœvre dans l'industrie pour la production et la commercialisation de solutions dialogiques s'est fortement réduite.

2.5.3 Communauté autour des *chatbots*

Nonobstant le scepticisme entourant cette technologie, de nombreuses communautés se sont organisées. Nous pouvons citer *PandoraBots*, qui au départ était une plateforme recensant différents *chatbots* et possédant un forum de discussion. Elle est devenue une société proposant des *chatbots* dans un langage de programmation spécifique à cette tâche, AIML [Wallace, 2004]. *Botlist* [27], un site récent, propose une liste de *chatbots* disponibles sur les réseaux sociaux. *Chatbots Magazine* [28], est un site très actif autour de la technologie, qui regroupe à la fois des articles sur la question, mais aussi des tutoriels et des études. Enfin *Chatbots.org* [29] reste la référence en terme de recensement d'agents conversationnels, qu'ils soient encore en usage ou obsolètes. Certains sont même testables sur la plate-forme. En ce qui concerne les *chatbots* dans l'industrie, nous constatons que les plus grands groupes (GAFAM + IBM) ne parviennent pas encore à développer des systèmes réellement satisfaisants, mais que la communauté ainsi que les développements publics autour des *chatbots* a beaucoup progressé ces dernières années, ce qui montre l'engouement général pour la technologie.

27. https ://botlist.co/
28. https ://chatbotsmagazine.com/
29. https ://chatbots.org/

2.6 Conclusion

Dans ce chapitre, nous avons proposé une nomenclature sur les différentes façon d'appeler un système de dialogue, ainsi qu'un état de l'art détaillé de la discipline, de son évolution, et de ses dernières avancées. Nous avons expliqué que, d'après la littérature, l'étude des différents modules d'un système de dialogue est peu à peu délaissée au profit des systèmes de bout-en-bout, d'autant plus depuis l'émergence des transformeurs. Nous avons également décrit plusieurs systèmes, notamment les memory networks, qui s'intéressent aux modèles cognitifs et à la représentation de soi et de l'utilisateur. Nous avons également expliquer le gouffre existant encore aujourd'hui entre les prouesses scientifiques et les réalisations industrielles, que nous pouvons attribuer à un manque de transfert technologique ainsi qu'à une absence, en contexte scientifique, de données représentatives de cas d'usage réels, mais attribuable également à une impossibilité de passage à l'échelle, en particulier dans le cas des derniers modèles de transformeurs. Dans nos travaux, nous essayons à la fois de proposer une solution basée sur les modèles mnémoniques qui soit industrialisable et pouvant concurrencer les transformeurs.

Chapitre 3

La mémoire dans le dialogue

> There are only two possible explanations : either no one told me, or no one knows
>
> Néo

Dans ce chapitre, nous nous intéressons aux modèles cognitifs de mémoire humaine ainsi qu'à la place de la mémoire pendant un dialogue et des différentes utilisations de celle-ci selon que l'interaction est orale ou écrite, humaine ou humain-machine.
Il est fréquent d'entendre ou de lire que l'être humain possède deux types de mémoire : la mémoire à long terme et la mémoire à court terme. Cela est partiellement vrai. [Atkinson and Shiffrin, 1968] définissent la mémoire comme "la faculté d'un système intelligent à enregistrer, conserver, et rappeler les expériences passées pour interpréter les expériences présentes". Leurs recherches leur ont permis de déterminer qu'il existait trois types de mémoire : la mémoire sensorielle [Klatzky, 1980], la mémoire à court terme [Sperling, 1967] et la mémoire à long terme [Rudner and Rönnberg, 2008]. [Squire and Zola, 1996] raffine ce modèle de mémoire et explique que la mémoire à long terme se décompose en mémoire déclarative (ou volontaire) et mémoire non déclarative (ou procédurale) [Miller, 1962] . La différence principale entre les mémoires à court et long terme s'exprime par le temps pendant lequel les informations ainsi que leur nombre maximal sont enregistrées dans le système cérébral. Alors que dans les mémoires à court et très court terme les informations sont oubliées au bout de quelques secondes à une minute et que la moyenne d'informations retenues varie entre sept et douze objets, les mémoires à long terme stockent les informations ad vitam eternam et en quantité illimitée [Cowan, 2008]. La différence entre les deux mémoires a été constatée lors d'études de cas pathologiques où les patients se mettent à oublier des informations de leur passé, ou au contraire ne sont pas capables de se souvenir du début de leur phrase [Gordon et al., 2001]. Au sein des mémoires à long terme, les mémoires déclaratives

décrivent le stockage et la restitution d'informations permettant le langage et la pensée, quand la mémoire procédurale peut se définir comme étant la mémoire des choses, des actions. Une étude [Owen et al., 1992] sur des cas de Parkinson et de Alzheimer a montré que la partie du cerveau endommagé lors de la maladie de Parkinson était le striatum, responsable de nos mouvements, alors que pour la maladie d'Alzheimer cet organe demeurait intact, mais c'était une atrophie du cortex temporal, responsable entre autres du langage, qui pouvait anticiper son apparition. [Greenberg and Verfaellie, 2010] parle de mémoire épisodique, qui se rapporte aux souvenirs autobiographiques ; et de mémoire sémantique, sorte d'encyclopédie des connaissances. [Ebbinghaus, 2013] décrit le concept de sauvegarde qui permet d'enregistrer dans la mémoire à long terme des informations.

Nous nous permettons une dernière remarque lexicologique sur la mémoire. Il est d'usage commun d'utiliser le verbe "mémoriser" ou son substantif "mémorisation" comme synonyme de "retenir" comme dans "Vous pouvez mémoriser le quai de votre bateau comme repère de balisage avant de naviguer". Cependant, il manque dans cette synonymie l'aspect de restitution de la fonction mémorielle. En effet, la rétention va de mise avec le rappel. Nous serions tentés de considérer que mémoriser c'est à la fois retenir et à la fois se souvenir.
Observons maintenant le fonctionnement de ces différentes mémoires.

3.1 Le fonctionnement de la mémoire

Pour comprendre le fonctionnement de la mémoire humaine, nous devons garder en tête trois concepts : la perception, l'action et la réaction. La perception correspond à l'arrivée dans nos capteurs de données brutes, l'action correspond à notre agissement au moment de cette perception, et la réaction (ou réponse) correspond à la génération d'une action décidée par notre système central exécutif [Baddeley, 1996] suite à l'analyse du couple perception-action.
Lorsque nous percevons quelque chose, par exemple quelqu'un qui nous salue, nos capteurs sensoriels sont activés et envoient dans la mémoire de saillance (aussi appelée mémoire sensorielle) le message reçu (le message lui-même, le son de la voix émettrice, l'origine visuelle de l'émission...) ainsi que l'action effectuée par nous au moment de la réception du message (marcher dans la rue, penser à quelqu'un, observer devant soi un visage familier...). . Le message, en moins d'une seconde, est transformé en informations qui sont encodées, stockées et interprétées par la mémoire de travail, pendant quelques secondes, le temps pour l'information d'être enregistrée puis restituée dans la mémoire à long terme. Ce sont les mécanismes de rétention et de rappel [Roediger, 1990, Bodner and Lindsay, 2003]. Lorsque les informations ont été encodées dans la mémoire à long terme, la mémoire de travail, tel un processeur informatique sert d'outil de décision sur quelles informations res-

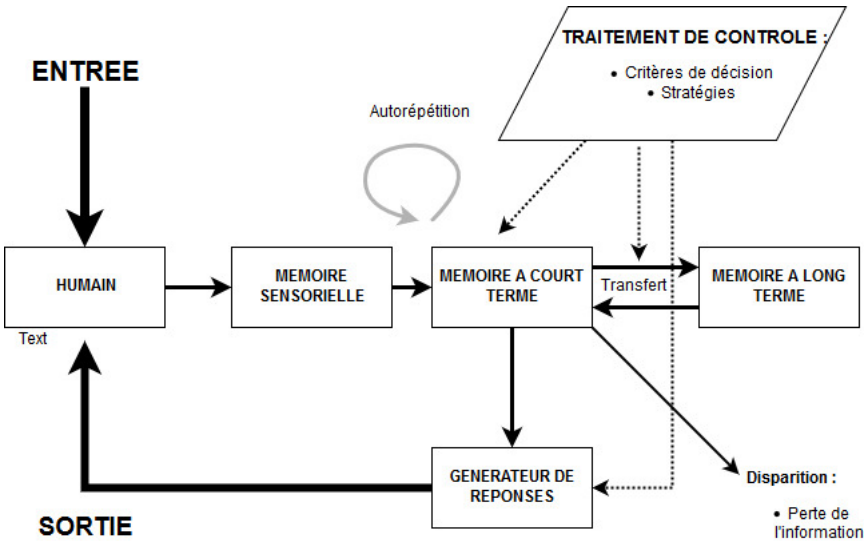

FIGURE 3.1 – Modèle de mémoire d'après [Atkinson and Shiffrin, 1968]. La **mémoire de travail**, au centre du fonctionnement cognitif, est constitué dans le schéma, de la mémoire dite à court terme et de l'appareil de traitement de contrôle également nommé central exécutif. La mémoire à long terme, dans le schéma, correspond aux mémoires **épisodique** et **sémantique**

tituer depuis, soit la mémoire épisodique (souvenir d'une réaction à avoir dans une situation (couple perception-action) similaire), soit la mémoire sémantique (savoir profondément ancré en soi, presque comme une vérité absolue, sur quelle réaction avoir lors de cette situation). Les informations sont alors renvoyées dans la mémoire de travail qui réalise un tri entre celles à garder et celles à oublier. La mémoire de travail sert également de système central exécutif qui décide également de la réaction (ou réponse) à générer [McCabe et al., 2010] . La figure 3.1 décrit ce schéma. Pour illustrer le fonctionnement des différentes mémoires nous prenons l'exemple de l'interprète.

3.1.1 Les mémoires à court terme

La mémoire à court terme est un terme générique qui regroupe d'une part la mémoire sensorielle (ou mémoire perceptive) et d'autre part la mémoire de travail [Baddeley, 2010] pour traiter les informations à court terme avec sa boucle phonologique [Baddeley et al., 1998] pour recevoir les sons et les interpréter, et son calepin visio-spatial [Baddeley, 1995] pour exécuter une représentation mentale des informations.

La mémoire sensorielle

Lorsque l'interprète traduit, il enregistre ce qu'il entend. Cet enregistrement s'effectue en moins d'une seconde et dure également moins d'une seconde. Il s'agit

d'un mécanisme de détection immédiate des éléments importants dans la ou les phrases entendues par l'interprète. La mémoire sensorielle envoie ensuite les enregistrements à la mémoire de travail puis elle se vide, prête à recevoir la prochaine perception auditive ou visuelle [Sams et al., 1993].

La mémoire de travail

Cette mémoire ne dure que quelques secondes le temps de visualiser les concepts associés aux enregistrements reçus par la mémoire sensorielle (ce qui ne permet de conserver en mémoire que quelques mots, entre sept et douze selon des expériences [1]). Elle est appelée mémoire de travail car c'est celle-ci qui possède le rôle central de contrôle des informations perçues et renvoyées. C'est cette mémoire de travail qui nous permet de retenir la commande d'un client, un numéro de téléphone, un code postal... Suffisamment longtemps pour que l'on effectue une action, avant d'être oubliées. Les mots reçus par l'interprète sont ensuite intégrés à la boucle phonologique qui est un système spécialisé dans le stockage et le traitement des informations verbales et symboliques : mots, chiffres, lettres, syllabes, etc. Elle est impliquée dans la lecture, l'écriture, la compréhension orale, et dans le calcul mental [Engle, 2002]. Composée d'un système de stockage et d'un système de répétition. Cette répétition mentale est celle qui permet à l'interprète de retenir les mots entendus, le temps d'aller chercher dans...

3.1.2 Les mémoires à long terme

...les mémoires déclaratives à long terme, les informations manquantes pour la bonne interprétation des mots, par exemple dans le cas des mots homonymes ou polysémiques, en fonction des contextes dans lesquels l'interprète, par le passé, a entendu ces mots et les connaissances sémantiques qu'il possède de ces mots. C'est aussi dans les mémoires à long terme que l'interprète va anticiper la fin de la phrase ou le début de la suivante qu'il n'a pas encore entendues, afin de se préparer à les traduire avant même de les avoir enregistrées et ainsi gagner du temps et de la place dans ses mémoires de travail et de perception.

La mémoire épisodique

La mémoire sémantique

Tout comme l'épisodique, la mémoire sémantique est pérenne, c'est là où sont stockés notre connaissance du monde, nos savoirs et nos acquis. C'est dans cette mémoire que sont stockés les triplets perception-action-réaction les plus fréquents et univoques : par exemple nous savons que mettre un doigt dans de l'eau bouillante

1. https ://www.cognifit.com/science/cognitive-skills/shortterm-memory

va nous faire mal. Pourtant, nous ne réfléchissons plus à un épisode de notre vie où cela nous est arrivé, nous le savons. La mémoire sémantique stocke les objets du monde devenus encyclopédiques pour nous. Pour l'interprète, c'est le lexique et le dictionnaire bilingue des mots et des variations syntaxiques de la langue qu'il écoute et de la langue dans laquelle il doit traduire.

Le système central exécutif

Longtemps associé à la mémoire de travail, il s'agit d'une fonction cognitive permettant le transfert d'informations de la boucle phonologique aux deux mémoires déclaratives jusqu'à ce que le central exécutif décide que les informations de la mémoire de travail ont été suffisamment enrichies par ces transferts. C'est également lui qui décide du lieu où doit être effectué le transfert : épisodique, sémantique, les deux ou aucun des deux. C'est enfin lui qui donne l'aval à la mémoire de travail pour générer une réponse, dans le cas de notre interprète, de générer une traduction. Un simple exemple de l'utilisation du transfert d'informations se trouve sur le site Neuropédagogie [2] :

"*Par exemple, essayez de vous souvenir d'un maximum de lettres ci-dessous en moins d'une minute. Au bout d'une minute, écrivez dans l'ordre les lettres sans plus voir le modèle. Top, c'est parti !*

QDFœIJHFGNMKENDIPIYDNLKNDPIYENPIDHGHBNEMIHDPIHHIMQDFDE

Vous ne devez pas avoir retenu beaucoup de lettres. Tentez maintenant l'expérience suivante. Nous avons autant de lettres dont il faut se souvenir en une minute que pour l'expérience précédente. Top, c'est parti !

FBICIANASANSACOCAFANTAPEPSIORANGINASONYIBMSAMSUNGSANYO

Ceux qui sont intéressés par les agences américaines, les boissons sucrées, les marques d'appareils électroniques (et qui se sont aperçus de l'astuce) auront sans doute noté plus de lettres que les autres :

FBI-CIA-NASA-NSA-COCA-FANTA-PEPSI-ORANGINA-SONY-IBM-SAMSUNG-SANYO

Et comme ces lettres sont classées par catégories, on s'en souvient mieux que si elles n'avaient pas été classées. De la même façon, en employant les chunks *afin de réunir des lettres en groupes ayant un sens, comme dans l'exemple précédent), on se souvient mieux d'un numéro de téléphone. Ainsi, le numéro de téléphone de l'Assemblée Nationale est 0140636000. On s'en souviendra mieux s'il est présenté ainsi : 01.40.63.60.00, encore mieux si l'on sait que l'indicatif de Paris est 01 (on a fait un chunk) : 40.63.60.00, encore mieux si on fait 406. 36000. Notons qu'au Royaume Uni les numéros de téléphones sont présentés sous forme de nombres à 3 chiffres, ce qui est beaucoup plus simple à retenir que les numéros de téléphone en France. En fait, il n'existe pas un empan mnésique, mais plusieurs, qui dépendent de*

2. https ://neuropedagogie.com/memoire-de-travail/capacite-de-la-memoire-de-travail.html

ce qu'il y a à mémoriser : chiffres, lettres, mots, pseudo-mots, séquences gestuelles ..."

La mémoire procédurale

C'est une mémoire dite implicite, non-déclarative, inconsciente. C'est l'enregistrement des épisodes de notre vie si répétés et similaires qu'ils sont devenus des automatismes. C'est la mémoire qui nous apprend à lire, écrire, faire du vélo. Son rôle dans l'apprentissage d'une langue n'est pas clair, tout comme son impact lors d'une conversation [Morgan-Short et al., 2014]. C'est également la mémoire qui survit le plus longtemps à l'altération des fonctions cognitives, qui résiste le mieux au vieillissement de l'humain. Cependant, c'est également la mémoire dans laquelle les informations mettent le plus de temps à être stockées, et pour lesquelles le rôle du sommeil est le plus important. En effet, pour que des actions deviennent si naturelles que nous les automatisions, il est nécessaire d'observer une période d'apprentissage plus ou moins longue selon nos capacités de rétention innées et notre âge. C'est pour cela qu'il est bien plus facile d'apprendre à nager en étant petit enfant, ou à faire du vélo "sans les mains" dès le plus jeune âge, qu'une fois l'âge adulte atteint [3]. De plus, le rôle du sommeil, déjà primordial pour les mémoires déclaratives [Ellenbogen et al., 2006], est indispensable au bon enregistrement des informations dans la mémoire procédurale.

3.1.3 Interférences

Après avoir résumé le fonctionnement des différentes mémoires, discutons désormais des dysfonctionnements. Outre les altérations de nos compétences mémorielles et cognitives liées au vieillissement ou à des pathologies [Small, 2002] voire à des facteurs extérieurs, il existe souvent des dysfonctionnements dans notre système mémoriel, appelés interférences que tout un chacun a déjà vécu. Cela peut se traduire par "Je l'ai sur le bout de la langue" en essayant de retrouver le nom ou le visage d'une personne. Ou encore, "Zut, j'ai oublié ce que je voulais te dire". Dans le premier cas, il s'agit de l'interférence rétroactive, dans le second, l'interférence pro-active.

Interférence rétroactive

C'est une interférence où les nouveaux souvenirs perturbent les anciens, soit car ils sont en cours de traitement (avant le sommeil par exemple) soit car ils sont très anciens. Lorsque l'on entend une voix, que l'on voit un visage, que l'on goûte un plat, et qu'on désire mettre un nom sur ces perceptions, la similitude entre plusieurs "candidats" possibles à être la bonne réponse, ou l'éloignement

3. Pour ce cas-ci, nous parlons en douloureuse connaissance de cause

temporel entre les perceptions et les informations correspondantes ancrées dans les mémoires à long terme provoque une latence voire une incapacité à fournir LA bonne réponse à notre questionnement. C'est l'effet "sur le bout de la langue". Le cerveau, lorsque le bon souvenir a été récupéré et transféré dans la mémoire de travail, active une fonction de récompense dans notre système central exécutif qui se traduit par un sentiment instantané de soulagement [Baddeley and Dale, 1966]. A l'inverse, si nous "abandonnons" notre recherche interne de la bonne réponse, s'active une punition sous forme de frustration.

Interférence pro-active

C'est l'inverse, où des souvenirs en cours de traitement viennent empêcher l'enregistrement d'informations nouvelles. [Szpunar et al., 2008] ont montré que lorsque l'on fait retenir à des volontaires plusieurs suites de mots, ils ont plus de mal à se rappeler les mots de la deuxième suite par rapport à la première. C'est également le cas lorsque l'on tente de retenir plusieurs actions, par exemple passer un coup de téléphone et mettre en marche la machine à laver, on a tendance à retenir la première des deux actions puis oublier la seconde action, tout en se souvenant que l'on avait bel et bien une seconde action à effectuer. La système de récompense-soulagement est néanmoins similaire à celui des interférences rétroactives.

3.2 La mémoire dans le dialogue

Lors d'une conversation humain-humain, [Ciaramelli et al., 2013, Denny et al., 2014, Brod et al., 2016] ont défini la capacité des personnes à dialoguer selon trois dimensions : connaissance, expérience et empathie, qu'ils regroupent dans la notion de mémoire. [Bangerter, 2004] parle d'attention jointe pour expliquer la focalisation de tous les interlocuteurs sur la même entité conversationnelle. Cette attention se porte sur le mot dans un dialogue. [Cintrón-Valentín and Ellis, 2016] la décrivent comme la saillance linguistique.
Dans le modèle de dialogue de [Clark and Marshall, 1981], on retrouve cinq concepts : l'environnement, le tour de parole, l'historique du dialogue, l'expérience des interlocuteurs et les connaissances des interlocuteurs. Or si l'on compare avec le modèle de mémoire humaine, on retrouve ces cinq composantes principales : la mémoire de saillance (tour de parole, très court, compréhension rapide et très éphémère), la mémoire de travail (historique du dialogue, la mémoire à court terme conserve les informations principales du dialogue en cours pour assurer sa cohérence), la mémoire épisodique (l'expérience des conversations passées pour permettre aux interlocuteurs de converser), la mémoire sémantique (la connaissance, la culture de l'interlocuteur) et la mémoire procédurale (l'environnement, les ac-

tions réalisées de façon presque automatique, inconsciemment).
[Poudade, 2006] a décrit un modèle de perception-action, où il définit la mémoire humaine comme une mémoire associative par le contenu. Il a montré l'impact de la taille de la mémoire de travail sur la complexité des représentations lexicales apprises par des agents lorsqu'ils jouent à des jeux de langage [Steels, 2000].
Un autre modèle utilisé pour décrire la mémoire dans le dialogue est le croyance-désir-intention (*Belief-Desire-Intention* BDI) [E. Bratman, 1987] qui explique que lors d'un tour de parole, l'interaction est guidée par la croyance de l'interlocuteur, son désir (le but final, ce qu'il n'a pas encore) et l'intention (la façon dont il parvient à son but). Ce modèle est adapté pour représenter la réalité d'une conversation, comme le montrent les réalisations de [Jokinen, 2018, Visser et al., 2015, Wobcke et al., 2005, Wadsley and Ryan, 2013].
Ce parallèle évoqué, *dans quelle mesure l'intégration de ce modèle de mémoire dans un système de dialogue peut-il améliorer les performances de ce dernier ?*

3.2.1 Mémoire parfaite

Il existe un lieu commun dans l'imaginaire collectif décrivant la capacité de certains individus à se souvenir d'une grande quantité d'informations observées dans les moindres détails, comme par exemple des listes entières de numéros de téléphones, des recueils entiers de poèmes, ou à se se souvenir, plusieurs secondes après son observation, d'un tableau, un paysage..permettant une reconstruction de ceux-ci presque à l'identique. La première capacité est souvent appelée mémoire photographique, la seconde, mémoire éidétique. Parfois, certains auteurs les considèrent interchangeables [la Brecque, 1972]. Des études ont montré que c'était certes une capacité existante chez certains enfants en bas-âge, mais il n'a jamais été prouvé qu'elle persistait à l'âge adulte malgré certains témoignages. Nous savons que la mémoire à long terme possède une capacité illimitée à la fois en espace et en temps. Cela signifierait que chez certains individus, la mémoire de travail posséderait également une capacité illimitée. Nous savons également que les transferts d'information (encodage et restitution) entre les mémoires de travail et à long terme ne s'effectuent pas à partir des neurones mais à partir des synapses connectant les neurones, provoquant parfois les interférences liées à des facteurs internes (âge, fatigue...) ou externes (bruit...). Cela signifierait que certains individus seraient capables de façon innée d'ignorer la totalité des interférences du monde et de restituer théoriquement l'intégralité de leur mémoire à long terme dans leur mémoire de travail. A notre connaissance, cela demeure encore de l'ordre du mythe. Il existe, il est vrai, une pathologie cognitive appelée hypermnésie forçant l'individu atteint par celle-ci à se remémorer en détails des épisodes très lointains dans son passé, à être perturbé par ces mêmes souvenirs au point d'être affecté socialement, en restant "coincé" dans son passé. Il a été démontré que les personnes atteintes d'hypermnésie n'avaient pas une mémoire de travail supérieure aux autres, que leur

capacité de rétention et de rappel ainsi que leur quotient intellectuel n'avait rien d'exceptionnel [Payne, 1987]. Cependant, il est difficile de considérer que le fait de se rappeler en détails des chaussettes mises le lendemain de notre quatrième anniversaire ainsi que du yaourt tombé par terre lors de ce même anniversaire constituent une mémoire parfaite. A l'inverse, les mnémonistes sont des personnes qui, par des techniques et beaucoup de travail [Luria, 1987], ont augmenté la capacité de leur mémoire de travail, au point d'être capables de mémoriser de grandes quantités d'information dans leur mémoire de travail. Cependant, cela demeure une capacité acquise, qui améliore la puissance normale de la mémoire de travail, sans la rendre illimitée pour autant. En définitive, il n'a pas encore été prouvé chez l'humain, l'existence d'une mémoire parfaite.

3.3 Les différents canaux dialogiques

Nous ne pouvons conclure ce chapitre sans évoquer les configurations dialogiques qui conditionnent la mémorisation (rétention et restitution, rappelons-le) des informations. En effet, pour bien comprendre le fonctionnement de la mémoire cognitive, et afin de construire un modèle de système de dialogue intégrant ces composantes, il nous semble important de décrire les différences linguistiques et neuropsychologiques entre les dialogues écrits et oraux, ainsi qu'entre les dialogues humains et humain-machine.

3.3.1 Du dialogue oral et écrit

[Rubin, 1978] décrit une taxonomie des six[4] principales différences entre le dialogue oral et écrit.

— **La Modalité** est la dimension audiovisuelle. C'est la différence la plus évidente mais également la plus étudiée. Là où la conversation orale permet des différences d'intonation, d'emphase, la conversation écrite se distingue par des manifestations visibles, comme des mots écrits en capitale, ou en gras, italique... Cependant, un même message ne sera pas mémorisé de la même façon et au même endroit dans la mémoire épisodique selon s'il est oral ou écrit. L'oral, par exemple, ne permet pas de visualiser une virgule, qui désambiguïse certaines propositions. Par exemple, "*Je vais manger, Paul*" et "*Je vais manger Paul*" n'ont plus le même sens en fonction de la présence ou pas d'une virgule. A l'inverse, l'écrit ne permet pas d'entendre la prosodie d'une phrase pour supprimer l'équivoque de celle-ci : "*Paul a trouvé Jean puis Marie l'a trouvé*", selon que l'intonation est mise sur Marie ou sur le pronom personnel, cela peut signifier, dans le premier cas, que

4. En realité ce sont sept et non six différences, mais la dernière, à savoir la séparabilité des discours ne s'applique pas car elle est évidente dans le cas d'une conversation

Marie a trouvé Jean, alors que dans le second cas, Marie a trouvé Paul[5].

— **L'interaction** est la dimension réflexive : comment chaque interlocuteur a-t-il conscience de l'autre ? L'état de croyance est bien différent si chacun voit et entend l'autre, si chacun entend uniquement l'autre, ou si chacun ne voit que la production écrite de l'autre.

— **L'engagement** est la dimension de connaissance de l'autre. Bien que proche, elle diffère de l'interaction en ceci qu'elle représente la connaissance que possède chaque interlocuteur de l'autre, c'est-à-dire l'importance qu'auront les pronoms personnels de deuxième personne sur la conversation. Ce n'est pas une dimension essentielle au niveau de la différence entre dialogue oral et écrit.

— **L'espace** est la dimension spatiale[6] . Les interlocuteurs partagent-ils le même espace physique, ou bien visualisent-ils l'espace de l'autre ? Quel impact cela possède-t-il sur la conversation ? La nécessité d'expliciter à l'écrit, à la fois l'espace visuel mais aussi auditif (son à décrire par exemple) change également la représentation cognitive que se fait chaque locuteur : celui qui écoute se représente la description faite par celui qui parle, et celui qui parle se représente la représentation de celui qui écoute.

— **La temporalité**. Il est difficile de citer l'espace sans évoquer la chronologie. Par exemple, dans la phrase "Je vous ai demandé hier de me rappeler demain !", *demain* s'agit-il du lendemain du jour de la demande, à savoir aujourd'hui, ou bien du lendemain du jour de la discussion, à savoir demain ? A l'oral, il serait possible de mettre l'emphase sur *hier* ou bien sur *demain* pour désambiguïser cette temporalité. A l'écrit, il est nécessaire d'avoir accès à un certain contexte.

— **La concrétion des référents** rejoint les dimensions spatio-temporelles. La référence à des objets du monde ne sera jamais la même selon que les interlocuteurs partagent un même espace ("la table devant nous"), ou non. De plus, ce mot "table", prononcé sans contexte, n'évoquera pas le même objet selon s'il est entendu (et donc prononcé) ou lu (donc écrit). Dans le premier cas, il évoquerait tantôt la table de classe ("mains sur la table") que la table de déjeuner ("nous passons à table"). Dans le second cas, il évoquerait plutôt des tables de multiplication ou la Table Ronde. Cela dépendra de la mémoire épisodique et sémantique de chaque interlocuteur, à savoir la représentation la plus saillante pour lui lorsqu'il perçoit le couple

5. Dans l'article original, d'autres exemples en anglais sont trouvables page 7

6. Nous espérons que le lecteur sera reconnaissant envers l'auteur de cette thèse pour cet éclairage, presque aussi important que celui de son chat qui découvrit que l'eau mouillait

signifiant-signifié du mot “table”.

[Sykes, 2005] ont montré dans une étude que lors de l'apprentissage d'une langue, des stratégies de production linguistique diffèrent selon si l'interaction est orale ou écrite, en face à face ou à distance. En effet, alors que pour les conversations orales, il est fréquent de voir des contextes plus élaborés et stratégiques pour corroborer les actes capitaux [Blum-Kulka et al., 1989, García, 1993], pour les conversations en face à face il n'est pas rare d'observer une plus grande diversité de mouvements de support des actes capitaux. Cependant, les conversations écrites, manquant de contexte visio-spatial et d'intonations, s'approprient les deux stratégies en utilisant à la fois des contextes et des mouvements de support complexes. D'autre part, alors qu'il est commun d'exprimer le refus ou le déni par le rire lors d'une conversation orale ou en face à face, ils en concluent que les apprenants à l'écrit produisent des expressions écrites plus complexes permettant de transcrire le refus ou le déni, conclusion corroborée par [Tang, 2020] . Dans la même étude, ils expliquent qu'une autre différence entre la conversation orale et écrite réside, pour la seconde, dans la présence d'un historique de dialogue, permettant aux interlocuteurs d'aller chercher des informations exactement comme elles ont été exprimées plus tôt dans la conversation, contrairement à la conversation orale, où une trace est conservée uniquement dans la mémoire de travail de chacun des locuteurs. Toutefois, pour la tâche d'apprentissage d'une langue seconde, malgré ces apparents avantages de la conversation écrite pour l'assimilation de celle-ci, [Tang, 2019] montrent que les conversations en face à face restent les plus efficaces pour l'acquisition de la compétence langagière.

3.3.2 Du dialogue humain-humain (DHH) et humain-machine (DHM)

Nous avons évoqué en 2.2.2 que le dialogue était une co-construction. Nous pouvons également parler de coopération des participants. Cette coopération est définie par [ALLWOOD et al., 2000] comme l'association de la considération cognitive, le but joint, la considération éthique et la confiance de chaque interlocuteur en l'autre. Cette coopération implique une certaine symétrie conversationnelle entre les participants, symétrie qui peut être altérée par des facteurs endogènes (participant malentendant, fatigué, à la voix cassée...) ou exogènes (bruit environnant, mauvaise connexion...). Il se crée alors une asymétrie de partenaire de dialogue [Bernsen et al., 1996]. Le dialogue humain-machine est un exemple de dialogue à partenaire asymétrique. Dès lors, la coopération entre l'humain et la machine ne peut se faire que si l'humain accepte le fait qu'il parle à une machine, et donc adapte son discours de façon à produire une expression linguistique permettant de faire avancer la conversation. Dans son livre *Talker Quality in Human and Machine Interaction*, [Weiss, 2020] évoque les différences d'analyse du dialogue entre hu-

mains et du dialogue humain-machine. Il explique d'une part que la machine sera toujours inférieure à l'humain en termes de compréhension sociale et contextuelle de l'information, mais au contraire possédera un accès unique à certaines données comme par exemple une localisation GPS. Il ajoute également que certains critères permettent d'évaluer à la fois le DHH et le DHM. En effet, la *structure de surface* (durée des tours de parole, nombre d'interaction, nombre de mots par locution...) les *phénomènes conversationnels* (changement des tours de parole, changement du ton acoustique ou linguistique) et les *évaluations et intentions* (actes de dialogue, réponses appropriées, évaluation de la coopérativité de l'utilisateur) sont autant de critères applicables au DHH comme au DHM. A l'inverse, la latence de réponse du système, les incohérences de dialogue (voir Chapitre 4), la qualité de la reconnaissance vocale sont autant de facettes qui sont exclusives au DHM.

3.4 Conclusion

En définitive, ce qu'il faut retenir de ce chapitre c'est que le fonctionnement mnémonique humain, basé sur cinq mémoires allant du sensoriel au procédural sont essentielles à l'apprentissage de l'exercice dialogique. Les mécanismes de transfert d'information si sophistiqués et spécifiques de l'humain sont-ils imitables chez la machine ? Peut-on apprendre à la machine à dialoguer tel que le ferait un humain, en lui faisant apprendre un modèle d'elle-même en train de dialoguer ? Si oui, quelle architecture conviendrait à ce modèle, et comment évaluer sa nécessité et ses performances ? Dans la suite de nos travaux, nous allons démontrer que l'ajout explicite de fonctionnalités mémorielles simulant à la fois un modèle de l'utilisateur et un modèle du système vu par l'utilisateur permet d'améliorer la cohérence du système lors de l'interaction.

Chapitre 4

Le modèle à trois axes

- Why do my eyes hurt ?
- Because you never used them before

Néo

Dans ce chapitre, nous traitons la problématique temporelle du dialogue orienté-tâche par le prisme des modèles mnémoniques humains. Comme suggéré dans le chapitre 3, la mémoire est au cœur de la fonction dialogique humaine, dans la mesure où elle permet à la fois la perception des informations transmises par le locuteur, la rétention de celles-ci le temps de récupérer les informations permettant de générer une réponse et l'encodage des informations pertinentes de façon pérenne afin de constamment améliorer sa capacité à converser.
Ce qui nous intéresse ici est de montrer la nécessité pour un système de dialogue de posséder un appareil mnémonique et de décrire un modèle permettant de simuler celui-ci. Dans ce chapitre, lorsque nous parlons de dialogue sans autre spécification, il s'agit de dialogue humain-machine orienté-tâche.

4.1 Motivation

Notre approche peut s'expliquer à partir de trois hypothèses. A travers l'étude des systèmes récents de dialogue, nous déterminons tout d'abord la présence d'une représentation de l'utilisateur dans les modèles, puis celle d'une temporalité dans l'apprentissage, et enfin la détection des incohérences.

4.1.1 La dimension réflexive

En effet, comme expliqué en 2, les modèles les plus performants de système de dialogue étaient représentés, jusqu'à la fin des années 2000, par un problème

de POMDP, c'est-à-dire un processus de décision markovien dans lequel l'état courant n'est pas connu par l'agent, mais au contraire il se trouve dans l'état de croyance (*belief state*) de l'agent. Une des limites de ce système était que le nombre d'états courants possibles est discret, alors que l'état de croyance, lui, est théoriquement continu. Cependant, la force du POMDP se résume au fait que pour un tour de parole donné de l'utilisateur, et d'un contexte (historique de dialogue), le système était capable d'apprendre, en fonction d'une récompense, la meilleure action à entreprendre et donc la meilleure réponse à fournir. Si l'on observe les différentes architectures récentes de système de dialogue, elles ont certes amélioré et complexifié les réponses du système par rapport au POMDP, mais elles n'ont pas fait évoluer le principe d'état de croyance lié aux observables courants et à l'historique de dialogue. Rien n'indique que les modèles de système dialogue, qu'ils soient construits à partir des auto-encodeurs variables, des antagonistes, des réseaux récurrents des transformeurs ou des réseaux de mémoire, fassent émerger, durant leur apprentissage, un modèle réflexif de soi, autrement dit, un modèle du système lui-même en train de se mettre à la place de l'utilisateur, donc en train de se représenter l'utilisateur en train de se mettre à la place du système lui-même à l'image du modèle COSMO [Barnaud et al., 2018]. Dans la mesure où lors de l'apprentissage, le système a non seulement accès à l'historique du dialogue qu'il est en train d'apprendre, mais aussi aux dialogues qu'il a déjà observés, afin de pouvoir décider, pour le tour de parole courant, de la meilleure action à mener, on peut émettre l'hypothèse que quelque part dans son apprentissage il existe une dimension réflexive représentant la projection que fait le système de l'état de croyance de l'utilisateur. Toutefois, puisque cette dimension n'est pas clairement explicitée, nous ne pouvons pas être certains de son existence. Une de nos contributions consiste à proposer un modèle explicitant cette dimension.

4.1.2 La dimension temporelle

Dans le dialogue, il existe une temporalité relative : chaque locuteur, sans pour autant pouvoir l'expliquer, sait à quel instant du dialogue il se trouve. Par instant, il faut entendre sous-dialogue tel que défini par [Bennacef et al., 1996] comme les symboles non-terminaux de la grammaire du dialogue, par opposition aux actes de dialogues qui sont les symboles terminaux. Plus précisément, les sous-dialogues contiennent, entre autres, les sujets (ou topiques) du dialogue tels que :

- les sous-dialogues d'ouverture et fermeture du dialogue.
- les sous-dialogues d'affirmation ou de continuation (lorsque l'utilisateur ou le système veulent passer à l'étape suivante, par exemple le choix d'un hôtel, ou la réservation d'un billet de train.
- les sous-dialogues de confirmation ou de vérification (lorsque le système s'assure qu'il a compris les intentions de l'utilisateur, souvent par le biais

d'un récapitulatif en langue naturelle des informations qu'il a récoltées, ou lorsque l'utilisateur s'assure que le système a fait ce qu'il devait faire, ou qu'il a besoin d'un rafraîchissent des informations).
— les sous-dialogues correctifs où l'utilisateur, peut, ou pas, corriger le système lorsque celui-ci se trompe dans sa tâche ou répond quelque chose d'incohérent.

Un sous-dialogue contient généralement un ou plusieurs actes de dialogue. L'utilisateur, en fonction du sous-dialogue courant, et de la longueur actuelle du dialogue, infère son positionnement, c'est-à-dire qu'il projette la durée restante du dialogue par rapport au sous-dialogue courant et à l'historique de la conversation. Cette projection lui permet de prévoir quelle sera la suite probable du dialogue, quelles questions sont susceptibles d'être posées, et quand le dialogue va se conclure [Jansen et al., 2013]. Or, cette dimension temporelle, présente chez l'utilisateur, est-elle également présente chez le système ? Y a-t-il, lors de l'apprentissage du modèle, une dimension temporelle qui émerge, permettant au système de mieux cerner le positionnement du sous-dialogue courant dans la temporalité du dialogue, et cette dimension permet-elle au système d'améliorer la qualité de sa réponse à l'utilisateur ? Nous contribuons à répondre à cette question en explicitant la dimension temporelle du dialogue lors de l'apprentissage, en identifiant des parties, ou tranches temporelles, clés du dialogue, aidant le système à se repérer et à mieux préparer sa réponse à l'utilisateur.

4.1.3 La dimension des incohérences

Cette dimension est corrélée avec la réflexivité et la projection sur le temps restant dans une dialogue. Cependant, la nature de la relation (cause, effet) n'est pas claire. La dimension des incohérences doit permettre de relier la réflexivité à la temporalité et de faire émerger la nature de ces relations.

4.2 Les méthodes de modélisation du dialogue

4.2.1 Les réseaux génératifs antagonistes

Il nous est apparu logique de penser qu'il fallait apprendre un modèle faisant émerger une certaine représentation de soi, une représentation de la temporalité du dialogue, et une gestion de la taille des informations retenues en mémoire. La première question que nous nous sommes posés est le type d'architecture la plus adaptée à l'intégration de ces représentations.
Pour illustrer le fonctionnement des différentes mémoires nous prenons l'exemple de l'interprète. Nous nous sommes tout d'abord intéressés aux réseaux génératifs antagonistes (GAN) [Goodfellow et al., 2020]. Un réseau génératif antagoniste 4.1

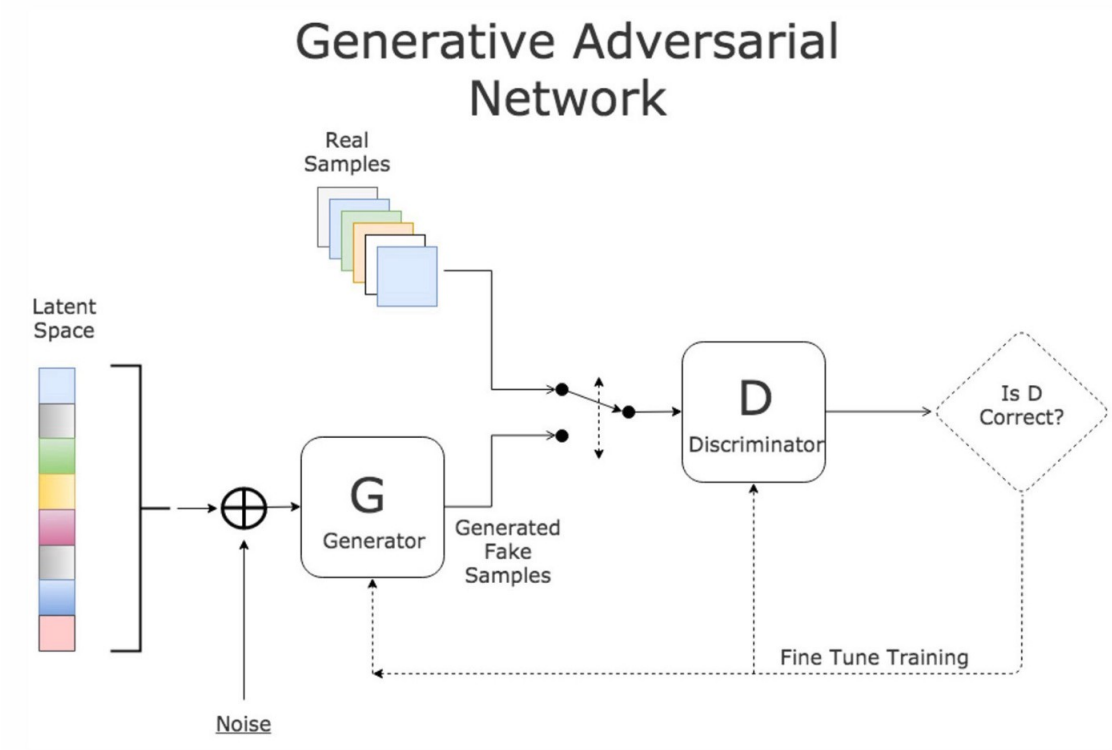

FIGURE 4.1 – Réseau génératif antagoniste de [Goodfellow et al., 2020]

se compose de deux éléments, un générateur et un discriminant. Le premier s'entraîne à créer de faux objets à partir d'un jeu de données et d'un vecteur de bruit, le second s'entraîne à différencier les vrais objets de ceux créés par le générateur. Lorsque chacun est appris, on met le discriminant au sommet du générateur. Les poids du discriminant sont alors figés, et le générateur doit tromper le discriminant en générant des faux objets les plus proches possibles des vrais. Le modèle est considéré comme appris lorsque le discriminant se trompe une fois sur deux, autrement dit, les faux objets ressemblent tellement aux vrais, que la décision du discriminant ressemble à un lancer de pièce de monnaie en l'air. C'est une technique d'apprentissage très populaire pour la création de fausses images, notamment dans des applications grand public récentes, telles que *Deepfake* [Nguyen et al., 2019] avec des résultats aussi impressionnants que dangereux [Westerlund, 2019]. Toutefois, pour la génération textuelle, le principal problème repose dans le fait que le générateur doit fonctionner avec des valeurs continues alors que dans une phrase, les valeurs sont discrètes (lemmes, caractères..), il n'est alors pas possible de modifier légèrement un mot ou une lettre, sans modifier sa sémantique. Ainsi, [Nie et al., 2019] ont mis au point une architecture de GAN permettant de transformer les séquences de phrases en valeurs continues afin de pouvoir effectuer la génération. Plus récemment, [Croce et al., 2020] ont combiné le transformeur BERT avec un GAN afin de réduire le nombre de données annotées nécessaires pour l'appren-

tissage et obtenir tout de même des résultats de génération comparables à l'état de l'art. Pour la génération de réponse dans un système de dialogue, [Su et al., 2018a] ont montré qu'il était en effet possible de générer des réponses pertinentes à des énoncés de l'utilisateur, bien que cela fonctionne mieux pour les dialogues ouverts que les dialogues orientés-tâche. Cependant, trois limitations apparaissent aux GAN pour les systèmes de dialogues :

— La convergence prend beaucoup plus de temps que pour un modèle basé sur les LSTM.
— Le modèle, tout comme celui basé sur des LSTM, a du mal à différencier les informations venant de l'historique de dialogue et celles de la base de connaissances externe.
— Lorsqu'un dialogue est multi-domaine[1], le GAN ne converge presque jamais.

Ces limitations se traduisent par des performances n'atteignant pas l'état de l'art sur les tâches de pistage d'état du dialogue et de génération de réponse, pour un temps d'apprentissage significativement plus long [Olabiyi et al., 2019].

4.2.2 Les réseaux de mémoire (MemNN pour *memory neural network*)

Nous avons déjà évoqué la limitation des LSTM en 2.2.2 pour la résolution de la tâche de dialogue de bout-en-bout. Nous venons de voir celles des GANs. Notre troisième piste architecturale a donc été les réseaux de mémoire (MemNN), dont la construction, à l'origine pensée pour les systèmes de question-réponse, a été adaptée par [Madotto et al., 2018], en s'inspirant des MemNN à porte [Liu and Perez, 2017] (eux-mêmes utilisant des réseaux routiers [Srivastava et al., 2015] pour optimiser la taille des couches du MemNN et donc la quantité d'information stockée à chaque saut de mémoire), pour générer des séquences et donc des réponses. A chaque saut, trois opérations sont effectuées : d'abord, l'historique de dialogue et l'énoncé de l'utilisateur courant sont concaténés puis encodés et les poids sont activés par une fonction $SOFTMAX$. Ensuite, ces poids sont multipliés par une copie l'historique de dialogue initial pour obtenir une sortie. Cette sortie est enfin additionnée à l'énoncé de l'utilisateur initial. L'opération est répétée k fois correspondant au nombre de sauts.
Le Mem2Seq [Madotto et al., 2018] est l'adaptation du MemNN pour la tâche de génération de réponse de dialogue. Plusieurs articles et didacticiels sur l'architecture détaillée des MemNN sont disponibles ici[2]. Pour le Mem2Seq 4.3, ici[3].

1. par exemple il traite un problème de billet train et de chambre d'hôtel
2. https ://www.tutorialexample.com/understand-end-to-end-memory-networks-part-1-a-simple-tutorial-for-nlp-beginners/
3. https ://github.com/HLTCHKUST/Mem2Seq

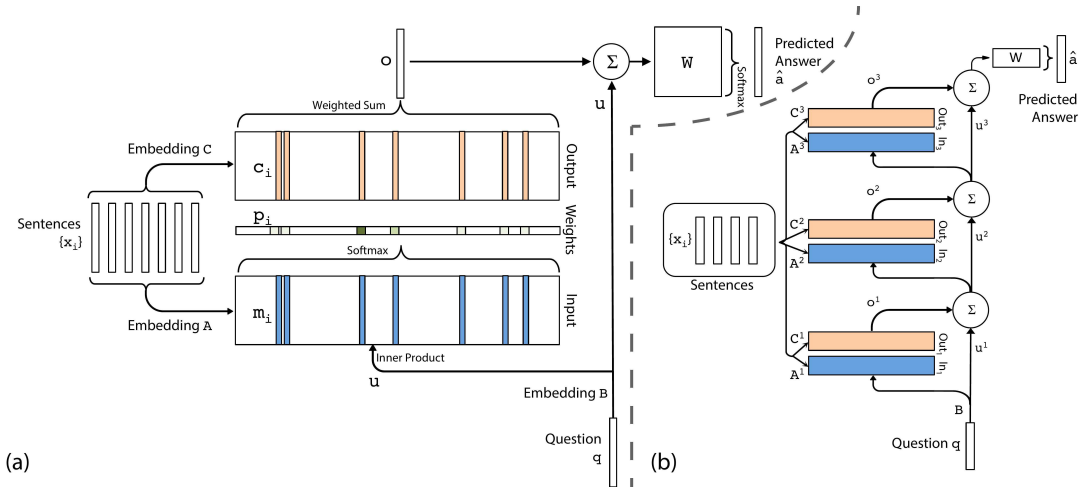

FIGURE 4.2 – Modèle de MemNN de [Sukhbaatar et al., 2015]

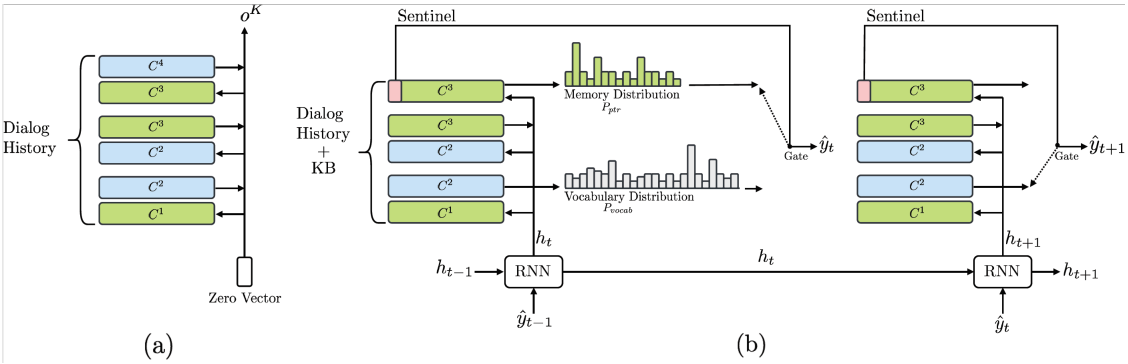

FIGURE 4.3 – Modèle de Mem2Seq de [Madotto et al., 2018]

$$A_i^k = C^k(x_i)$$

FIGURE 4.4

$$\overline{A_i^k} = \text{TRANS}(C^k(x_i))$$

FIGURE 4.5

4.2.3 Modèle mémoriel du WMM2Seq

Suite aux travaux sur le Mem2Seq, [Chen et al., 2019b] ont développé une architecture de système de dialogue inspiré des travaux en psychologie et en neuroscience sur les modèles de mémoire cognitive. De ce fait, on retrouve dans leur architecture des mémoires à long terme : une mémoire épisodique, une mémoire sémantique, ainsi qu'une mémoire de travail à court terme. Les deux premières sont des MemNN entraînées à apprendre respectivement les séquences de dialogue et les informations provenant d'une base de connaissance externe. La mémoire de travail est une unité récurrente à porte (GRU pour *gated recurrent unit*) permettant de choisir le nombre des informations à stocker et à récupérer pour répondre à l'énoncé courant de l'utilisateur. C'est également cette mémoire de travail qui génère la réponse du système, mot à mot, en décidant de la source de génération du mot : dans la mémoire épisodique, dans la mémoire sémantique, ou dans le vocabulaire stocké en mémoire à court terme. C'est enfin cette mémoire de travail qui met à jour les poids des deux mémoires à long terme et qui apprend une fonction de perte sur la décision prise de générer le mot suivant, et sur la qualité générale de la réponse fournie une fois tous les mots générés.
Pour tenter de résoudre le problème des mots inconnus, le *Working Memory Model to Sequence* (WMM2Seq) remplace la fonction de copie de la représentation de l'historique (eq. 4.4 au sein de chaque saut où A et C représentent les matrices de plongement pour chaque élément x de l'historique.

Elle la remplace par une fonction appelée $TRANS$ (eq. 4.5 qui permet à chaque saut de mieux raisonner sur le contenu de l'historique de dialogue.

Elle se matérialise soit par un Bi-GRU soit par un réseau de convolution (eq. 4.6) où hi est la représentation sensible au contexte lié à l'historique de dialogue et Φ^e une fonction entraînable pour la matrice de plongement.

Cependant, ce modèle n'explicite pas l'apprentissage d'un pistage d'état du dialogue à un tour de parole donné ni la classification des actes de dialogue des locuteurs, ce qui limite l'explicabilité des réponses générées par le système. D'autre part, comme le modèle génère les mots de la réponse les uns après les autres, il faut attendre la fin de la phrase générée avant de pouvoir l'évaluer. L'absence

$$
\begin{aligned}
h_i &= \mathrm{TRANS}(\phi^e(x_i)) \\
&= \mathrm{CNN}([\phi^e(x_{i-2}), \ldots, \phi^e(x_{i+2})])'
\end{aligned}
$$

ou

$$
\begin{aligned}
h_i &= \mathrm{TRANS}(\phi^e(x_i)) \\
&= \begin{bmatrix} \vec{h}_i \\ \overleftarrow{h}_i \end{bmatrix} = \begin{bmatrix} \overrightarrow{\mathrm{GRU}}(\phi^e(x_i), \vec{h}_{i-1}) \\ \overleftarrow{\mathrm{GRU}}(\phi^e(x_i), \overleftarrow{h}_{i+1}) \end{bmatrix}
\end{aligned}
$$

FIGURE 4.6

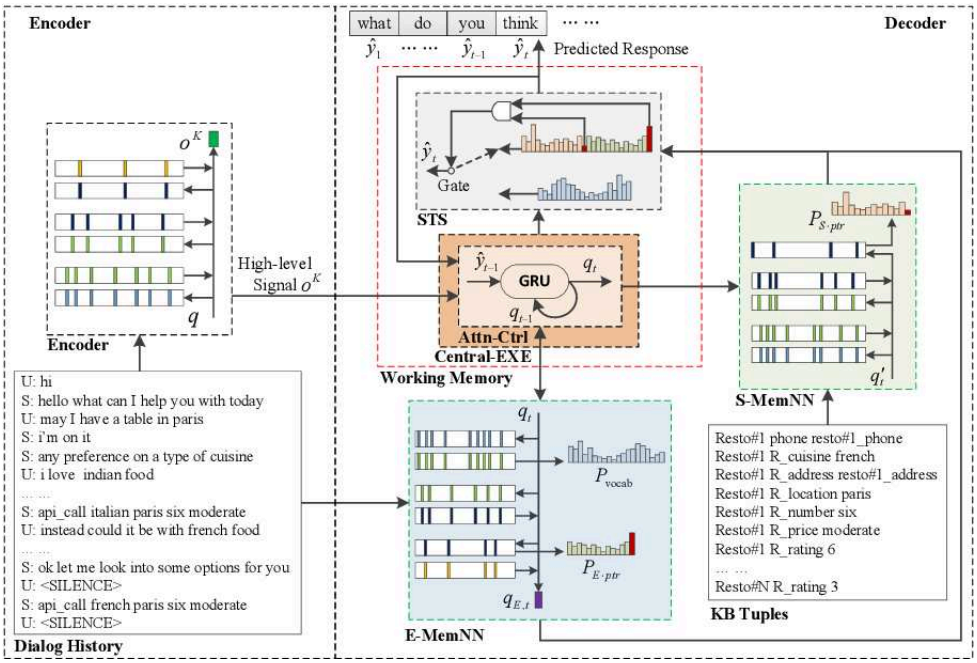

FIGURE 4.7 – Modèle de WMM2Seq [Chen et al., 2019b]

d'apprentissage au niveau des n-grammes de caractères ou des syntagmes limite la performance du système dès lors que la phrase à générer est de grande taille [Wang et al., 2020b]. Toutefois, la séparation des mémoires retenant l'information, en plus d'être une idée originale, permet de penser que lors de l'apprentissage de la mémoire épisodique, il pourrait émerger une dimension temporelle latente du dialogue, bien distincte des informations de la mémoire sémantique, ainsi qu'une dimension latente contenant une image du système lui-même tel qu'il s'imagine être représenté par l'utilisateur. On peut également supposer que la mémoire de travail possède la capacité d'apprendre à filtrer l'historique de dialogue en ne conservant en mémoire que les informations pertinentes pour la suite de la conversation. Notre travail a consisté à faire émerger la dimension réflexive, la dimension temporelle et la détection des incohérences dans l'apprentissage du modèle.
Enfin, un des aspects oubliés par le WMM2Seq, mais également par les modèles les plus récents tels que MarCo [Wang et al., 2020c] ou GPT-2 [Kulhánek et al., 2021] est celui du dialogue humain-machine. En effet, la plupart des systèmes actuels utilisent des corpus simulant des conversations humain-machine en particulier par la technique du dialogue avec compère (Magicien d'Oz). De ce fait les dialogues, bien que très scénarisés voire robotiques côté système, restent constitués par deux humains. A notre connaissance, dans le domaine du dialogue orienté-tâche, il n'y a guère que le corpus DSTC2 qui soit véritablement humain-machine. Or, les systèmes peuvent générer des énoncés que nous ne produirions jamais car ils nous sembleraient trop incohérents, sauf dans le cas où nous jouerions de ces incohérences, par exemple, à des fin comiques ou déstabilisantes. Il s'agit ici comme dans toute interaction d'un agent autonome ancré situé avec le monde, d'un problème d'*affordance* [Raubal and Moratz, 2008], ou potentialité en français. A part dans DSTC2, nous ne trouvons pas de dialogue contenant des énoncés du système avec des incohérences. Elles peuvent contenir des erreurs ou des imprécisions, mais elles ne vont pas à l'encontre de la cohérence du dialogue. Notre modèle utilise un classifieur d'incohérences, que nous détaillons en 4.4, entraîné à prédire si une réponse du système à un énoncé de l'utilisateur est cohérente ou pas. Ce classifieur d'incohérences prend en entrée un énoncé du système ainsi que tout le reste du dialogue, passé comme futur. Ainsi, non seulement en fonction de ce qui a été dit, mais également sur ce qui sera prononcé ensuite, le classifieur est capable de déterminer la présence d'une incohérence ou pas dans l'énoncé du système, ainsi que son type.

4.3 Proposition d'un Bi-WMM2Seq : modèle à trois axes

Nous expliquons ici plus en détails comment expliciter une dimension réflexive dans l'architecture originale du WMM2Seq, puis comment ajouter la temporalité

dans l'encodage de l'énoncé courant de l'utilisateur, et enfin comment détecter des incohérences pour doter le système d'un garde-fou, interne dans notre cas [Maes, 1987].

4.3.1 Dimension réflexive

Par dimension réflexive, nous ne prétendons pas être capable d'implémenter de la réflexivité dans un réseau de neurones, ni de pouvoir faire émerger une conscience de soi au sein du système. Il faut comprendre dimension réflexive comme la complexification du modèle de l'utilisateur en simulant un utilisateur qui possède un modèle de la machine. Autrement dit, la dimension réflexive que nous souhaitons expliciter représente l'image que le système se fait de l'image que l'utilisateur peut avoir du système lui-même. Comme nous l'avons évoqué en 4.1.1, dans la mesure où lors de l'apprentissage de la mémoire épisodique, le système a accès à tout l'historique de dialogue, il est fort probable que cette dimension existe de façon latente quelque part, ou pour parler avec des termes propres aux humains, de façon subconsciente. En explicitant cette dimension de façon consciente, nous nous rapprochons d'un modèle cognitif du fonctionnement de la mémoire humaine dans le dialogue dans lequel la représentation de soi fait partie des informations déclaratives.
Pour ce faire, le modèle a appris à inférer le tour de parole suivant de l'utilisateur à partir de la réponse du système générée. En effet, si nous parvenions, à partir de cette réponse du système, à prédire quelle serait la réponse de l'utilisateur, nous serions en mesure de permettre au classifieur d'incohérences d'être plus précis et permettre au système de se corriger sur le moment, en demandant à l'utilisateur une reformulation, ou en mettant fin au dialogue. Dans le WMM2Seq, la génération de chaque mot de la réponse se fait en fonction du précédent. Dans le cas où nous voulons que le système se corrige en temps réel, il est nécessaire de rétropropager l'information sur l'incohérence de la phrase générée dans la mémoire de travail du WMM2Seq afin qu'il génère une nouvelle réponse, en tenant compte de l'incohérence de la réponse précédemment générée.
Le tour de parole suivant de l'utilisateur est généré en implémentant un WMM2Seq interne (WMM2SeqI) dans le WMM2Seq global (WMM2SeqG). Ce WMM2SeqI est utilisé lorsque le WMM2SeqG, avec sa mémoire de travail, a généré la réponse du système. Celle-ci devient l'équivalent de la requête pour le WMM2SeqI, et l'historique de dialogue est concaténé avec le tour de parole courant utilisateur. L'imbrication de deux WMM2Seq telles des poupées russes nous a mené à nommer ce modèle Bi-WMM2Seq. La réponse générée est l'énoncé de l'utilisateur censé suivre la réponse du système, au tour de parole suivant. Cet énoncé de l'utilisateur au tour de parole suivant est ensuite concaténé avec la réponse du système et l'historique du dialogue, afin de tenter de prédire une incohérence dans la réponse du système (voir 1).

Algorithm 1 Algorithme du Bi-WMM2Seq

Require : $len(response) \leq max(len(possible\_responses))$
Ensure : $inc = 1$ **or** $inc = 0$
  $hop \leftarrow 3$
  $response \leftarrow null$
  $q \leftarrow user\_query$
  $c \leftarrow dialog\_history$
  while $inc \neq 0$ **and** $counter > 0$ **do**
    while $PAD \notin response$ **do**
      $word \leftarrow WMM2SeqG.predict\_next\_word(response, q, c)$
      $response \leftarrow response + word$
    end while
    $next\_user \leftarrow null$
    $c \leftarrow c + q$
    $q \leftarrow response$
    while $PAD \notin next\_user$ **do**
      $word \leftarrow WMM2SeqI.predictNextWord(next\_user, q, c)$
      $next\_user \leftarrow next\_user + word$
    end while
    $inc \leftarrow inc\_predictor(response, q, c, nex\_user)$
    $hop \leftarrow hop - 1\ Wmm2SeqG.update\_weights(response, q, c)$
  end while
  if $inc = 0$ **then**
    return $(response)$
  end if
  if $hop = 0$ **then**
    return $(default\_response)$
  end if

En outre, à l'instar des autres travaux sur les MemNN, nous utilisons trois sauts de réflexion lors de la détection d'une incohérence, c'est-à-dire que le système a la possibilité de se corriger trois fois, si une incohérence dans la réponse qu'il allait générer est détectée. A chaque saut, les poids de la GRU de la mémoire de travail, servant à prédire chaque mot, sont mis à jour afin de ne plus produire une réponse incohérente. Trois sauts de réflexion, d'après les autres travaux sur les MemNN, sont un bon compromis entre un apprentissage avec un seul saut où le système n'a pas le temps de se corriger et six sauts qui n'est pas recommandé [Bordes et al., 2017].

4.3.2 Dimension temporelle

Afin d'expliciter la dimension temporelle dans l'apprentissage du modèle, nous avons trouvé un modèle prédictif capable de trouver les tranches temporelles correspondant aux différents ensembles de sous-dialogues présents au sein d'un même dialogue. Il existe plusieurs travaux comme [Pasupat et al., 2019, Qin et al., 2021] qui tentent de prédire les séquences temporelles dans un dialogue, il existe même un jeu de données appelé TimeDial annoté en séquences temporelles[4]. Nous nous sommes inspirés des travaux de [Papangelis et al., 2017b] sur l'utilisation de plusieurs indices, acoustiques comme sémantiques, pour déterminer la longueur d'un dialogue à partir d'un tour de parole donné et du contexte précédent (l'historique de dialogue). Bien que les traits les plus discriminants pour prédire la longueur du dialogue soient les indices acoustiques [Lykartsis and Kotti, 2019], l'utilisation de l'entropie conditionnelle de l'état de croyance d'un tour de parole donné se révèle efficace pour cette même tâche [Papangelis et al., 2017b]. En plus de sa simplicité, nous l'avons préférée à un tf-idf ou à des plongements de mots car l'entropie est un modèle universel de calcul de la quantité d'information, adaptable à des problématiques dépassant le cadre du traitement du dialogue, voire du traitement de la langue. De plus, sa faible complexité [Wu, 2012] permet de l'ajouter dans notre système général sans alourdir le temps d'apprentissage global du modèle. En imitant [Huang et al., 2018] qui implémentent un Bi-LSTM pour modéliser la temporalité dans les intentions de l'humain durant une conversation, nous combinons la mesure de l'entropie avec un Bi-LSTM pour prédire la taille restante du dialogue à un tour de parole donné.

L'entropie conditionnelle

L'entropie de Shannon[Shannon, 1948] peut être vue comme une mesure représentative de la quantité d'information contenue dans les messages issus d'une source, quantité qui est calculée en fonction des symboles constituants les messages et des hypothèses que l'on pose sur le modèle de la source. Elle peut aussi s'interpréter comme une mesure de la prévisibilité du contenu des messages produits par cette source, ou encore comme une mesure de la limite de la capacité des messages de cette source à être représentés sous une forme plus compacte sans perte d'information, par exemple le nombre de bit minimal requis pour un message. Concernant la langue anglaise,[Montemurro and Zanette, 2011] estime que son entropie est en moyenne environ 9 bits par mot, alors que celle du finnois est de l'ordre de 10 bits par mots. Les méthodes de calcul pour estimer la valeur de l'entropie d'une source sont nombreuses et dépendent entre autres choses, des échantillons de message dont on dispose, de ce qu'on connaît de la source et de bien sûr de l'objectif pour lesquels on calcule l'entropie. Il est naturellement possible de

4. https ://github.com/google-research-datasets/timedial

recourir à des calculs statistiques effectués sur les échantillons de messages dont on dispose, mais on peut aussi utiliser avoir recours à des algorithmes de compression de données [Montemurro and Zanette, 2011]. Un des facteurs-clé pour déterminer la méthode d'estimation est la manière dont on souhaite rendre compte des contraintes imposées par le "langage" défini par la source. Elles portent sur l'ordre d'apparition des différents symboles au sein d'un message et de leur interdépendance, en particulier lorsqu'on travaille sur des textes[Manning and Schutze, 1999]. Dans ce cas, il est souvent fait recours à l'entropie conditionnelle, que nous avons choisie pour mesurer la quantité d'information contenue dans les dialogues. Nous reprenons ci-dessous les formules de [Dušek et al., 2020] pour le calcul de l'entropie de Shannon (Eq. 4.1) et de l'entropie conditionnelle dans l'esprit des modèles de langage n-gramme (Eq. 4.2) que nous avons utilisées pour nos expériences.

$$H(text) = -\sum_{x \in text} \frac{freq(x)}{len(text)} log_2 \left(\frac{freq(x)}{len(text)} \right) \tag{4.1}$$

$$H_{cond}(text) = -\sum_{(c,w) \in text} \frac{freq(c,w)}{len(text)} log_2 \left(\frac{freq(c,w)}{freq(c)} \right) \tag{4.2}$$

Dans l'équation 4.2 de l'entropie conditionnelle, la paire (c, w) représente tous les types de n-grammes uniques, présents dans le $texte$. Elle est composée du préfixe du n-gramme (contexte) c et du dernier *token* w du n-gramme.

Enfin, nous avons eu recours à un lissage par interpolation linéaire, comme dans [Zhai and Lafferty, 2017], afin de prendre en compte la présence éventuelle dans les dialogues du corpus d'évaluation de bi-grammes de mots absents du corpus de développement qui nous a servi pour estimer l'entropie conditionnelle. Cela se traduit par la formule 4.3 où la fonction $Interpol$ se définit par 4.4 et où $\lambda 1$, $\lambda 2$ et $\lambda 3$ sont des paramètres réels d'optimisation et $vocab$ le nombre total de mots dans le corpus de développement. Dans nos expérimentations, $\lambda 1 = 0,9$, $\lambda 2 = 0,09$ et $\lambda 3 = 0,01$.

$$\widehat{H}_{cond}(text) = -\sum_{(c,w) \in text} \frac{freq(c,w)}{len(text)} log_2 \left(Interpol \left(\frac{freq(c,w)}{freq(c)} \right) \right) \tag{4.3}$$

$$Interpol \left(\frac{freq(c,w)}{freq(c)} \right) = \lambda 1 \times \frac{freq(c,w)}{freq(c)} + \lambda 2 \times \frac{freq(w)}{len(vocab)} + \lambda 3 \tag{4.4}$$

Les tranches temporelles

Dans le chapitre 5, nous expliquons en détails les différents traits utilisés pour calculer l'entropie, le nombre optimal de tranches temporelles par dialogue en fonction : du domaine de celui-ci, de la longueur des tours de parole et des états du dialogue, de la présence ou non d'une incohérence, de l'historique de dialogue, de l'entropie de caractères, de bigrammes de caractères et de mots.
Nous faisons l'hypothèse que nos pouvons distinguer cinq tranches temporelles dans les dialogues, tout domaine confondu.

1. Les salutations : elles se caractérisent par une entropie faible de l'utilisateur et du système. Les mêmes phrases sont souvent échangées. Elle est très courte (un tour de parole)
2. les informations de l'utilisateur : elles se caractérisent par une entropie plus forte des énoncés de l'utilisateur que celles du système car c'est la tranche où l'utilisateur explique à ce dernier de quoi il a besoin, quelles sont ses demandes. Son vocabulaire est plus diversifié et moins prévisible. Elle peut être relativement longue, elle s'étend jusqu'à la moitié du dialogue.
3. l'échange d'information, les confirmations et les corrections : c'est la phase la plus floue et la plus complexe à décrire. L'entropie est élevée chez l'utilisateur comme chez le système. C'est là où le système a tendance à demander plus d'informations à l'utilisateur, ou l'avertir qu'il effectue les recherches, ou que l'utilisateur corrige le système qui a mal compris sa demande. Elle se situe en général au milieu du dialogue, et s'étend jusqu'aux trois quarts du dialogue.
4. les requêtes de l'utilisateur : c'est la phase où le système répond aux demandes de l'utilisateur concernant des requêtes sur des objets (un numéro de téléphone, une adresse...). L'entropie du système est plus élevée que dans le reste du dialogue eu égard de la diversité de vocabulaire dans les possibles réponses à fournir à l'utilisateur. C'est une phase assez courte, néanmoins, dans le cas des dialogues qui s'éternisent (plus de vingt échanges), on s'aperçoit que c'est souvent cette phase qui est allongée.
5. la fin du dialogue : elle se caractérise par un seul échange. Elle est difficile à différencier de la phase de salutation en cela que leur entropie est très similaire. Nous utilisons le numéro du tour de parole pour pouvoir les distinguer : si ce n'est pas le premier ou le deuxième tour de parole, c'est que nous sommes dans la fin du dialogue.

Cette prédiction de la tranche temporelle à laquelle appartient un tour de parole, nous l'intégrons dans notre MemNN de décodage du contexte, à chaque saut de réflexion qui permet de déterminer les pointeurs de la mémoire épisodique, afin de faire émerger de façon explicite la dimension temporelle dans le modèle, ainsi que dans la MemNN épisodique. De cette façon, nous sommes capables de fournir au système une information supplémentaire quant à la position relative du tour de parole courant dans la suite du dialogue. En plus de permettre au modèle de générer avec plus de précision la réponse du système, cette information de position limite la possibilité de générer une incohérence dans la mesure où le système peut estimer sa progression dans le dialogue. En outre, cette prédiction n'est utilisée par le système uniquement si la confiance dans son estimation est supérieure à un seuil relatif fixé à 70%. D'autre part, la prédiction du tour suivant de l'utilisateur permet au modèle de mieux détecter la temporalité du tour courant de l'utilisateur, et donc d'être plus précis dans la réponse à générer. De la même façon, le fait de

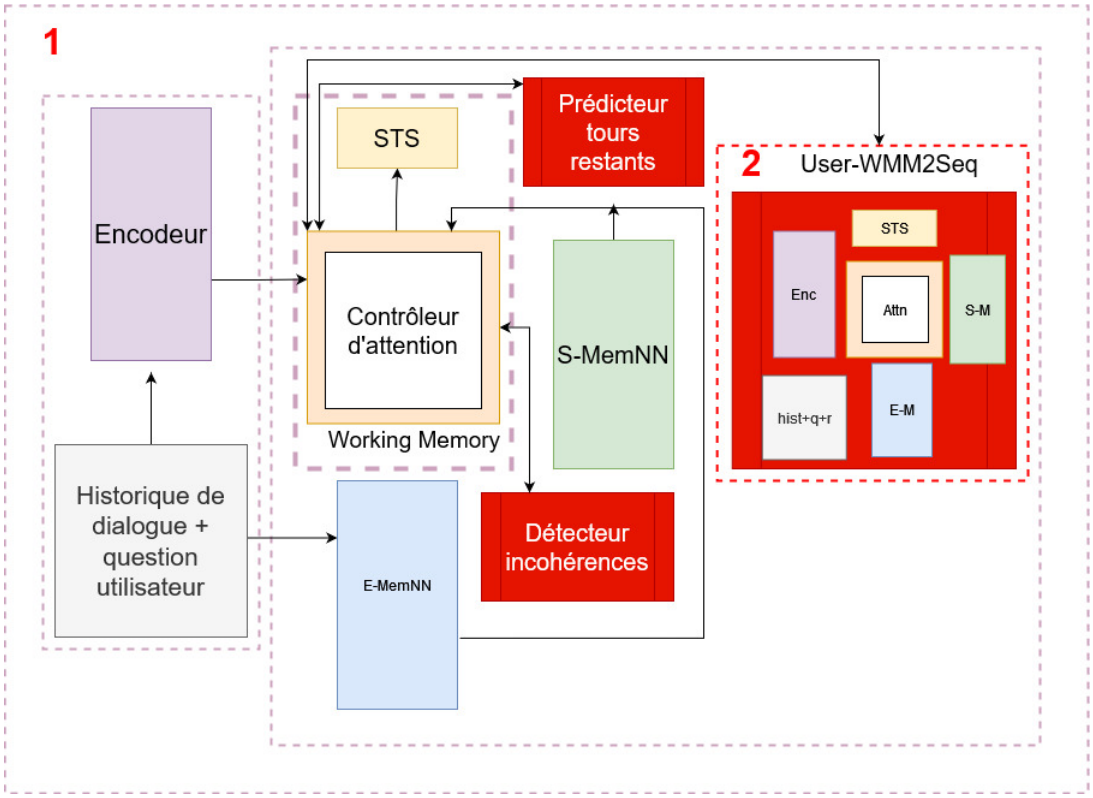

FIGURE 4.8 – Le système Bi-WMM2Seq

connaître la temporalité dans laquelle se situent le tour courant de l'utilisateur ainsi que la réponse du système permet au modèle de mieux prédire le tour suivant de l'utilisateur. Les deux modules se renforcent ainsi mutuellement. La figure 4.8 résume le modèle que nous désirons mettre en place.

1. Le système reçoit l'énoncé courant de l'utilisateur et il récupère l'historique de la conversation

2. Il prédit la temporalité de l'énoncé courant de l'utilisateur

3. Il filtre l'historique de la conversation pour ne garder que les informations que peut stocker la mémoire de travail

4. L'énoncé courant de l'utilisateur et l'historique de la conversation sont encodés dans un MemNN

5. La MemNN épisodique est alimentée par l'énoncé courant de l'utilisateur et l'historique de la conversation

6. Les données de la base de données externe (KB) sont encodées dans la MemNN sémantique

7. Les mots de la réponse candidate sont générés les uns après les autres à la manière du WMM2Seq

8. La mémoire de travail sort la concaténation de l'énoncé courant de l'utilisateur et de l'historique de la conversation sous forme de nouvel historique de conversation

9. La réponse candidate ainsi que la concaténation de l'énoncé courant de l'utilisateur et de l'historique de la conversation sont données en entrée d'un WMM2Seq interne (WMM2SeqI) qui imite les étapes 2 à 7

10. Le tour suivant de l'utilisateur est prédit, il est utilisé pour prédire à nouveau la temporalité de l'énoncé courant de l'utilisateur et de l'historique de la conversation. Les étapes 2 à 10 sont répétées jusqu'à ce que le système soit certain de la temporalité.

11. Sortie de la réponse candidate finale et les poids sont à jour

12. Sortie du tour suivant de l'utilisateur final

13. Le détecteur d'incohérence se met au travail pour la réponse candidate finale

    — Si détection d'une incohérence, une correction est effectué (demande de reformulation ou redirection vers un agent humain). Si correction, on effectue une mise à jour des poids de la GRU de la mémoire de travail. Les étapes 11 à 13 sont répétées

    — Si pas d'incohérence, la réponse finale est envoyée à l'utilisateur

4.4 Les incohérences dans le dialogue

Afin de développer nos hypothèses, nous avons imaginé un moyen de montrer la nécessité de voir émerger explicitement une dimension réflexive dans l'apprentissage du modèle. Pour y parvenir, nous nous sommes penchés sur les incohérences

dans les réponses du système aux énoncés de l'utilisateur. Nous supposons que l'utilisation d'un historique de dialogue limité et la prédiction du prochain tour de l'utilisateur peuvent améliorer la classification des incohérences.
Aucune des architectures évoquées dans les chapitres précédents ne traite le problème des incohérences du système, inhérent à toute tâche de génération de dialogue. En effet, au cours d'une conversation humain-machine, il n'est pas rare d'observer la machine dire quelque chose d'inattendu ou d'incohérent [Litman et al., 2006, Engelbrecht and Möller, 2010]. La détection et la correction de ces incohérences est difficile, mais constitue une amélioration importante car elle permet au système de se corriger lui-même [Zhang et al., 2019b], nous rapprochant ainsi d'une architecture d'apprentissage de type *lifelong learning* [Veron, 2019, Hancock et al., 2019].
Nous avons expliqué dans le chapitre 3 que la mémoire est une fonction essentielle pour maintenir la continuité d'un dialogue et on peut légitimement se demander laquelle de ses caractéristiques a un impact sur la cohérence d'un dialogue, par exemple on sait que la mémoire de travail humaine a une taille très réduite [Baddeley, 2010] et des études en neurosciences ont montré que certaines parties du cerveau d'un locuteur consacrées à l'écoute sont actives pendant qu'il parle et vice versa [Kuhlen et al., 2017]. L'un des aspects saillants du rôle de la mémoire au cours d'un dialogue [Schaub and Vaudapiviz, 2019a] est de permettre à chaque participant de se faire une représentation de l'autre, ce qui contribue à fluidifier l'échange [Laurent, 2014]. Pour un système de dialogue, cela signifie apprendre une représentation de l'utilisateur et de ses demandes, mais aussi une représentation de lui-même tel qu'il est vu par l'utilisateur. En d'autres termes, pour un dialogue donné, à un tour t, lorsque l'utilisateur parle, le système prépare non seulement sa réponse mais aussi la prochaine requête de l'utilisateur qui suivra sa réponse, voire sa propre réponse à $t + 1$.
En ce qui concerne l'analyse des erreurs du système, [Whitney et al., 2017] modélisent avec un POMDP [Sammut and Webb, 2010] l'incertitude d'un agent de dialogue lorsqu'il répond à une question de l'utilisateur afin d'améliorer la précision de la réponse. [Welleck et al., 2019] utilisent un modèle d'inférence du langage naturel pour améliorer la cohérence d'un système dans un dialogue, [Li et al., 2020a] intègrent ensuite la cohérence dans le signal d'entraînement du système. [Dziri et al., 2019] appliquent une approche similaire basée sur l'inférence pour l'évaluation des systèmes de dialogue. Au cours de la tâche partagée DSTC6 [Hori et al., 2019], la détection d'incohérence pour les dialogues ouverts était l'un des problèmes étudiés ; cependant, les incohérences trouvées restent assez spécifiques à ce type de dialogue. [Gao et al., 2019] montrent que lorsqu'une conversation dépasse un certain nombre de tours de dialogue, les systèmes de bout en bout voient leurs performances diminuer, ce qu'ils attribuent à l'historique de la conversation qui devient bruité s'il est trop grand. Le travail le plus proche du nôtre est celui de [Li, 2020], qui crée un modèle du locuteur à partir des comportements

incohérents de chacun des participants à un dialogue. Il se sert du corpus Twitter Persona dataset, un corpus de dialogues ouverts annotés en personnalités du locuteur [Li et al., 2016]. Il montre que l'annotation en cohérence du dialogue favorise la réduction du taux d'erreur du système. À notre connaissance, il n'y a pas eu de système qui prédise le tour de l'utilisateur suivant et qui filtre l'historique des dialogues orientés-tâche pour anticiper les incohérences du système.

4.4.1 Typologie des erreurs

Nous devons faire la distinction entre les erreurs de compréhension ou de décision dans les dialogues humains, et les incohérences propres aux robots dans un dialogue humain-machine. En effet, lors d'une conversation orientée vers une tâche entre deux humains, les erreurs ou les problèmes peuvent provenir de malentendus liés à la tâche elle-même, ou à un vocabulaire connu uniquement par l'un des locuteurs. Ces erreurs, dans la théorie conversationnelle, conduisent à des co-corrections quasi systématiques entre les deux interactants. Des co-corrections auto-initiées et auto-réparées ou hétéro-initiées sont appliquées dès qu'une erreur se produit [Chernyshova, 2018]. Cependant, ces erreurs sont corrigées grâce à l'inférence, c'est-à-dire que les erreurs explicites d'un locuteur amènent l'auditeur à interpréter et à inférer ce que le premier locuteur a réellement voulu dire. De la même manière, le locuteur qui a commis l'erreur déduit ce que l'auditeur aurait pu comprendre et déduire, afin de s'auto-corriger : [Deppermann, 2018, Fernandez et al., 2006]. Dans l'interaction humain-machine, cette inférence est entravée par le fait que l'humain sait qu'il parle à une machine dont la capacité d'inférence et de correction est a priori limitée. Par conséquent, lorsque l'utilisateur ou la machine commet des erreurs de type humain, cette limitation diminue inévitablement la confiance et la patience de l'utilisateur. Cependant, notre travail se concentre principalement sur les *incohérences* - les erreurs du système qui ne sont pas de type humain mais endémiques au processus d'apprentissage du système. Les incohérences peuvent souvent amener l'utilisateur à finaliser le dialogue sans succès. Pour que la machine acquière une certaine capacité de co-correction, c'est-à-dire pour qu'elle soit capable de détecter les incohérences dans ses propres réponses, elle doit avoir une certaine représentation de l'utilisateur qui les interprète. Les dialogues DSTC2 se déroulent entre un système réel et des utilisateurs [Henderson et al., 2014b], et comprennent donc de nombreuses incohérences du système, notamment basées sur des erreurs de reconnaissance vocale. Un grand nombre de ces incohérences sont récupérables (par exemple, l'utilisateur répète sa requête et celle-ci est reconnue correctement). Le dialogue n'est donc pas menacé dans son ensemble. Sur la base d'une analyse empirique du corpus, [Schaub et al., 2021] ont établi une liste des types d'incohérences les plus fréquents dans l'interaction humain-machine, avec un exemple pour chacun d'entre eux (voir Tableau 4.1 pour les statistiques) :

1. **Intention incomprise** (*Ex : U -> dans quelle zone est-il situé ? S -> Eraina est un grand restaurant.*) Ici, le système n'a pas prédit l'intention correcte de l'utilisateur (*request*) ni l'entité (*zone*).
2. **Mauvaise entité** (*Ex : U -> Méditerranée. S -> Il n'y a pas de restaurant jamaïcain ici.*) Nous distinguons ce cas du précédent car, dans ce cas, le système a effectivement prédit l'intention mais il a ensuite généré une sortie incorrecte.
3. **Repetition** (*Ex : U -> gastropub. S -> Quel type de restaurant recherchez-vous ? U -> Gastropub. S -> Ok laissez-moi regarder.*) Le système a forcé l'utilisateur à répéter la même énonciation.
4. **"au revoir" incompris** (*Ex : U -> au revoir. S -> Le restaurant Paris sert de la cuisine française.*) Il s'agit d'un type particulier d'intention incomprise où le système ne comprend pas que l'utilisateur veut mettre fin à la conversation. Nous l'avons noté parce que dans les cas d'utilisation réels, c'est l'une des incohérences qui agace le plus l'utilisateur.
5. **Insultes ou grossièretés de l'utilisateur** (*Ex : U -> je m'en fous si hk_fusion sert de la cuisine chinoise, je cherche de la cuisine thaïlandaise. S -> je suis désolé mais il n'y a pas de restaurant indien qui corresponde à votre demande.*) En plus de fournir un indice sur l'incompréhension de l'intention, cette catégorie fournit également un retour sur le ressenti de l'utilisateur, ce qui est utile lorsqu'on envisage un déploiement commercial [Kumar, 2010]. Par conséquent, nous avons décidé de distinguer les insultes des autres types d'incohérence.
6. **Mauvais appel API** (*Ex : U -> dans le sud. S -> appel api = Cuisine, ANY_location, cher*). Comme l'appel API est explicite, nous pouvons vérifier facilement s'il est cohérent avec les informations données par l'utilisateur.
7. **Incomplétude** (*Ex : U -> téléphone et adresse. S -> Le téléphone est ****) Le système ne répond qu'à certaines des demandes de l'utilisateur.
8. **Erreur de négation** (*Ex : U -> pas de cuisine française. S ->Il y a des restaurants servant de la cuisine française. Quelle zone voulez-vous ?*) Le système est incapable de traiter une négation dans la requête de l'utilisateur.

4.4.2 Classification des erreurs

Un grand nombre des incohérences ci-dessus peuvent être détectées à l'aide de règles simples (par exemple, détecter que l'énoncé du système se répète deux fois de suite). Des règles plus élaborées peuvent concerner l'inadéquation des types d'entités entre les énoncés de l'utilisateur et du système, etc. à un moment particulier d'un dialogue, etc. La principale source d'information pour toutes ces règles

type d'incohérence								tours incohérents	tours corrects	tours total
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**			
783	245	1,360	275	11	242	780	64	3,760	26,179	29,939

TABLE 4.1 – Types et nombres d'incohérences dans le corpus DSTC2/bAbi

est l'historique du dialogue. Nous notons que l'annotation des incohérences n'est pas indépendante du tour. Par exemple, pour détecter que le système dit deux fois la même phrase et que cela dérange l'utilisateur, nous devons connaître les tours t et $t+1$. L'anticipation de l'énonciation de l'utilisateur suivant au tour $t+1$ pourrait permettre au système de fournir une réponse plus précise au tour t.[5] Nous verrons dans les expérimentations du chapitre 5 dans quelle mesure une réduction de la taille de l'historique du dialogue, combinée avec la connaissance du prochain tour de l'utilisateur et un indice sur la temporalité, peut permettre au système de mieux détecter son propre comportement incohérent.
Une fois ce détecteur entraîné, nous voulons l'intégrer dans l'architecture globale en tant que dernière étape avant la génération d'une réponse finale. Si une incohérence y est détectée, la réponse n'est pas générée et le système tente de se corriger en mettant à jour les poids de sa mémoire de travail. Au bout de trois sauts de réflexion, si une incohérence persiste, une réponse par défaut ou une demande de reformulation est générée à la place, et dans ce dernier cas seuls les poids de la GRU sont maintenus, les autres informations (temporalité, sauts de réflexion, tour de parole suivant utilisateur..) sont rétablis comme avant l'énoncé courant de l'utilisateur.

4.5 Modules non implémentés et limites du modèle

4.5.1 Réponses alternatives

Nos réflexions se sont aussi portées sur de multiples réponses possibles à un même énoncé de l'utilisateur. En effet, [Zhang et al., 2020b], après une étude des conversations humaines, montrent que pour une question donnée de l'utilisateur, plusieurs réponses du système sont possibles 4.9.

La plupart des modèles génèrent la réponse possédant le taux de confiance le plus élevé. [Zhang et al., 2020b] ont choisi de combiner les réponses potentielles pour en fabriquer une nouvelle, contenant plus d'information. Ils obtiennent ainsi des dialogues raccourcis de plusieurs échanges et limitent donc la taille de l'histo-

5. Nous supposons que le système initie le dialogue. Par conséquent, nous prenons l'énoncé suivant de l'utilisateur dans le même tour. C'est le cas pour DSTC2 (voir chapitre 5)

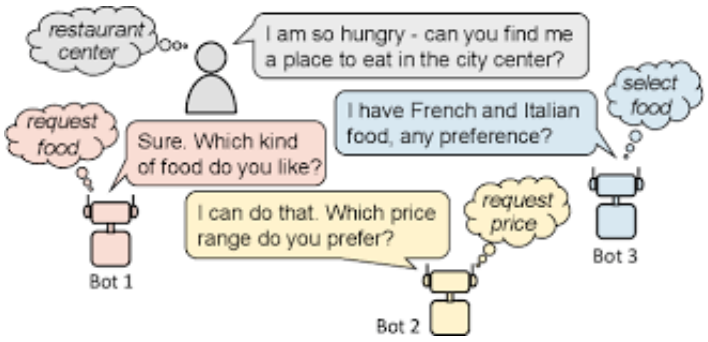

FIGURE 4.9 – Illustration des réponses multiples de [Zhang et al., 2020b]

rique de dialogue, minimisant ainsi le risque d'incohérence du système. Une de nos perspectives est d'implémenter ce module afin d'améliorer la précision du système lorsqu'il génère sa réponse. En effet, si le système était capable de non seulement choisir entre plusieurs réponses, mais de traiter les réponses possibles comme les membres d'une autre réponse, faisant office de résumé des premières, il pourrait fournir des réponses permettant de limiter le risque d'intentions mal comprises, d'incomplétude et de répétition, trois des incohérences les plus pénalisantes pour le succès du dialogue.

4.5.2 Limites du modèle

Ce modèle contient plusieurs limitations. Tout d'abord, il se base sur la détection des incohérences dans les réponses du système à l'utilisateur. Or l'immense majorité des modèles sont appris en utilisant des corpus de conversations entre des humains, soit créés par la technique du dialogue avec compère, soit par des corpus de conversations écrites (sites de messagerie instantanée ou de clavardage, sites de messagerie asynchrone...) ou orales. Dans la mesure où les incohérences sont propres à la machine, il est difficile de corréler les imperfections des modèles appris sur ces corpus avec les incohérences potentiellement présentes dans les réponses du système. Une des solutions envisagées est d'utiliser des simulateurs conversationnels de haut niveau afin de créer des jeux de données humain-machine et de pouvoir trouver des incohérences exploitables.
D'autre part, nous partons de l'hypothèse que les incohérences sont corrélées avec les échecs du dialogue. Cela se traduirait par le fait que les incohérences empêchent le dialogue de se poursuivre, et la tâche sous-jacente à celui-ci n'est jamais résolue. Dans le corpus duquel nous nous sommes inspirés pour générer la typologie des incohérences, les dialogues sont étiquetés comme étant des succès ou des échecs. Or, après une analyse des annotations, nous avons découvert plusieurs discordances dans les critères définissant un dialogue échoué ou réussi. Après ré-annotation d'un échantillon représentatif du jeu données, nous nous sommes aperçus que 20% des dialogues annotés comme des succès présentaient tous les critères d'un échec

d'après le manuel officiel d'annotation [6]. Il est donc important de prendre en compte cette marge d'erreur, et d'imaginer que le taux de corrélation entre les incohérences et l'échec des dialogues est encore plus élevée que celui que nous avons mesuré.
De plus, le corpus DSTC2 ne compte que trois mille dialogues, ce qui est peu quand on considère le nombre de textes nécessaires pour les méthodes d'apprentissage les plus récentes. En outre, il n'y a qu'un seul domaine qui soit traité : la réservation de restaurant. Sans compter le fait que l'on perd en diversité linguistique car, dans la réalité,la grande majorité des conversations orientés-tâche sont multi-domaine c'est-à-dire qu'au sein d'un même dialogue, plusieurs domaines sont traités, parfois sans linéarité : l'utilisateur ou le système peuvent converser sur un domaine A, puis un domaine B, puis l'un des deux peut revenir au domaine A, avant de converser d'un domaine C ou à nouveau du domaine B. Il est à notre avis impératif de construire de nouvelles ressources dialogiques dans le cadre d'une interaction humain-machine afin de vérifier la robustesse de notre modèle.
Dans un tout autre registre, celui de l'évaluation, nous pensons que les évaluations automatiques actuelles ne permettent pas de révéler les deux facteurs qui, inspirés par le protocole PARADISE, nous semblent essentiels pour le succès d'un dialogue orienté-tâche :

— La satisfaction utilisateur : est-ce que l'utilisateur a trouvé le dialogue utile, agréable. Le conseillera-t-il aux autres utilisateurs ? Le réutilisera-t-il ? Est-il prêt à fournir un retour permettant d'améliorer le système ? Toutes ces caractéristiques ne sont pas évaluées automatiquement.
— Le risque d'attrition : c'est une mesure corrélée à la satisfaction de l'utilisateur. Derrière tout système de dialogue se cache un service défini par la tâche (réservation d'une table de restaurant, location d'une voiture ou d'un appartement, consultation d'un compte bancaire...). Il n'existe pas aujourd'hui de mesure évaluant le risque d'attrition, autrement dit la possibilité que l'utilisateur, après avoir conversé avec le système de dialogue, non seulement soit insatisfait par le dialogue et donne un retour négatif, mais en plus décide de ne plus jamais utiliser le service. En d'autres termes, à notre connaissance, on ne sait pas évaluer le risque que le service ayant intégré le système de dialogue perde, à cause de ce dernier, un client.

Un troisième facteur, non inclus dans PARADISE, est également à prendre en considération, est celui de l'origine de l'incohérence. En effet, nous sommes capables de prédire l'incohérence dans la réponse du système. Nous sommes également capables de mesurer la corrélation entre cette incohérence et le tour de parole suivant de l'utilisateur ainsi qu'avec l'historique de conversation. Cependant, nous ne sommes pas encore capable d'établir l'origine véritable de l'incohérence. Où se situe son origine ? Quels mots de l'utilisateur ou du propre système l'ont déclenchée ? Quel est le contexte exact de son apparition ? Nous ne pouvons qu'effectuer des observations quant à certaines intentions ou certaines entités plus propices à provoquer une

6. https ://github.com/matthen/dstc/blob/master/handbook.pdf

incohérence, mais cela ne nous donne pas l'information exacte quand à l'origine de l'incohérence. Cette origine pourrait être primordiale, lors de l'apprentissage, pour renforcer la vigilance du modèle lors de tel ou tel contexte dont nous savons qu'il est l'origine de plusieurs incohérences.
Enfin, nous n'exploitons pas les actes de dialogue et le pistage d'état. Un des prochains ajouts à cette architecture serait d'avoir une partie de la mémoire épisodique dédiée uniquement aux actes de dialogue et à l'état de croyance.

4.6 Conclusion

En définitive, nous proposons un modèle de système de dialogue basé sur les réseaux de mémoire de travail de séquence à séquence permettant de faire émerger explicitement, au cours de l'apprentissage, une représentation fine du passé en convoquant

1. l'historique de dialogue,
2. une conscience précise du présent, grâce à l'estimation de la progression dans le dialogue basée sur les tranches, temporelles
3. et une anticipation efficace du futur, grâce à la prédiction du tour de parole suivant de l'utilisateur.

Ces trois dimensions permettent au modèle, une fois déployé, de mieux détecter une incohérence dans la réponse à générer et de se corriger en temps réel.

Chapitre 5

Résultats expérimentaux

Let us hope that I was wrong

Morpheus

Dans ce chapitre, nous présentons les résultats expérimentaux que nous avons obtenus à la fois sur la détection des incohérences dans différents jeux de données, sur la prédiction de la fin du dialogue et sur la qualité de la réponse du système générée par notre modèle à trois axes. Dans la section 5.1, nous décrivons les différents corpus sur lesquels nous avons travaillé, notamment DSTC2 et MultiWOZ, et leurs spécificités. Dans la section 5.2, nous présentons nos travaux sur la constitution d'une nouvelle ressource pour le dialogue : nous expliquons notre méthode pour fusionner jusqu'à quatre corpus de dialogue orienté-tâche pour différents domaines, afin de pouvoir tester la capacité d'un modèle comme le nôtre à utiliser des dialogues de différentes sources pour devenir plus robuste. Section 5.3 et 5.4 nous démontrons l'importance de la détection des incohérences pour connaître le succès d'un dialogue, et l'intérêt de l'entropie pour prédire au niveau du tour de parole le nombre de tours à venir. Section 5.5 nous décrivons les modèles que nous utilisons comme référence pour comparer les performances de notre modèle à trois axes et présentons les résultats obtenus. Nous évaluons notre modèle avec différentes mesures pour la génération de réponse.

5.1 Corpus de dialogue actuels

Pour nos travaux, nous avions besoin d'exploiter des dialogues orientés-tâche (cf 2.1.1), entre un humain et une machine, en langue française. Cependant, comme nous l'avons mentionné en 2.3, il n'existe pas beaucoup de corpus de dialogues orientés-tâche disponibles en libre accès. Il restait donc deux solutions : constituer un corpus de dialogues à partir de données telles que des corpus de conversation

de clavardage, ou abandonner l'idée de réaliser nos expériences sur le français. Dans le chapitre 6, nous relatons notre tentative de constituer une ressource en français pour des dialogues orientés-tâche, et les difficultés qui se sont révélées malheureusement trop chronophages pour permettre à notre tentative d'aboutir.
Nous avons alors opté pour la seconde solution : utiliser des ressources existantes en langue anglaise, pour laquelle les deux corpus les plus utilisés actuellement sont : DSTC2 et MultiWOZ.

5.1.1 DSTC2 : erreurs et spécificités

Nous l'avons déjà évoqué en 2.3, DSTC2 est un vrai corpus humain-machine créé par l'université de Cambridge et datant de 2013, par opposition aux corpus créés avec compère. Il est composé d'environ trois-mille dialogues, d'une taille moyenne de douze échanges (énoncé utilisateur et énoncé système). Il contient plusieurs annotations complémentaires :

- Le but du dialogue : il apparaît sous la forme d'un dictionnaire présentant les informations que doit demander l'utilisateur et celles que doit demander le système. Il est accompagné d'un texte en langue naturelle résumant ce but.
- Le succès du dialogue : c'est un score déterminant le taux de réussite du système à compléter la tâche et à achever le dialogue correctement.
- Les actes de dialogue de l'utilisateur ainsi que ceux du système, pour chaque tour.
- L'état de croyance du système, également pour chaque tour.

Comme tout jeu de données, DSTC2 comporte aussi son lot de défauts. On peut en distinguer deux sortes : ceux liés au système de reconnaissance vocale et au POMDP après un énoncé de l'utilisateur, et ceux liés aux annotations fournies par les étiqueteurs. Pour le premier cas, nous distinguons les erreurs, des incohérences (4.4) en cela que nous considérons les premières comme "humaines", liées à la tâche et à son succès (demande de l'utilisateur inaccessible au système, erreur sur un produit ou une information...) alors que les secondes sont "robotiques", liées à la fluidité du dialogue et d'origine linguistique (répétition de la même phrase sans raison, information fournie sans que l'utilisateur l'ait demandée...).
Pour le second cas, il peut s'agir soit d'une erreur dans l'annotation des actes de dialogue, soit d'une erreur dans le succès de la tâche et la qualité du dialogue [Devillers et al., 2002]. En effet, après avoir mené une étude empirique sur le succès de la tâche, nous nous sommes aperçus que près de 20% des dialogues, dont la tâche à été annotée comme réussie, se révèlent avoir échoué. Nous avons récupéré un échantillon de trois cents dialogues de DSTC2, soit 10% de la taille totale du corpus. Sur ces trois cents dialogues, 63 étaient annotés comme des succès sans en être. Pour réaliser cette vérification, deux linguistes ont annoté les trois cents dialogues séparément. Les critères de succès de la tâche, conformément au manuel

d'annotation de DSTC2, étaient les suivants :

— informations données par l'utilisateur bien comprises par le système
— informations demandées par l'utilisateur bien fournies par le système

Nous avons ajouté empiriquement plusieurs critères permettant objectivement de déterminer le succès du dialogue :

— Le nombre d'incohérences du système
— Le taux d'exaspération de l'utilisateur (jurons, réflexions...)
— Le nombre de tours de parole (faisant partie des coûts du dialogue décrits par [Walker, 1996] et l'outil PARADISE [Walker et al., 1997b])

L'accord entre les deux linguistes a été mesuré avec le Kappa de Cohen [Cohen, 1960] obtenant un score inter-annotateurs de $\kappa = 0.78$. Ce score relativement peu élevé montre la difficulté même pour des experts d'unifier les critères permettant d'affirmer ou non le succès de la tâche. Il suffit pour s'en convaincre de lire le manuel d'annotation de DSTC2 [1] : *feedback :*

— succès : le succès subjectif du dialogue (booléen)
— commentaire : les impressions de l'utilisateur sur la qualité de l'échange (chaîne de caractères)
— questionnaire : une éventuelle liste de questions-réponses pour améliorer le système.

Cependant, malgré ses défauts, DSTC2 est encore aujourd'hui utilisé pour l'évaluation de nombreux modèles de système de dialogue comme en attestent les bancs d'essai [Mi et al., 2021].

5.1.2 MultiWOZ erreurs et spécificités

MultiWOZ, lui, est un corpus de dialogue de l'université de Cambridge datant de 2018 où le POMDP est remplacé par un humain. L'utilisateur, lui, ne sait pas qu'il dialogue avec un humain, il pense dialoguer avec une machine.
Contrairement au corpus DSTC2, MultiWOZ est multi-domaines, c'est-à-dire qu'au sein du corpus, on trouve des dialogues traitant des sujets différents, mais cela est également possible au sein d'un même dialogue. C'est la majorité des cas puisque sur dix mille dialogues, plus de sept mille sont multi-domaines.
Tout comme DSTC2, MultiWOZ contient des annotations sur l'état du dialogue à un tour donné, ainsi que sur les actes de dialogue sous-jacents aux réponses du système. Cependant, MultiWOZ originel ne contient pas d'annotation sur les actes de dialogue de l'énoncé de l'utilisateur. Celles-ci ont été ajoutées ultérieurement par [Eric et al., 2019].
Par ailleurs, MultiWOZ n'inclut que des dialogues dont la tâche a été un succès, en tout cas pour l'évaluation [2]. Il est également précisé que tous les dialogues sont cohérents, ce qui implique que le succès du dialogue ainsi que le ressenti de

1. https ://github.com/matthen/dstc/blob/master/handbook.pdf page 17 A.2 *feedback*
2. https ://github.com/budzianowski/multiwoz

l'utilisateur est positif dans la grande majorité des cas. Les principales erreurs dans MultiWOZ proviennent de l'impossibilité de l'agent de mener à bien une sous-tâche dans le dialogue (réservation d'une chambre d'hôtel, sélection d'un restaurant correspondant à tous les critères de l'utilisateur...), impliquant des erreurs dans les annotations en état de croyance.
Cependant, dans la mesure où MultiWOZ est lui aussi annoté en but, on peut en déduire le succès ou pas de la tâche avec les mêmes critères que pour DSTC2.
Ce n'est pas le seul point commun entre DSTC2 et MultiWOZ. En effet, hormis les dialogues eux-mêmes, chaque corpus contient une base de données créée dans le but de permettre au système d'apprendre à aller y chercher les informations demandées par l'utilisateur. Nous nous sommes demandés si certains éléments de la base de données de l'un se trouvait dans celle de l'autre. DSTC2 ne contenant que des informations sur des restaurants, nous avons observé la base de données de restaurants de MultiWOZ, et sans surprise, l'intégralité des restaurants de la base de données de DSTC2 se trouvent à moins de trois kilomètres de l'université de... Cambridge !
Concernant les deux autres corpus qui nous intéressent, le CamRest676 et le Schema-Guided Dialogue dataset (SgD), les annotations sont presque les mêmes, à cela près que pour SgD, l'information sur le but du dialogue n'est pas fournie, ce qui rend l'évaluation du succès de la tâche plus compliquée. La base de données de CamRest est sensiblement la même que celle de DSTC2, tandis que SgD n'en contient pas. Une base de données artificielle peut cependant être générée en récupérant les résultats d'API, appelés "service" dans le corpus, et en partant du postulat que l'intégralité des résultats d'API du corpus représente l'intégralité des éléments disponibles dans une hypothétique base de données.

5.2 Construction de nouvelles ressources

Nous avons décidé en collaboration avec des collègues de l'université de Prague, d'étudier l'unification de ces corpus en un unique jeu de données de dialogue orientés-tâche, que nous avons baptisé DIASER. La motivation est double. Tout d'abord, nous voulions créer une base de données et une ontologie unifiées, permettant à un système de mieux généraliser les tâches de pistage d'état de génération de réponse pour un ensemble de domaines, en particulier celui du restaurant qui est présent dans les quatre corpus originaux. Ensuite, nous désirions vérifier si les incohérences "naturellement" présentes dans DSTC2 peuvent servir à prédire des incohérences dans les réponses du système générées par un modèle entraîné sur l'ensemble des quatre corpus.
Dans le domaine de la vision, la fusion de données a montré une certaine efficacité pour la conduite automatique [Yang et al., 2018] et la détection d'objets dans la

nature [Rame et al., 2018].
L'idée d'unifier des ensembles de données multi-domaines a été appliquée dans de nombreuses branches du traitement de la langue naturelle. [Gao and Zhang, 2005] a fusionné trois ensembles de données pour obtenir une extraction efficace de textes. Le projet OPUS ([Tiedemann and Nygaard, 2004]) consiste en la fusion de dizaines de corpus différents et est toujours actif aujourd'hui. Le projet Universal Dependencies [Nivre et al., 2016] unifie l'annotation syntaxique de plus de cent corpus et de nombreuses langues.
[Fortuna et al., 2018] a fusionné avec succès différents ensembles de données de réseaux sociaux pour réaliser une identification des écrits haineux. Dans le domaine de la détection des émotions, [Bostan and Klinger, 2018] a obtenu des résultats proches de l'état de l'art en fusionnant des ensembles de données. [Acedo et al., 2018] a proposé un modèle pour la fusion de jeux de données ainsi qu'une méthode pour détecter les jeux de données textuelles similaires [Acedo et al., 2019] et pour identifier les mots représentatifs de chaque jeu de données [Fernández-Sellers et al., 2019].
Malgré quelques efforts théoriques de normalisation des annotations [Bunt et al., 2020], les tentatives d'unification des données de dialogue se sont surtout limitées à fournir des données pour différentes langues selon un schéma commun, sans tentative de fusion des ensembles de données [Chen and Kan, 2013, Noh et al., 2015]. Récemment, ConvLab-2 [Zhu et al., 2020b] a mis à disposition un ensemble de données d'environ 106 000 dialogues qui fusionne quatre ensembles de données de dialogue orientés vers des tâches différentes [Serban et al., 2018] : CamRest [Wen et al., 2016a], MultiWOZ [Budzianowski et al., 2018a], DealOrNoDeal [Lewis et al., 2017] et CrossWOZ [Zhu et al., 2020a] dans une boîte à outils de dialogue avec un support logiciel pour les expériences multi-domaines, mais sans schéma d'annotation commun unifié, qui est l'objectif principal de notre travail. De plus, dans ConvLab-2, les sous-corpus MetaLWOz [Shalyminov et al., 2020] et Taskmaster [Byrne et al., 2019] (70 686 dialogues) ne sont pas utilisables pour notre objectif car le premier n'a pas d'annotation et le second a été construit en utilisant une procédure avec compère, avec deux parties : un sous-corpus parlé et un sous-corpus écrit, avec le compère jouant les deux rôles en même temps. Comme nous nous concentrons sur les dialogues humain-machine, la partie exploitable de ConvLab-2 est donc de taille similaire au corpus que nous présentons ici, le nôtre étant légèrement plus grand d'environ deux mille dialogues.

5.2.1 Fusion de corpus libres

La conception de l'ensemble de données DIASER[3] n'est pas un processus anodin dans lequel nous aurions simplement combiné les différents ensembles de données. Il y a eu des problèmes techniques que nous ne discuterons pas dans cette thèse, mais aussi des questions plus théoriques, comme le format final qui doit être conforme au modèle de dialogue de 2.1.1 et compatible avec les différentes ontologies des quatre corpus initiaux. Le tableau 5.1 résume les statistiques des quatre corpus utilisés pour la fusion.

Data	SgD	MultiWOZ	DSTC2	CR	Total
Domaines	18	7	1	1	19*
Entités	145	29	10	7	166*
Dialogues**	22,8	10,4	3,2	0,7	37,1
Tours**	463,3	143,0	51,0	5,5	662,8
Tours par Dial.	20,30	13,71	15,77	8,12	17,83
taille moy. tour	9,86	13,23	8,47	10,71	10,49
Hapax**	32,3	23,2	1,3	1,7	49,9
Shannon entr.	8,96	8,54	7,04	7,69	9,01
Cond. entr.	4,76	4,41	2,14	2,95	4,83

TABLE 5.1 – Composition de notre corpus avec quelques statistiques
*les doublons sont ignorés
** en milliers

L'ensemble du processus de fusion se compose de plusieurs étapes consécutives. Elles sont décrites succintement dans la liste ci-dessous et ensuite détaillées plus loin dans le texte.

Processus de fusion Dans le but de fusionner ces quatre corpus, nous avons mis en place un processus complet de transformation des données :

— **Prétraitement de DSTC2 et du CR pour correspondre au schéma du SgD**. Le problème de DSTC2 et CamRest est dû au fait que les annotations de l'état de croyance sont extraites à la fois de l'énoncé de l'utilisateur et du système, alors que dans le jeu de données MultiWOZ et SgD, l'état de croyance est seulement extrait de l'énoncé de l'utilisateur. Nous avons dû filtrer automatiquement les annotations de DSTC2 et CamRest jusqu'à ce qu'elles correspondent à la représentation de l'état de croyance de MultiWOZ.

— **Ajout de but, d'incohérence et d'erreurs d'annotation** à partir des jeux de données originaux. DSTC2, CamRest et MultiWOZ contiennent le but du dialogue (également appelé tâche) sous la forme d'un acte de

3. La ressource, ainsi que les schémas d'annotation, les utilitaires python et toutes les expériences associées seront disponibles en libre accès sur Github à partir du 5 avril 2022

dialogue contenant les contraintes (p.e. restaurant cher au sud) et les informations que l'utilisateur doit demander au système (p.e. numéro de téléphone et adresse). Ils contiennent également un petit texte résumant cet objectif (p.e. Vous voulez un restaurant turc dans le sud de la ville et il doit être dans la gamme de prix élevés. Vous demandez alors le numéro de téléphone et l'adresse). Nous ajoutons également, pour DSTC2, les incohérences que nous avons décrites dans 4.4. Enfin, par rapport à MultiWOZ 2.3 [Han et al., 2020b] et MultiWOZ 2.4 [Ye et al., 2021b], la version que nous utilisons, MultiWOZ 2.2, contient plusieurs erreurs d'annotation. Nous ne les avons pas corrigées en copiant les corrections des versions postérieures, mais nous avons ajouté les corrections comme un objet *Corrigé* dans les méta-données. De même, nous avons récupéré dans la version plus ancienne MultiWOZ 2.1 les erreurs d'annotations (informations superflues ou manquantes) pour enrichir chaque tour de parole d'une balise *Erreurs d'annotation*.

— **Convertir tous les ensembles de données** dans notre structure d'annotation. La structure que nous proposons est similaire à la structure SgD et MultiWOZ 2.2, mais nous abandonnons les cadres sémantiques [El Asri et al., 2017] que nous considérons trop lourds pour notre objectif, et créons un objet *Tour* qui contient soit l'énoncé de l'utilisateur et les actes de dialogue, soit ceux du système. Deux *Tours* différents peuvent partager le même numéro de tour, nous considérons qu'un *Tour* est un énoncé unique et non un échange entre l'utilisateur et le système (c'est-à-dire une paire d'énoncés).

— **Pour MultiWOZ, générer des annotations NLU** à partir des actes de dialogue : nous créons également un objet *NLU* qui contient tous les quadruplets domaine-intention-entité-valeur extraites de l'énoncé du locuteur. Dans les données originales de MultiWOZ2.2, les actes de dialogue du système sont dans un ensemble de données séparé (nous ne savons pas pourquoi), nous avons donc dû les fusionner.

— **Unifier** les noms d'entités, de domaine et d'intention pour les raisons expliquées dans 5.2.2.

— **Concaténer** toutes les données traitées ensemble et préparer les ensembles d'entraînement, de développement et de test.

5.2.2 Fusion d'ontologies et de bases de connaissances

La difficulté lors de la fusion d'ensembles de données est non seulement de trouver un modèle de dialogue commun qui sera utilisé comme abstraction pour unifier les sous-ensembles de données, mais aussi d'unifier les différentes ontologies. Nous définissons les ontologies des ensembles de données par les différents objets (entités) et attributs (valeurs) qui sont contenus dans les actes de dialogue. Plus

précisément, les ontologies des corpus contiennent tous les domaines possibles pour le dialogue, toutes les entités possibles pour chaque domaine, et toutes les valeurs possibles pour chaque entité. Il s'agit d'un triplet domaine-entité-valeur. Mais nous ne pouvons pas considérer les entités indépendamment du domaine auquel elles appartiennent. En effet, une entité qui représente la *gamme de prix* ne correspondra pas au même ensemble de valeurs s'il s'agit d'un restaurant ou d'un billet d'avion. Nous avons identifié deux problèmes principaux lors de l'analyse des ontologies des différents jeux de données.

1. **Mêmes valeurs, entités différentes, mêmes entités, valeurs différentes**
   Dans le jeu de données MultiWOZ, il existe un nom d'entité *day*, et un nom d'entité *book-day* mais les deux contiennent les mêmes valeurs (chaque jour de la semaine). Le contexte dans lequel les entités sont utilisées est le même, par exemple pour *réservation de restaurant* ou *réservation d'hôtel*. D'autre part, dans SgD, nous pouvons trouver une entité *nommé start-day* et un autre appelé *day* mais dans le premier cas, il s'agit d'une date et dans le second, du jour de la semaine.
   Dans MultiWOZ encore, on trouve une entité nommée *leave-at*, et indépendamment du domaine, sa valeur est toujours une plage horaire, mais l'entité symétrique appelée *arrive-by* peut être soit une plage horaire, soit un jour de la semaine selon le domaine.
   Dans le jeu de données SgD, toutes ces entités ont exactement les mêmes valeurs : *nombre de billets, nombre d'adultes, personnes, nombre total de personnes, personnes, nombre de personnes, taille du groupe, nombre de sièges*. Il n'existe pas d'ontologie claire qui unifie toutes les entités à valeurs identiques en un seul nom.
2. **SgD n'a ni ontologie ni base de données** L'objectif du SgD était de construire des modèles capables de récupérer les appels API et la base de données utilisée par le système à partir des résultats API. Par conséquent, aucune métadonnée n'a été publiée avec le corpus, ce qui nous a obligés à créer une ontologie et une base de données uniquement à partir des données.

Bien que nous pensions que ces quatre corpus seraient très similaires, leur fusion n'est pas une tâche évidente, en raison des différences qu'ils présentent à la fois dans la conformité du modèle de dialogue et dans l'ontologie. Aucun d'entre eux ne respecte réellement la norme ISO d'annotation des actes de dialogue [Bunt et al., 2020]. L'une des parties les plus difficiles de ce travail a été d'unifier des ontologies qui n'étaient pas construites sur les mêmes dimensions.
Tout d'abord, les ontologies DSTC2 et CamRest distinguent deux types d'entités : les entités renseignées et les entités renseignables.

— **Entités renseignées** : Ce sont les entités dont la valeur est connue pendant le dialogue parce qu'elle est donnée par l'utilisateur, par exemple : *Je veux*

Intent	SgD	MultiWOZ	DSTC2	CR	all
inform	151,467	84,259	33,451	5,786	274,963
request	81,786	31,888	13,620	1,516	128,810
offer	67,095	4,497	23,763	0	95,355
bye	35,089	12,380	0	0	47,469
else	31,030	13,779	0	0	44,809
select	33,820	2,834	243	0	36,897
confirm	30,327	0	5,756	0	36,083
thank	23,172	9,064	0	0	32,326
negate	132,01	4,399	4,413	0	22,023
total	435,957	163,100	81,346	7,312	718,645

TABLE 5.2 – Intentions les plus fréquentes dans notre schéma, par source et au total

un restaurant bon marché qui sert de bons sushis[4].

— **Entités renseignables** : ce sont les entités dont l'utilisateur veut obtenir la valeur, par exemple *Pouvez-vous me donner le téléphone s'il vous plaît ?*.

Deuxièmement, il y avait des noms d'entités ambigus tels que *nom* et *type*. Par exemple, pour le domaine des taxis, l'entité appelée *nom du taxi* dans MultiWOZ avait les mêmes valeurs que l'entité *taxi-type* dans SgD. Pour le domaine des restaurants, l'entité *restaurant-type* dans MultiWOZ était le même que *food-type* dans DSTC2, cependant, le *restaurant-type* de SgD se référait au type de restaurant (restaurant, bar, fast-food, izakaya) et non au type de cuisine.
Troisièmement, comme nous l'avons dit précédemment, le corpus de SgD ne contient aucune ontologie, nous avons donc dû la construire à partir de zéro. Le problème est que dans ce corpus, il existe plusieurs schémas par domaine, et plusieurs dialogues par schéma. Chaque schéma possède ses propres noms d'entités, mais dont les valeurs sont les mêmes (par exemple "number_of_people" et "number_of_person"). Dans d'autres cas, il existe une ambiguïté liée au nom des entités. Pour le domaine des vols, dans certains schémas, nous avons les entités *original-airport,departure-airport,origin-city,departure-city*, mais dans un autre schéma, nous trouvons *origin*. Il n'existe aucune règle que nous pourrions appliquer afin de savoir si *origin* fait référence à l'aéroport ou à la ville.

Enfin, nous voulons aborder les erreurs d'annotation dans les actes de dialogue concernant les valeurs des entités. Par exemple l'entité *start-time* peut prendre la valeur **[trente minutes plus tôt]** ou l'entité *hotel-name* peut prendre la valeur **[le premier que vous avez mentionné]**, cela devient un problème lorsque nous essayons de faire correspondre les actes de dialogue des données avec l'ontologie unifiée, qui est censée contenir toutes les informations du corpus.

4. (nous pourrions discuter de l'incohérence de cette demande, mais il existe peut-être une ville dans le monde où les bons sushis sont bon marché)

entité	SgD	MultiWOZ	DSTC2	CR	all
name	65,999	42,215	3,135	2	111,351
area	8,133	48,285	37,697	4,178	98,293
date	81,600	20,872	0	0	102,472
price	1,950	33,084	32,267	4,032	71,333
type	41,146	28,562	0	0	69,708
leave	32,625	26,684	0	0	59,309
arrive	26,180	26,228	0	0	52,408
food	13,501	20,963	3,007	571	38,042
book	17,618	11,723	0	0	29,341
total	298,752	232,388	76,106	8,783	532,257

TABLE 5.3 – Entités les plus fréquentes dans notre schéma, par source et au total

Intent	SgD	MultiWOZ	DSTC2	CR	all
inform	2,671	6,608	33,451	5,486	48,216
offer	707	0	23,763	0	24,470
request	1,615	2,275	13,620	1,516	19,026
confirm	819	0	5,756	0	6,575
bye	744	2,023	0	0	2,767
thank	496	750	0	0	1,246
negate	259	733	4,413	0	5,405
select	302	305	243	0	850
total	7,613	12,694	81,246	7,002	108,555

TABLE 5.4 – Intentions les plus fréquentes dans le domaine du restaurant

entité	SgD	MultiWOZ	DSTC2	CR	all
area	146	4,315	37,697	4,178	46,336
price	103	4,278	32,267	4,032	40,680
food	2,564	4,494	3,007	571	10,636
name	2,811	2,594	3,135	2	8,542
postcode	0	0	2,499	610	3,109
phone	0	0	2,317	503	2,820
book	513	1,424	0	0	1,937
count	367	1,446	888	0	2,701
date	513	1,436	0	0	1,949
total	7,017	19,987	81,810	9,896	118,710

TABLE 5.5 – Entités les plus fréquentes dans notre schéma, par source et au total.

5.2.3 Impact de la fusion évalué avec des modèles de référence

Nous choisissons de comparer les performances de deux méthodes de modélisation de dialogue : MarCo et GPT-2. Nous les évaluons à la fois par rapport

au pistage d'état et aux mesures de qualité du langage sur le corpus unifié. Nous essayons diverses combinaisons entraînement-test pour comprendre l'influence de chaque partie des données sur le résultat. Notez que dans nos expériences, les données de CamRest676 ont toujours été incluses dans l'ensemble d'entraînement et n'ont pas été utilisées pour l'évaluation car leur taille est trop petite. Le but est de prouver que l'unification de données telle que nous l'avons réalisée permet de produire des modèles de dialogue plus robustes et performants que si nous produisions un modèle de dialogue par corpus.

Modèles utilisés

MarCo [Wang et al., 2020c] est un modèle qui génère conjointement les actes de dialogue et la réponse du système dans le même réseau. Il traite la séquence d'actes de dialogue générée comme un plan sémantique pour la réponse finale. MarCo utilise l'état de croyance comme une caractéristique, il dépend donc d'un pisteur d'état externe, ce qui constitue sa principale limitation. L'architecture est inspirée du modèle à deux étapes appelé HDSA (Hierarchical disentangled self-attention) [Chen et al., 2019a], un transformeur avec un mécanisme de self-attention mais où les têtes ne sont pas toutes activées au moment de concaténer les vecteurs d'une couche en un seul. MarCo diffère de HDSA car il calcule conjointement l'acte de dialogue et la génération de réponse au lieu d'un calcul en deux étapes, et améliore les résultats par rapport à HDSA. [5]

GPT-2 Ces dernières années, l'utilisation de modèles de langue (LMs) pré-entraînés s'est largement développée. Ces architectures sont principalement basées sur l'architecture Transformer [Vaswani et al., 2017] et ont une certaine capacité d'apprentissage de modèle de langue à partir de corpus de taille exceptionnelle. L'architecture GPT [Radford et al., 2019] fait partie des approches les plus utilisées et a été appliquée avec succès à une grande variété de tâches dans le domaine du TAL. Le GPT utilise le composant décodeur du Transformer qui représente essentiellement un LM auto-régressif. Le décodeur pré-entraîné peut être ajusté sur n'importe quelle tâche en aval exprimée comme un problème de séquence à séquence. Récemment, des travaux ont appliqué l'architecture GPT-2 à la tâche de modélisation du dialogue : [Peng et al., 2020, Zhang et al., 2020c, Kulhánek et al., 2021]. Nous suivons l'approche [Peng et al., 2020] et utilisons le modèle pré-entraîné à la fois pour le pistage des états de croyance et la génération de réponses. Notre approche comporte deux étapes. Tout d'abord, l'état de croyance est généré sur la base du contexte du dialogue. L'état de croyance est décodé mot par mot et nous entraînons le modèle de sorte que la séquence décodée puisse être analysée de manière déterministe. Ensuite, nous effectuons une recherche dans la

5. Les codes sources de MarCO et de HDSA sont disponibles en ligne : https://github.com/InitialBug/MarCo-Dialog.

base de données sur la base de l'état de croyance analysé. Puis nous concaténons le contexte, l'état de la croyance et un résumé du résultat de l'entropie consultation de la base de données et nous générons la réponse du système avec le même modèle. Nous travaillons avec des versions délexicalisées des énoncés du système. Nous évaluons les performances des modèles à l'aide d'un ensemble de mesures basées sur le corpus. Notre jeu de données se concentre sur les systèmes orientés-tâche, nous évaluons donc la précision des états de dialogue générés qui sont essentiels pour l'interaction avec la base de données et l'exactitude du système. Pour évaluer les états, nous utilisons des mesures classique : *exactitude jointe* et *f-mesure*. L'**exactitude jointe** donne le pourcentage de correspondance exacte sur les états de dialogue, c'est-à-dire qu'elle reflète combien de dialogues ont un état correctement prédit. Au contraire, la f-mesure est calculée sur le nombre et la précision des entités correctement détectées par le système dans l'énoncé de l'utilisateur. On la nomme **f-mesure d'entité**. En outre, nous calculons le score **BLEU** [Papineni et al., 2002b] entre les énoncés du système généré et la vérité de base pour obtenir une approximation de la fluidité du texte produit. Le score BLEU est calculé sur des versions délexicalisées des énoncés, ce qui correspond au schéma d'évaluation commun. Bien que l'utilisation du score BLEU pour l'évaluation des systèmes de dialogue soit controversée, nous décidons de l'utiliser car il est couramment rapporté sur des ensembles de données orientés-tâche et pour mesurer la fluidité de la sortie.

apprentissage			évaluation			mesures					
						f-mesure d'entité		**exactitude jointe**		**BLEU**	
DSTC2	**MultiWOZ**	**SgD**	**DSTC2**	**MultiWOZ**	**SgD**	MarCo[†]	GPT	MarCo [†]	GPT	MarCo [†]	GPT
	✓		✓			0,85	0,47	0,45	0,11	10,47	17,90
		✓	✓			0,51	0,19	0,05	0,01	3,66	4,11
✓	✓		✓			0,46	0,84	0,05	0,54	10,01	46,31
✓		✓	✓			0,56	0,69	0,18	0,26	23,46	43,47
	✓	✓	✓			**0,88**	0,46	**0,57**	0,11	6,33	16,55
✓	✓	✓	✓			0,50	0,85	0,13	0,36	27,31	**46,47**
	✓			✓		0,70	0,89	0,48	0,53	17,05	18,61
		✓		✓		0,46	0,16	0,11	0,02	2,87	4,01
✓	✓			✓		0,74	0,89	0,49	**0,55**	16,17	19,67
✓		✓		✓		0,51	0,17	0,15	0,03	4,27	5,68
	✓	✓		✓		0,64	0,89	0,39	0,52	13,19	19,92
✓	✓	✓		✓		0,65	**0,90**	0,38	0,54	14,59	**21,09**
	✓				✓	0,42	0,04	0,04	0,01	2,97	5,63
		✓			✓	0,68	0,59	0,19	0,21	9,72	**28,17**
✓	✓				✓	0,51	0,03	0,10	0,01	2,97	5,51
✓		✓			✓	**0,85**	0,58	**0,52**	0,21	7,49	27,96
	✓	✓			✓	0,71	0,63	0,17	0,23	6,17	27,54
✓	✓	✓			✓	0,77	0,63	0,32	0,22	1,72	27,72
✓	✓		✓	✓	✓	–	0,28	–	0,12	–	15,30
✓		✓	✓	✓	✓	0,62	0,55	0,23	0,22	3,95	27,28
	✓	✓	✓	✓	✓	**0,79**	0,65	**0,40**	0,25	8,70	25,13
✓	✓	✓	✓	✓	✓	0,65	0,70	0,20	0,28	14,49	**29,73**

TABLE 5.6 – Performance de nos modèles entraînés sur la combinaison de corpus. [†] état du dialogue en temps qu'indice supplémentaire

Analyse des erreurs

Le tableau des résultats 5.6 résume nos différentes expérimentations.

GPT-2 Nous pouvons observer une tendance générale dans les résultats qui suggère que les modèles de réseaux neuronaux actuels ne parviennent pas à généraliser sur différents ensembles de données lorsqu'un sous-ensemble de données sur lequel nous évaluons n'est pas inclus dans un ensemble d'entraînement (en termes de SgD, MultiWOZ, DSTC2 et CR). Une baisse significative des performances peut être observée pour toutes les métriques enregistrées. Les résultats individuels sont détaillés dans les paragraphes ci-dessous.

Nous pouvons observer que l'omission de DSTC2 des données d'entraînement conduit à une performance assez mauvaise. En termes de score de f-mesure des entités et d'exactitude jointe, l'explication de la mauvaise précision du modèle est à attribuer dans une distribution significativement différente des entités entre les jeux de données SgD/MultiWOZ et DSTC2, comme on peut le voir dans les tableaux 5.3 et 5.5. Plus précisément, lorsqu'il est entraîné uniquement sur SgD, le modèle est capable de n'apprendre presque rien. D'autre part, l'entraînement sur MultiWOZ obtient de bons scores sur la f-mesure d'entité et l'exactitude jointe. Il semble également que la concaténation de SgD à MultiWOZ n'apporte qu'une amélioration très marginale. Une fois que DSTC2 est inclus dans les données d'entraînement, les performances entre les différents sous-ensembles de données sont similaires. Le score de f-mesure le plus élevé (0,72) et l'exactitude jointe (0,17) sont obtenus avec les données d'entraînement DSTC2+MultiWOZ, tandis que les modèles formés sur les données DIASER sont les plus performants en termes de score BLEU (46,61).

L'analyse des performances du modèle évalué sur MultiWOZ est assez semblable. L'entraînement du modèle uniquement sur SgD nous donne un modèle qui n'est pas capable de généraliser sur les données MultiWOZ. La concaténation des jeux de données MultiWOZ avec SgD, DSTC2 ou les deux conduit à une légère amélioration, principalement en termes de score BLEU. Dans ce cas, le modèle obtient un score BLEU de 12,29 lorsque le GPT-2 est entraîné sur MultiWOZ. Nous obtenons les meilleurs résultats en utilisant l'ensemble des données DIASER pour l'entraînement et nous obtenons le modèle avec de meilleures capacités de génération reflétées par le score BLEU de 13,53.

L'évaluation du modèle sur les données SgD suit le même schéma que celui décrit pour DSTC2 et MultiWOZ. Dans ce cas, il est assez intéressant d'observer que l'entraînement du modèle ne semble pas tirer profit de l'ensemble de DIASER comme il a pu le faire pour l'évaluation de DSTC2 et de MultiWOZ.

Enfin, si nous utilisons l'ensemble des données DIASER de test pour l'évaluation, il est clair que le modèle le plus performant est obtenu lorsque utilisons

également l'ensemble des données d'entraînement (DSTC2 + MultiWOZ + Camrest676 + SgD). D'après les résultats présentés dans le tableau 5.6, nous pouvons voir que l'inclusion des données SgD est cruciale pour obtenir des valeurs plus élevées de BLEU, tandis que l'entraînement sur MultiWOZ aide le modèle à prédire les valeurs des entités.

MarCo Les résultats du modèle MarCo sont plus difficiles à interpréter. Nous constatons souvent un comportement incohérent par rapport aux données d'entrée. En général, MarCo surpasse souvent le modèle GPT-2 en termes de f-mesure des entités.
Cependant, il semble qu'il ne soit pas capable d'utiliser cette connaissance pour générer avec succès des réponses significatives. La raison de ce comportement pourrait être que MarCo utilise l'état de croyance comme une caractéristique d'entrée et obtient ainsi de bons résultats lorsqu'il est évalué avec la f-mesure d'entité et l'exactitude jointe, mais il a du mal à obtenir des réponses significatives. Le modèle MarCo semble également être très vulnérable aux incohérences des réponses du système présentes dans DSTC2. Ce phénomène a été analysé et discuté dans la communauté [Schaub et al., 2021].

Il se peut également que le modèle MarCo soit beaucoup plus sensible au choix des hyperparamètres, qui doit donc être fait avec soin pour chaque ensemble de données, voir la section 5.2.3.

5.2.4 Conclusion

En résumé, nous observons que chaque sous-ensemble de données a ses propriétés spécifiques qui ne sont pas interchangeables entre les sous-ensembles. Cependant la combinaison des données d'entraînement avec des exemples supplémentaires est profitable. Dans l'ensemble, nous pouvons dire que le modèle GPT fournit des résultats plus cohérents que MarCo, mais qu'il a du mal à réaliser le pistage d'état du dialogue avec une ontologie complexe et une annotation de qualité inférieure, ce qui est le cas de SgD.

5.3 Incohérences

L'incohérence se définit par "Manque d'accord, d'unité, de lien logique entre des parties, entre les éléments d'un ensemble" [6].

Pour le dialogue, la cohérence peut se traduire par une forme d'unité, d'équilibre linguistique entre les énoncés des deux participants et à l'intérieur de chacun des énoncés. L'incohérence, peut venir du manque d'équilibre dans cet échange ou

6. https ://www.cnrtl.fr/definition/incohérence

au sein d'un énoncé. Nous avons proposé en 4.4 une typologie des incohérences dans le dialogue humain-machine. Cependant, dans cette typologie, nous ne montrons que les incohérences du système. Or, comme nous en discutons en 6.5.2, les incohérences peuvent également venir des énoncés de l'utilisateur, ou de l'échange entre les deux, et peuvent chevaucher plusieurs tours de parole avant d'être identifiées. Nous proposons ici différentes méthodes pour détecter les incohérences.

5.3.1 Corrélation entre incohérences et échec du dialogue

Nous avons calculé la corrélation entre les dialogues marqués comme des échecs et l'occurrence des incohérences. Sans surprise, presque tous les dialogues en échec contiennent des réponses incohérentes du système. Le test exact de Fisher [Fisher, 1936] montre qu'il existe une dépendance très probable entre les dialogues ayant échoué et la présence d'incohérences – les dialogues comportant des incohérences sont environ trois fois plus susceptibles d'échouer (odds ratio 0,0328, $p <$ 1e-20). Nous avons cherché à savoir quelles étaient les caractéristiques déterminantes pour décider si un dialogue était un échec ou non. Nous avons utilisé un modèle de classification simple (Bayésien naïf gaussien) [Chen et al., 2009] de l'outil Scikit-Learn [Pedregosa et al., 2011] pour prédire le succès du dialogue.

Le tableau 5.7 résume certaines des différentes caractéristiques. utilisées pour améliorer la détection des dialogues ayant échoué. Nous calculons le score f-mesure des dialogues infructueux (les cas infructueux sont traités comme positifs). Les meilleurs résultats sont obtenus avec une vectorisation des textes basée sur le TF-IDF, couplée au nombre d'incohérences totales et au nombre d'incohérences apparaissant avant le premier appel API du système dans le dialogue. Les résultats

indices	precision	rappel	f-mesure
texte	0,56	0,52	0,53
texte+nbr incohérences	0,57	0,62	0,60
texte+nbr cohérences+incohérences arrivant avant l'appel API	**0,65**	0,57	**0,61**

TABLE 5.7 – Prédictions des échecs des dialogues

confirment que l'information sur la présence ou non d'une incohérence rend la prédiction d'échec du dialogue plus efficace.

5.3.2 Tâche de détection d'incohérence

Nous avons mené deux séries d'expériences séparées sur la détection des incohérences : l'une de façon binaire et l'autre de façon multi-classe. Le détecteur binaire précise s'il existe une incohérence dans le tour de parole courant (en rejetant systématiquement la faute sur le système). Le détecteur multi-classe, quant à lui, tente de prédire quelle incohérence a été détectée. Dans le cas où il existerait

plusieurs incohérences au sein du tour de parole courant, nous conservons la première. Il y a au moins deux autres configurations que nous n'avons pas traitées : la tâche de prédiction du nombre d'incohérences, sous forme soit de classification (de 0 au nombre maximal d'incohérences pour un seul tour de parole) soit de régression, ainsi que la tâche multi-étiquettes, la plus difficile, où il faut prédire toutes les incohérences d'un tour de parole avec leur type. Notre annotation automatique basée sur des règles (voir Section 4.4) utilise l'ensemble du dialogue annoté pour détecter et classer les incohérences, c'est-à-dire qu'elle peut décider en fonction du contexte futur sur le tour de parole courant contient une incohérence. Cependant, dans les cas d'utilisation réels, nous ne pouvons observer que l'historique du dialogue existant au moment où nous le regardons. Par conséquent, nous nous demandons s'il est possible d'obtenir les mêmes performances en formant un modèle de classification avec moins d'informations à sa disposition. Nous avons entraîné quatre classifieurs différents sur notre annotation pour prédire les incohérences :

— **Bi-LSTM avec attention** [Jang et al., 2020] est un modèle simple pour la classification séquence/texte mais assez efficace lorsqu'il doit traiter des informations à long terme [Hochreiter and Schmidhuber, 1997], comme l'historique des dialogues.
— **DIET classifier**, le transformeur d'entités et d'intentions de dialogue (*Dialogue Intent and Entities Transformer*), est un classifieur d'intentions de dialogue basé sur le transformeur [Devlin et al., 2019] [Wu et al., 2020a] qui surpasse la plupart des classifieurs récents dans la tâche de détection des intentions de l'utilisateur. Il est utilisé dans le cadre de RASA [Bocklisch et al., 2017b].
— **WMM2Seq** [Chen et al., 2019c] est un modèle basé sur un réseau de mémoire qui utilise deux modules de mémoire différents : le contexte (l'historique du dialogue en tant que mémoire épisodique) et la base de connaissances (les appels API en tant que mémoire sémantique) pour générer les réponses du système, un mot à la fois.
— **GPT-2** [Radford et al., 2019] est une architecture à base de transformeurs composée de plusieurs blocs de décodeurs de transformeurs [Vaswani et al., 2017], empilés les uns sur les autres. L'architecture est pré-entraînée pour la modélisation de la langue sur un énorme corpus et est capable d'un réglage fin efficace pour de nombreuses tâches en aval. Nous affinons le modèle dans un cadre multi-tâche, c'est-à-dire que nous optimisons à la fois la perte de classification des incohérences et la perte de génération de réponses.

Nous utilisons l'exactitude et la moyenne macro pondérée des scores f-mesure comme mesures d'évaluation, et nous entraînons les modèles d'abord pour la classification binaire (incohérence ou non) puis pour la classification multi-classe (prédiction d'un type d'incohérence spécifique ou de l'absence d'incohérence). Nous utilisons la prédiction de l'étiquette la plus fréquente pour obtenir une valeur de

référence (aucune incohérence, présente dans 87% des exemples, soit une exactitude de 87%). Les résultats sont présentés dans le tableau 5.8 et discutés ci-après. Nous utilisons 2 117 dialogues pour l'entraînement et 1 115 pour les tests.

	Bi-LSTM				**DIET**				**WMM2Seq**				**GPT-2**			
	binaire		multi		binaire		multi		binaire		multi		binaire		multi	
combinaison	exa	F1	exa	F1	exa	F1	exa	F1	exa	F1	exa	F1	exa	F1	exa	F1
c	59,0	52,3	83,9	9,53	85,7	66,6	83,4	52,6	86,0	50,1	83,7	35,7	64,2	77,2	62,1	8,1
c+h_1	80,5	65,0	84,0	10,2	85,2	58,4	83,1	43,0	86,9	48,1	82,6	41,8	84,3	90,5	84,2	38,8
c+h_2	77,0	59,8	84,8	9,18	83,2	55,2	82,6	50,0	**87,0**	46,2	82,6	39,6	85,2	91,0	85,1	40,7
c+h_f	71,0	48,9	85,4	6,57	79,3	53,7	83,6	50,7	85,9	44,4	82,6	44,6	99,9	99,9	**98,4**	**93,7**
c+n	79,2	71,1	80,2	9,57	89,7	64,1	**88,9**	54,1	**90,2**	57,3	85,1	26,1	25,2	8,2	10,2	22,9
c+n+h_1	88,0	82,1	84,2	8,48	87,9	76,6	**87,6**	52,0	**89,1**	53,4	81,3	39,3	80,6	87,7	79,2	40,8
c+n+h_2	87,4	81,4	81,8	8,41	89,0	77,6	85,6	40,2	**89,1**	55,1	81,4	39,6	85,2	90,9	85,1	40,6
c+n+h_f	79,0	65,9	85,7	5,72	87,0	70,2	86,2	48,6	86,4	46,9	82,6	40,8	**99,9**	99,9	**98,5**	**93,5**

TABLE 5.8 – Précision de la classification des incohérences et scores f-mesure moyens pondérés (mode binaire et multiclasse) de nos modèles. La performance de base obtient une précision de 87 %. Nous présentons les résultats obtenus avec différentes combinaisons de données d'entrée. Les entrées possibles sont : le tour actuel (c), le prochain énoncé de l'utilisateur (n), le ou les deux derniers tours de l'historique du dialogue (h_1,h_2) ou l'historique complet (h_f).

Les résultats montrent que, même si la performance de base est élevée (87%), elle est surpassée par tous les modèles. Les meilleurs résultats (99%) sont obtenus par le modèle de type GPT-2 lorsqu'il utilise l'historique complet du dialogue (*h*) et l'énoncé suivant de l'utilisateur (*n*), bien qu'il obtienne déjà de bonnes performances sans *n*. Cependant, lorsque *h* n'est pas utilisé, les performances diminuent. Nous pensons que GPT-2 est capable d'extraire les informations d'entrée les plus pertinentes pour la classification des incohérences, il bénéficie donc d'un long historique. Au contraire, DIET et WMM2Seq ont obtenu les meilleurs résultats (0,90) avec l'énoncé de l'utilisateur suivant et sans aucun historique, même si la différence de performance sans l'énoncé de l'utilisateur suivant est inférieure à celle de GPT-2. De même, un simple Bi-LSTM surpasse la performance de base en utilisant à la fois *h1* et *h2* avec *n* en mode binaire, mais ne parvient pas à capter les caractéristiques nécessaires en mode multi-classe. Une explication possible de ce comportement serait que les réseaux basés sur les LSTM sont incapables de conserver l'historique suffisamment longtemps pour détecter le type correct d'incohérence lors du filtrage de l'historique du dialogue. En général, nous pouvons dire que l'ajout de l'énoncé de l'utilisateur suivant améliore les performances de classification des incohérences. Enfin, nous observons qu'avec moins d'informations, WMM2Seq obtient la meilleure précision après GPT-2. Cependant, comme l'entraînement GPT-2 est à la fois coûteux et nécessite l'ensemble du dialogue pour obtenir les meilleures performances, les résultats confirment la nécessité de prédire l'énoncé suivant de l'utilisateur pour obtenir un modèle plus précis, dans le cas où un modèle plus petit est requis ou lorsque l'historique complet du dialogue n'est pas disponible. Nous comparons également les prédictions des meilleures variantes du modèle aux annotations de référence et mesurons le Kappa de Cohen pour vé-

model	**bi-LSTM**	**DIET**	**WMM2Seq**	**GPT-2**
κ	0,74	0,76	0,67	**0,97**

TABLE 5.9 – Valeurs du Kappa de Cohen pour la comparaison des prédictions des meilleurs modèles avec les étiquettes de la vérité terrain

rifier que les performances des modèles sont supérieures au hasard. Les résultats sont présentés dans le tableau 5.9.

5.3.3 Conclusion

Avec cette première série d'expériences, nous fournissons une première preuve de l'avantage que nous pourrions obtenir en ayant des systèmes de dialogue possédant un oracle pour prédire le prochain tour de l'utilisateur tout en calculant la sortie actuelle du système, un premier pas vers une future architecture de dialogue avec un double modèle système et utilisateur.

5.4 Entropie et progression dans le dialogue

Dans cette section, nous expliquons l'intérêt du calcul de l'entropie afin de mesurer la progression de l'information au long du dialogue, et donc la progression temporelle d'un tour de parole. Après avoir présenté le problème de l'incohérence du système liée non seulement aux tours de parole passés mais aussi à ceux à venir, intéressons-nous maintenant au lien pouvant exister entre la progression dans le dialogue et ces incohérences. En d'autres termes, existe-t-il une corrélation entre le nombre de tours de parole restants et l'apparition d'une incohérence du système ? En fonction de quels facteurs le risque d'une incohérence est-il plus ou moins élevé ? Enfin, dans quelle mesure pouvons-nous apprendre à la machine à prévoir la durée restante d'un dialogue à un tour de parole donné ?

5.4.1 Calcul de l'entropie conditionnelle

Nous avons comparé la croissance d'entropie conditionnelle dans les différents sous-corpus et sur chaque domaine de DIASER. Nous avons également étudié le cas particulier du domaine du restaurant car il est, comme nous l'avons déjà évoqué, le seul à être présent dans les quatre sous-corpus. L'objectif de cette comparaison est de détecter des irrégularités (accélération ou ralentissement) de la croissance d'entropie d'un tour de parole à l'autre, et d'identifier les périodes où se situent ces irrégularités. Notre première hypothèse est que ces irrégularités permettront d'annoter chaque dialogue en tranches temporelles afin de posséder un indice de la progression du dialogue par tour de parole. Cela permettra à un système de prédire,

à travers cette progression, le nombre approximatif de tours de parole restants dans le dialogue, et donc de se projeter une image de ce qu'il pourrait s'y dérouler, à la façon du système mémoriel humain. Notre seconde hypothèse, quant à elle, est qu'il existe une corrélation entre la progression dans le dialogue et la détection d'une incohérence. Nos expériences explicitent cette corrélation et tentent de faire émerger une relation de causalité entre les deux informations.

Méthodologie

Afin de récupérer ces tranches temporelles, nous avons dans un premier temps calculé l'entropie conditionnelle de mots de chacun des sous corpus (dix-neuf domaines et quatre jeux de données différents). Nous n'effectuons pas ici de séparation (apprentissage, validation, test) dans les sous-corpus car le but n'est pas de créer un modèle mais uniquement des sources pour l'entropie. Pour chaque sous-corpus, nous utilisons six types de source d'entropie : l'énoncé de l'utilisateur seul, l'énoncé du système seul, l'état de croyance de l'utilisateur, l'énoncé et l'état de croyance de l'utilisateur, l'énoncé de l'utilisateur et celui du système, et les trois éléments précédents concaténés.
Dans un second temps, nous avons comparé la quantité d'entropie contenue dans chacune des tranches temporelles, en prenant en compte le tour de parole de l'utilisateur, son état de croyance, le tour de parole du système et les trois ensemble. Nous avons observé, pour chacune des tranches temporelles (1 à 5), la croissance d'entropie de caractères, de mots et de bigrammes de caractères en fonction du domaine et des jeux de données. Les figures 5.1 et 5.2 décrivent cette croissance.

Nous sommes capables de tirer certaines conclusions de ces graphiques. :

— On distingue clairement les cinq tranches temporelles dans les différentes mesures d'entropie. Les 2 et 3 sont généralement les pics, souvent la tranche 2 est dominée par l'utilisateur, et la 3 l'est le système. Contrairement à ce que nous pensions, la tranche 4 montre une entropie plus faible de la part du système que de l'utilisateur mais dans les corpus où un humain joue le rôle du système, c'est l'utilisateur qui domine. Notre hypothèse énoncée en 4.3.2 est en revanche confirmée pour DSTC2.
— Les tranches 1 et 5 sont celles qui montrent le moins d'entropie, les creux dans les deux graphiques représentent toujours la tranche 5.
— Dans le domaine du restaurant, on peut facilement distinguer les corpus humains du corpus humain-machine car ce dernier est le seul dont l'entropie de caractère est proportionnellement la plus importante, côté système. Cela est probablement dû aux nombreuses incohérences de ce dernier qui rendent plus imprévisibles les mesures d'entropie conditionnelle.
— Dans DSTC2, côté utilisateur, on s'aperçoit que le pic d'entropie est à la tranche 2 pour progressivement s'éteindre, confirmant que le schéma des cinq tranches temporelles et de l'alternance de la croissance d'entropie en

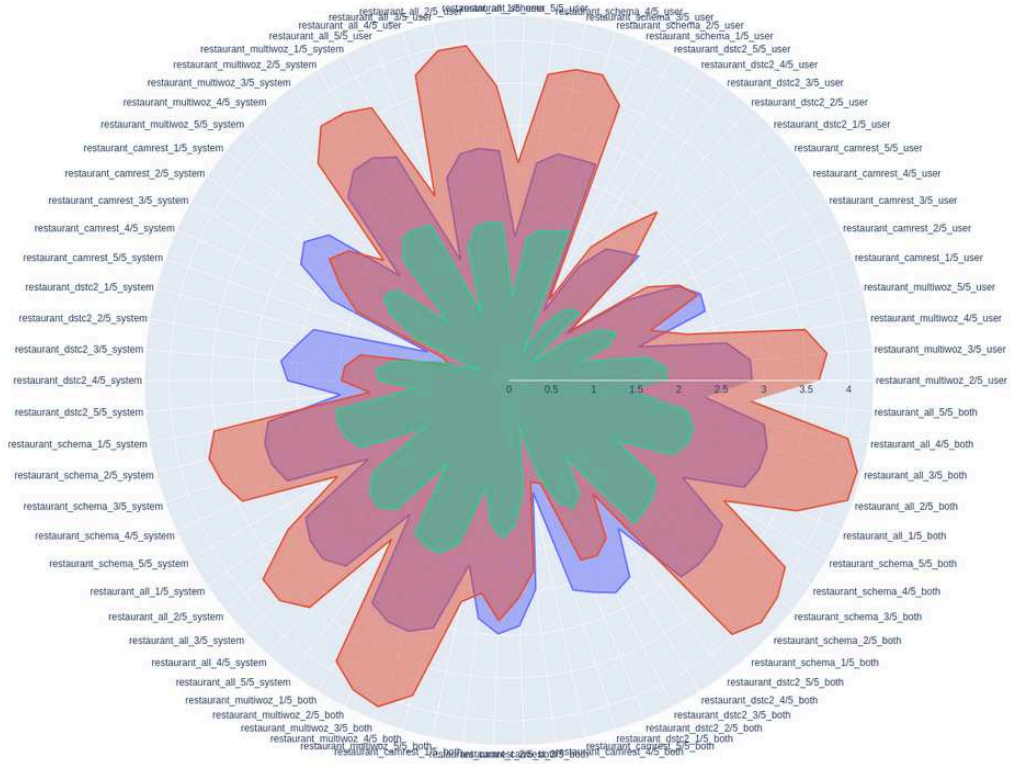

FIGURE 5.1 – Mesure d'entropie conditionnelle pour les cinq tranches temporelles, pour le restaurant, pour chaque sous-corpus, pour les différentes sources textuelles (voir 5.4.1) concaténées.En *rouge : l'entropie de mots*, en *bleu : l'entropie de caractères*, en *vert : l'entropie de bigrammes de caractères*. Notez que le violet représente soit le bleu lorsque le rouge est d'entropie supérieure, soit le rouge lorsque le bleu est d'entropie supérieure.
Chaque pétale contient cinq sommets, représentant la quantité d'entropie contenue dans la tranche correspondante. Le sens de lecture est celui du sens inverse des aiguilles d'une montre.

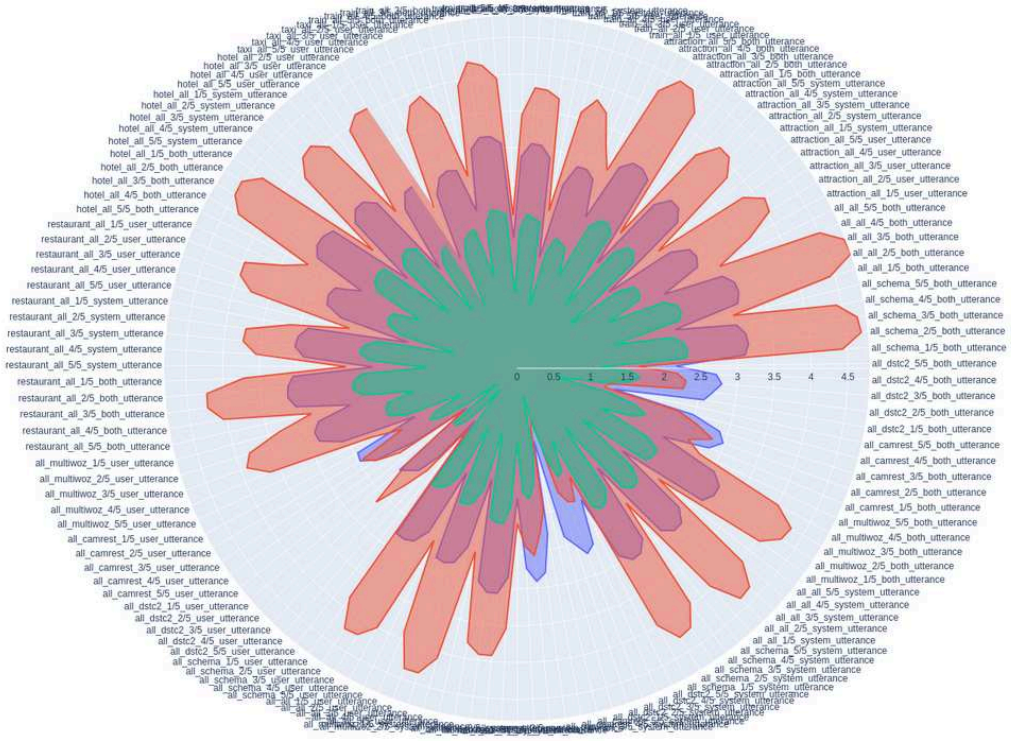

FIGURE 5.2 – Mesure d'entropie conditionnelle pour les cinq tranches temporelles, pour chaque domaine, pour chaque sous-corpus, pour les énoncés uniquement (voir 5.4.1). En *rouge : l'entropie de mots*, en *bleu : l'entropie de caractères*, en *vert : l'entropie de bigrammes de caractères*. Notez que le violet représente soit le bleu lorsque le rouge est d'entropie supérieure, soit le rouge lorsque le bleu est d'entropie supérieure.
Chaque pétale contient cinq sommets, représentant la quantité d'entropie contenue dans la tranche correspondante. Le sens de lecture est celui du sens inverse des aiguilles d'une montre.

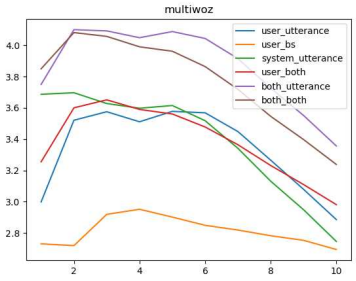

FIGURE 5.3 – Entropie conditionnelle MultiWOZ de chaque tour. Les différentes courbes représentent l'évolution de l'entropie contenue dans chaque source de données. Par exemple, *user_bs* constitue la quantité d'entropie contenue dans l'état de croyance de l'utilisateur. *both_utterance* constitue la quantité d'entropie contenue dans la concaténation des énoncés de l'utilisateur et du système.

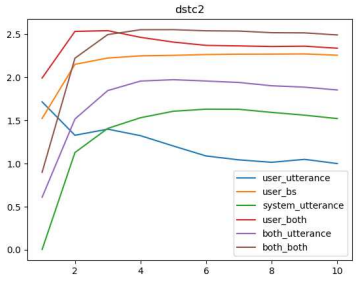

FIGURE 5.4 – Entropie conditionnelle de DSTC2 de chaque tour. Les différentes courbes représentent l'évolution de l'entropie contenue dans chaque source de données. Par exemple, *user_bs* constitue la quantité d'entropie contenue dans l'état de croyance de l'utilisateur. *both_utterance* constitue la quantité d'entropie contenue dans la concaténation des énoncés de l'utilisateur et du système.

fonction du locuteur est avérée pour le corpus humain-machine.

— Pour le domaine du restaurant, contrairement à MultiWOZ, où l'entropie est la plus forte, utilisateur ou système, à la tranche 3, pour SgD, c'est plutôt dans la tranche 2 où l'entropie est la plus forte. Cela peut être expliqué par les schémas pré-conçus dans ce dernier qui favorisent l'apparition des informations importantes assez tôt dans le dialogue.

— D'une façon générale, pour les corpus humains, la tendance est d'avoir un pic entre les tranches 2 et 3, qui affichent une entropie similaire. Sans doute devrions-nous diviser en plus de tranches temporelles pour tenter de les distinguer mieux. La seule exception est pour le domaine du train (tout en haut du graphique 5.2, peu visible, nous présentons nos excuses) où le pic se situe à la tranche 3.

— En général également, l'entropie de caractères est proportionnelle à celle de mots, et suit le même schéma de croissance.

Analyse des entropies et corrélations entre information et domaine

Ensuite, nous avons étudié l'évolution de l'entropie, à chaque tour de parole. Pour ce faire, nous avons considéré n sources, chacune correspondant à l'ensemble des $n-èmes$ tours de parole des dialogues d'un sous corpus. n allant de 0 à la taille du plus grand dialogue (en terme de tours de parole) de chaque sous corpus. Les figures 5.4 5.6 5.3 5.5 5.8 5.7 5.11 5.12 5.9 et 5.10 illustrent l'évolution de l'entropie par tour de parole dans les différents sous-corpus et pour l'ensemble DIASER. Pour toutes ces figures, nous observons les dix premiers tours de parole, qui, à l'exception de DSTC2, représentent la taille moyenne des dialogues dans les autres corpus.

Voici les faits que nous avons observés :

— Une des premières choses à remarquer dans cette croissance d'entropie, est

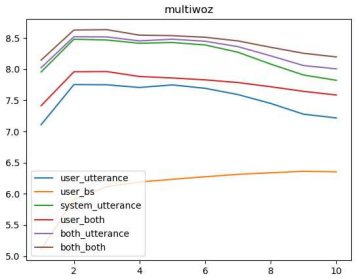

FIGURE 5.5 – Entropie de Shannon de MultiWOZ de chaque tour. Les différentes courbes représentent l'évolution de l'entropie contenue dans chaque source de données. Par exemple, *user_bs* constitue la quantité d'entropie contenue dans l'état de croyance de l'utilisateur. *both_utterance* constitue la quantité d'entropie contenue dans la concaténation des énoncés de l'utilisateur et du système.

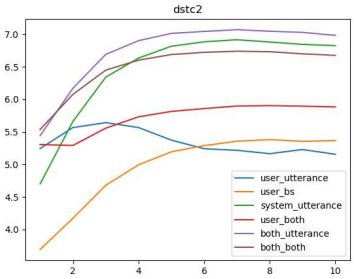

FIGURE 5.6 – Entropie de Shannon de DSTC2 de chaque tour. Les différentes courbes représentent l'évolution de l'entropie contenue dans chaque source de données. Par exemple, *user_bs* constitue la quantité d'entropie contenue dans l'état de croyance de l'utilisateur. *both_utterance* constitue la quantité d'entropie contenue dans la concaténation des énoncés de l'utilisateur et du système.

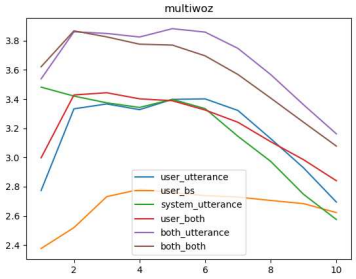

FIGURE 5.7 – Entropie conditionnelle de MultiWOZ pour le domaine du restaurant de chaque tour de parole du corpus. Les différentes courbes représentent l'évolution de l'entropie contenue dans chaque source de données. Par exemple, *user_bs* constitue la quantité d'entropie contenue dans l'état de croyance de l'utilisateur. *both_utterance* constitue la quantité d'entropie contenue dans la concaténation des énoncés de l'utilisateur et du système.

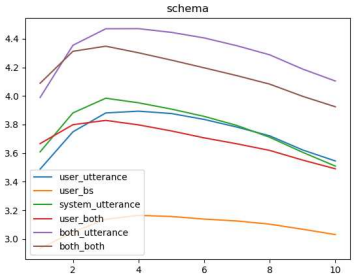

FIGURE 5.8 – Entropie conditionnelle de SgD de chaque tour de parole du corpus. Les différentes courbes représentent l'évolution de l'entropie contenue dans chaque source de données. Par exemple, *user_bs* constitue la quantité d'entropie contenue dans l'état de croyance de l'utilisateur. *both_utterance* constitue la quantité d'entropie contenue dans la concaténation des énoncés de l'utilisateur et du système.

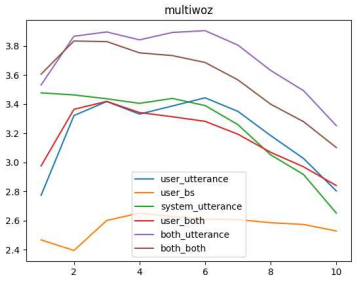

FIGURE 5.9 – Entropie conditionnelle de MultiWOZ pour le domaine de l'hôtel de tour de parole du corpus. Les différentes courbes représentent l'évolution de l'entropie contenue dans chaque source de données. Par exemple, *user_bs* constitue la quantité d'entropie contenue dans l'état de croyance de l'utilisateur. *both_utterance* constitue la quantité d'entropie contenue dans la concaténation des énoncés de l'utilisateur et du système.

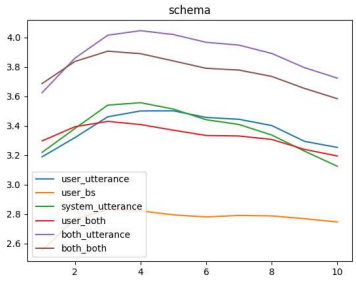

FIGURE 5.10 – Entropie conditionnelle de SgD pour le domaine de l'hôtel de chaque tour de parole du corpus. Les différentes courbes représentent l'évolution de l'entropie contenue dans chaque source de données. Par exemple, *user_bs* constitue la quantité d'entropie contenue dans l'état de croyance de l'utilisateur. *both_utterance* constitue la quantité d'entropie contenue dans la concaténation des énoncés de l'utilisateur et du système.

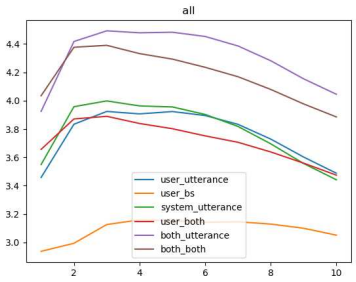

FIGURE 5.11 – Entropie conditionnelle de DIASER de chaque tour de parole du corpus. Les différentes courbes représentent l'évolution de l'entropie contenue dans chaque source de données. Par exemple, *user_bs* constitue la quantité d'entropie contenue dans l'état de croyance de l'utilisateur. *both_utterance* constitue la quantité d'entropie contenue dans la concaténation des énoncés de l'utilisateur et du système.

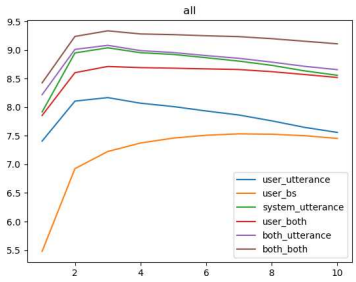

FIGURE 5.12 – Entropie de Shannon de DIASER de chaque tour de parole du corpus. Les différentes courbes représentent l'évolution de l'entropie contenue dans chaque source de données. Par exemple, *user_bs* constitue la quantité d'entropie contenue dans l'état de croyance de l'utilisateur. *both_utterance* constitue la quantité d'entropie contenue dans la concaténation des énoncés de l'utilisateur et du système.

la différence d'augmentation de celle-ci entre les énoncés de l'utilisateur et ceux du système. En effet, nous remarquons que d'une manière générale, dans les premiers tours de parole ce sont les énoncés du système qui possèdent l'entropie conditionnelle la plus élevée, puis elle est rattrapée par celle des énoncés de l'utilisateur et enfin elle est dépassée par cette dernière. Autrement dit, nous distinguons trois "époques" dans un dialogue, qui correspondent aux phases du dialogue orienté-tâche : présentation de la tâche par l'utilisateur avec les informations nécessaires pour le système, échange d'information, confirmation, correction de la part des deux interlocuteurs et résolution de la tâche par le système en fournissant les réponses attendues par l'utilisateur. A cela nous pouvons ajouter deux phases, la salutation, uniquement observée dans les données à partir du corpus DSTC2, et la conclusion, caractérisée par un remerciement de la part de l'utilisateur ainsi qu'une prise de congé.

— De la même façon que ce phénomène se produit pour chaque type d'interlocuteur, il se produit également pour chaque type de domaine, mais pas au même moment. Certains domaines (restaurant) voient le croisement entre l'augmentation de l'entropie de l'utilisateur et celle du système arriver plus tôt, d'autres plus tard (hôtel). Il serait alors également envisageable, dans un second temps, de prédire le domaine dans lequel évolue un dialogue en fonction de la mesure de l'information.

— Ensuite, il est intéressant de constater que ce phénomène ternaire dépend de l'origine des interlocuteurs, car on ne le retrouve pas pour les dialogues avec compère et avec schéma mais seulement pour les dialogues humain-machine. En effet, le cas des cinq tranches temporelles est avéré pour le cas de DSTC2 que ce soit en calculant l'entropie de Shannon ou l'entropie conditionnelle. Il existe un certain déséquilibre entre la quantité d'information dans les énoncés de l'utilisateur et ceux du système. En outre, dans le cas de DSTC2, l'entropie de l'état de croyance est supérieure à celle de l'énoncé de l'utilisateur, montrant la faible quantité d'information non prévisible contenue dans ce dernier. Nous pouvons observer la façon dont les courbes des énoncés utilisateur et système se croisent, montrant que le système finit par générer plus d'information non prévisible que l'utilisateur, à partir du quatrième tour de parole. La question que l'on peut se poser ici est la corrélation entre ce quatrième tour de parole et l'apparition progressive d'incohérences dans les tours de parole.

— Dans le cas des trois autres sources de dialogues, il existe au contraire une harmonie entre l'entropie de l'utilisateur et celle du système. L'état de croyance est assez prévisible avec une entropie faible. Nous pourrions penser que l'utilisation des corpus humain-humain pour modéliser les dialogues humain-machine a ses limites car l'apparition d'information non attendue n'est pas reflétée lorsqu'un humain joue le rôle du système.

— Dans le cas de l'entropie de Shannon, l'information non attendue est plus importante chez le système, quel que soit le corpus d'origine. Par contre, dans le cas de la conditionnelle, pour les corpus humain-humain, les énoncés utilisateur deviennent moins prévisibles que ceux du système au fur et à mesure que l'on avance dans le dialogue. Nous émettons l'hypothèse que, malgré tout, les compères (Magicien d'Oz) suivent un script lors de la conversation, avec un certain nombre de réponses-types aux énoncés de l'utilisateur, identifiables grâce aux bigrammes, et donc faisant baisser la quantité d'entropie.
— Le domaine du restaurant, pour MultiWOZ et Schema, ressemble à l'ensemble des données, ce qui nous conduit à penser que ce domaine est représentatif de l'ensemble des jeux de données, et donc que l'on peut l'utiliser comme exemple pour nos résultats et comme domaine pour l'évaluation.
— Dans la mesure où c'est l'interaction humain-machine et donc DSTC2 qui nous intéresse, nous gardons les annotations des dialogues en cinq tranches temporelles. Cependant, nous modifions les tailles de celles-ci en fonction de l'appartenance d'un dialogue à une source . En effet, la partie "échange", où l'entropie des deux locuteurs est similaire, occupe la majeure partie des dialogues issus des conversations humain-humain, il semble donc raisonnable de redistribuer la taille de chaque tranche en fonction de nos observation précédentes.
— Contrairement à ce que nous imaginions, l'état de croyance de l'utilisateur n'est pas un indicateur fort de la progression de l'information dans le dialogue, il n'est pas suffisant pour pouvoir prédire la progression dans le dialogue. Cela est peut-être dû, si l'on met de côté les erreurs d'annotation [Higashinaka et al., 2015], au fait que l'état de croyance ne reflète pas les négociations à l'intérieur du dialogue [Yamaguchi et al., 2021], par exemple les éventuelles confirmations demandées par le système, les demandes de reformulation, les incohérences du système et la réaction de l'utilisateur...
— L'entropie de caractère nous semble moins significative que celle de mot, dans la mesure où la croissance d'entropie peut venir de la diversité des valeurs possibles pour les entités. En effet, l'entropie de mots est moins sensible aux fautes d'orthographe ou aux variations lexicales[Bentz et al., 2017]. C'est pourquoi les expériences suivantes sont élaborées en n'utilisant comme indices que les entropies issues des bigrammes de mots.
— Si l'on compare les domaines du restaurant et de l'hôtel, on s'aperçoit que les schémas sont similaires. La principale différence est entre MultiWOZ et SgD, où dans le premier, on voit que l'entropie du système et de l'utilisateur suivent le schéma des cinq tranches temporelles, alors que pour SgD, le pic se situe assez tôt dans le dialogue et l'entropie décroît de façon continue ensuite.

— D'une façon générale, si l'on observe DIASER, on s'aperçoit que pour l'entropie conditionnelle, les courbes d'entropie de l'utilisateur et du système présentent des formes et des valeurs similaires, alors que pour l'entropie de Shannon, la courbe d'énoncé de l'utilisateur est plus faible que celle du système, montrant que lorsqu'on analyse les mots comme des variables indépendantes, les énoncés du systèmes montrent une plus grande perplexité que ceux de l'utilisateur.

En définitive, le calcul de l'entropie conditionnelle nous a permis de constater que nos hypothèses sur les époques au sein d'un dialogue étaient acceptables pour l'interaction humain-machine, mais pas pour les dialogues humain-humain. Cela nous permet de considérer que l'entropie conditionnelle peut nous servir d'annotateur pour étiqueter les tours de parole en tranches temporelles. Celles-ci étant au nombre de cinq, en nous appuyant sur la moyenne de la croissance des entropies, nous considérons que la première et la dernière tranche, que nous décidons de qualifier de tranches extérieures sont toujours représentées par le premier et le dernier tour de parole, et que les trois tranches intérieures sont de taille équivalente dans le cas de DSTC2. La moyenne de tours de parole par dialogue étant supérieure à 10 pour DSTC2, pour un dialogue de huit tours de parole, nous considérons le premier tour de parole comme appartenant à la tranche "salutations", les deux suivants comme dans la tranche "utilisateur", les quatre et cinq, dans la tranche "échange", les deux suivants dans la tranche "système" et enfin le dernier dans la tranche "congé". Toutefois, l'observation des croissances d'entropie montre qu'au sein des tranches intérieures, il peut exister des irrégularités, et ce pour plusieurs domaines. Il est possible que cela provienne du fait que certains dialogues traitent plusieurs domaines, en particuliers ceux supérieurs à dix tours de parole, et donc que l'on peut subdiviser les tranches intérieures.

5.4.2 Corrélation entre l'entropie et les erreurs dans le dialogue

Nous avons pu constater que la croissance d'entropie au long du dialogue suit le schéma suivant : à partir du deuxième tour de parole jusqu'à la moitié du dialogue, l'entropie croît, puis chute, avec le manque de nouvelles informations de la part de l'un ou l'autre des locuteurs. Nous avons alors cherché à savoir si cette croissance d'entropie, qui pourrait faire penser à une loi normale, peut nous apprendre aussi à anticiper les moments compliqués du dialogue pour le système.
Pour ce faire, nous avons évalué un système de dialogue entraîné sur le corpus MultiWOZ en termes de scores INFORM et REQUEST, correspondant respectivement au nombre d'informations données par l'utilisateur correctement identifiées par le système [7], et au nombre d'informations demandées par l'utilisateur correctement

7. par exemple le fait que l'utilisateur veut un hôtel, pas cher, avec WI-FI dans la chambre : les entités

fournies par le système[8]. Il s'agit donc d'une évaluation concernant l'ensemble du dialogue et non pas chaque tour de parole, comme peut l'être l'évaluation de la génération de réponse du système. Ces scores établis, nous avons analysé le corpus afin de relever, pour chaque tour de parole, les erreurs dans l'état du dialogue prédit par le système par rapport à celui attendu. Nous utilisons les quatre composants de l'état du dialogue : le domaine, l'intention de l'utilisateur, les entités et les valeurs et nous calculons, pour chacun d'entre eux, le nombre d'éléments manquants (silence) ou superflus (bruit). Ce calcul nous permet d'obtenir le graphique 5.13, "all failed" signifie, pour un dialogue, que les tests INFORM et REQUEST ont tous deux échoué. "all success" signifie l'inverse.
Nous proposons quelques remarques :

- Que ce soit pour les dialogues réussis ou échoués, le premier tour de parole est celui qui concentre le plus d'erreurs, et dans des proportions comparables pour les deux types de dialogue.
- D'une façon générale, les dialogues échoués montrent une grande quantité d'informations superflues dans les cinq premiers tours de parole, en particulier au niveau des intentions de l'utilisateur. Tandis qu'à partir du quatrième tour de parole, on constate beaucoup d'informations manquantes, notamment au niveau du domaine du dialogue. Si l'on additionne les informations manquantes avec celles qui sont superflues, on s'aperçoit que le pic d'erreurs se situe entre les tours trois et sept de dialogue.
- Il est intéressant de constater que pour les quatre premiers tours de parole, dans les dialogues réussis, nous avons une quantité non négligeable d'informations superflues, ainsi qu'un nombre constant d'informations manquantes durant ce même intervalle.

En définitive, il est intéressant de constater que l'accumulation des erreurs dans les dialogues ne réussissant pas les évaluations INFORM et REQUEST forme une loi normale assez proche de celle de l'évolution de l'entropie pendant un dialogue.

5.4.3 Prédiction des tranches temporelles

Nous précisons ici, que le tranchage temporel est une modélisation approximative de la progression dans le dialogue. Nous avons expliqué que le chevauchement de ces tranches pouvait se faire au sein d'un même tour de parole. Nous pouvons également considérer que l'époque de "l'échange" n'est qu'une succession de tranches courtes et entremêlées pour ce qui est de leur origine : "utilisateur" ou "système". Une étude plus approfondie sur ces époques nous semble judicieuse ici, cependant, pour des raisons de simplification, et comme cette prédiction de la progression dans le dialogue n'est qu'un module parmi d'autres du système global Bi-WMM2Seq, nous n'avons pas approfondi davantage la question.

8. le nom et le numéro de téléphone dudit hôtel, son code postal : la valeur des entités

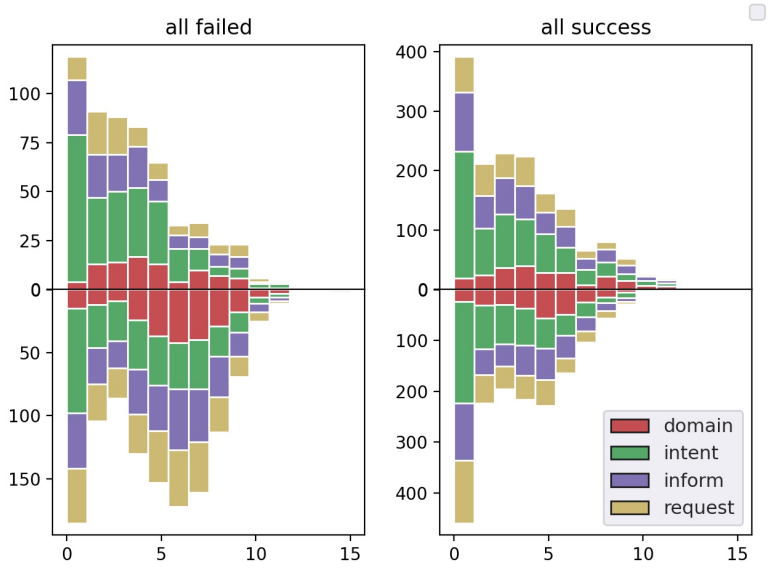

FIGURE 5.13 – Distribution des erreurs de l'état de croyance en fonction des scores INFORM et REQUEST. L'abscisse correspond au numéro du tour de parole dans le dialogue, l'ordonnée correspond au nombre de dialogues ayant, vers le haut, des informations en trop, et vers le bas, des informations manquantes. Pour le graphique de gauche, cela signifie qu'au tour 0 de dialogue, il y a cent cinquante dialogues pour lesquels il manque des entités informables, et soixante-quinze pour lesquels on trouve des informations en trop sur les intentions de l'utilisateur.

Protocole d'annotation des tranches temporelles

Nous avons donc étiqueté l'intégralité des données DIASER avec ces tranches temporelles de un à cinq.

L'objectif est maintenant de modéliser ces tranches temporelles, donc d'entraîner un modèle capable de prédire, en fonction d'un tour de parole, le nombre de tours restants dans le dialogue.
Pour ce faire, nos indices sont les suivants : l'entropie de Shannon (calculée sur le tour de parole courant, avec et sans son historique de dialogue), l'entropie conditionnelle (dans le cas du corpus d'évaluation, l'entropie conditionnelle est croisée avec l'entropie du corpus d'apprentissage, voir 5.4), le domaine (nous considérons qu'en situation réelle l'utilisateur et le système savent à l'avance le thème de leur dialogue, et que donc cette information n'a pas à être devinée par ce dernier) et la taille de l'historique de dialogue.
Nous avons mentionné en 5.3.1 que la détection d'incohérences dans un tour de parole était possible non seulement grâce à l'historique de dialogue, mais aussi grâce à l'analyse des tours de parole suivants. Ainsi, dans le cas du corpus DSTC2, nous ajoutons comme indice (en tant qu'oracle) la présence, ou non, d'une incohérence dans l'échange d'énoncés entre l'utilisateur et le système.
Pour entraîner les données, nous utilisons un bi-LSTM, qui s'est révélé être une référence moderne pour les tâches de classification de tours de parole de dialogue

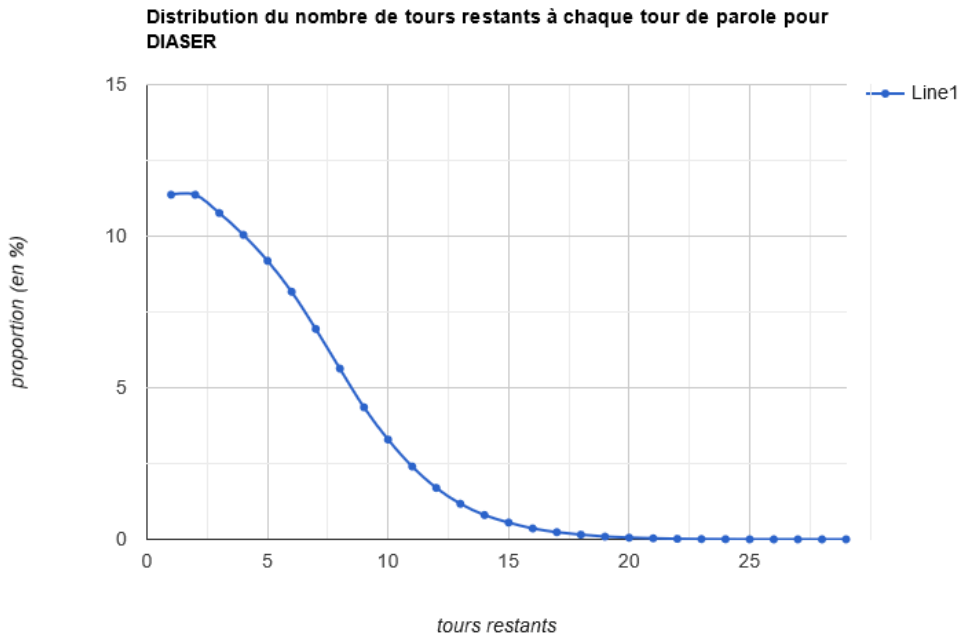

FIGURE 5.14

[Skantze, 2017].
Nous étiquetons chaque tour de parole courant avec le nombre de tours de parole restants dans le dialogue. Cependant, la distribution des tours de parole restants n'est pas homogène, dans la mesure où les dialogues ne sont pas tous de même longueur. Ainsi pour un tour de parole t, la probabilité qu'il reste cinq tours de parole, ou moins, est bien plus élevée que la probabilité qu'il en reste six ou plus. Les distributions pour DSTC2 et MultiWOZ étant presque identiques à DIASER, nous ne présentons que les statistiques de ce dernier (voir 5.14).

Modèles de référence et résultats pour la prédiction du nombre de tours restants

Pour finir sur cette étude de l'entropie dans le but de prédire le nombre de tour de paroles restants dans le dialogue, nous avons désiré comparer la capacité d'un modèle à effectuer cette tâche en fonction de la nature du dialogue (humain-humain ou humain-machine), du domaine, mais également de la présence ou non d'une incohérence et de la croissance d'entropie de chaque tour de parole.

Nous avons comparé, pour le seul corpus humain-machine que nous possédons, à savoir DSTC2, l'impact de la mesure d'entropie mais aussi, à chaque tour de parole, la présence ou non d'une incohérence, dans la prédiction du nombre d'énoncés restants. Pour corroborer notre hypothèse sur la nécessité d'anticiper le tour de parole suivant, nous avons intégré dans le contexte de dialogue, l'énoncé de l'utili-

Tours **Corpus**	**0 ou 1** f-mesure	**2 ou 3** f-mesure	**4 à 6** f-mesure	**7 à 10** f-mesure	**plus de 10** f-mesure	**Tout** exactitude
DSTC2	0,81	0,61	0,42	0,53	0,01	0,54
DSTC2+$t+1$	**0,99**	0,73	0,40	0,46	0,26	0,59
DSTC2+tranches	0,82	0,60	0,42	0,53	0,01	0,52
DSTC2+inc	0,98	0,73	0,40	0,47	0,26	0,59
DSTC2+inc+$t+1$	**0,99**	0,75	0,48	**0,56**	0,04	0,62
DSTC2+tranches+inc	0,98	**0,76**	0,42	0,51	**0,29**	0,62
DSTC2+tranches+inc+$t+1$	0,98	**0,76**	0,50	0,53	0,25	**0,63**
DSTC2+tranches+inc+entropie	**0,99**	0,74	**0,54**	0,54	0,00	0,60
DSTC2+tranches+entropie+$t+1$	0,98	0,69	0,46	0,51	0,09	0,61
DSTC2+all	0,98	0,74	0,46	0,53	0,12	0,61
MultiWOZ	0,86	0,61	0,39	0,59	0,00	0,62
MultiWOZ+$t+1$	0,93	0,65	0,34	0,58	0,00	0,64
MultiWOZ+tranches	**0,99**	**0,73**	0,48	**0,62**	0,02	0,68
MultiWOZ+tranches+$t+1$	0,95	0,71	**0,49**	0,61	**0,08**	**0,69**
MultiWOZ+tranches+entropie	0,89	0,62	0,44	0,60	0,00	0,64
MultiWOZ+tranches+entropie+$t+1$	0,95	0,67	0,51	**0,62**	0,04	0,68

TABLE 5.10 – prédiction tours de parole restants pour DSTC2 et MultiWOZ

sateur à $t+1$. Pour avoir un élément d'analyse significatif, nous avons également cherché à prédire le nombre de tours de parole restants pour MultiWOZ. Bien sûr, pour ce dernier, nous n'avons pas pu ajouter comme indice la présence potentielle d'une incohérence.

5.4.4 Résultats de la prédiction des tours de parole restants

Nous modélisons le problème de la prédiction du nombre de tours de parole restants comme une tâche de classification. En effet, nous étions tentés de la modéliser comme un problème de régression, à l'image de [Cao et al., 2020]. Cependant, comme il s'agit de valeurs relativement petites (de 0 à 29) et que l'immense majorité des valeurs se situe entre 0 et 9, nous avons décidé de le traiter comme un problème de classification à cinq valeurs :

- valeur 0 = tour de parole restant : 0 ou 1
- valeur 1 = tours de parole restants : 2 ou 3
- valeur 2 = tours de parole restants : 4 à 6
- valeur 3 = tours de parole restants : 7 à 10
- valeur 4 = tours de parole restants : plus de 10

Le but étant, pour un tour de parole dans un dialogue, prédire le nombre de tours de parole restants jusqu'à la fin de celui-ci. Nous avons utilisé, comme cité dans 4.3 le Bi-lstm, efficace pour prédire la durée restante d'une séquence. Ce Bi-LSTM est combiné, à des indices numériques représentant les tranches temporelles et, dans le cas de DSTC2, d'un booléen représentant la présence d'une incohérence[9]. Nous utilisons comme mesure d'évaluation l'exactitude ainsi que la f-mesure pondérée. Comme les classes sont a priori équitablement représentées, il n'est pas nécessaire

9. https ://towardsdatascience.com/how-to-combine-textual-and-numerical-features-for-machine-learning-in-python-dc1526ca94d9

de biaiser le calcul de la f-mesure. Les résultats sont dans le tableau 5.10

5.4.5 Analyse des résultats sur les tours de parole restants

Encore une fois, plusieurs éléments nous paraissent assez importants pour que nous nous y attardions :

- Tout d'abord, si nous observons les résultats classe par classe, nous rendons compte que malgré nos efforts pour équilibrer les tours de parole restants, seuls les derniers tours de parole possèdent un bon score d'exactitude indépendamment de la combinaison d'éléments et de corpus. Cela est probablement dû au fait que lors des derniers tours de parole, certains termes ou tournures ont tendance à revenir (salutations, remerciements..) qui permettent au modèle de bien apprendre à les prédire. A l'inverse, dès qu'il reste plus de quatre tours de parole, le modèle est bien incapable de prédire s'il en reste cinq ou neuf.
- Dans le cas de DSTC2, on s'aperçoit que les trois indices (incohérence, tranche temporelle et tour de parole suivant) permettent d'obtenir les meilleurs résultats d'exactitude, montrant l'intérêt de les ajouter lors de cette tâche.
- Contrairement au DSTC2, le modèle appris avec MultiWOZ est plus performant lorsqu'on utilise uniquement les tranches alors qu'il l'est moins lorsqu'on ajoute tous les indices. En effet, dans DSTC2, les tranches temporelles non seulement n'aident pas à améliorer le Bi-LSTM mais obtiennent de moins bons résultats qu'avec le corpus sans ajout d'information.
- De façon surprenante, le Bi-LSTM parvient mieux à prédire les tours de parole restants pour MultiWOZ que pour DSTC2. Il est vrai qu'en moyenne les tours de parole de MultiWOZ sont plus courts que ceux de DSTC2. Cependant, comme MultiWOZ est un corpus multi-domaine et que dans un même dialogue, plusieurs domaines peuvent être concernés, nous nous attendions à ce que la tâche de prédiction soit plus difficile que pour DSTC2. Les incohérences de ce dernier sont probablement la raison pour laquelle le modèle montre plus de difficultés à prédire le nombre de tours de parole restants.
- Dans la mesure où le tranchage temporel n'est pas représentatif de la distribution de l'information dans le dialogue pour MultiWOZ, il est étonnant de constater que cet indice améliore la prédiction du nombre de tours de parole restants.
- Comme attendu, la présence du tour de parole suivant améliore significativement la précision du modèle, en particulier pour la classe 0 (o ou 1 tour de parole restant) et la classe 1 (2 ou 3 restants)
- En ce qui concerne les classes, moins il y a de tours restants à prévoir

et meilleur est le modèle. La classe 0 est presque toujours au dessus de 95% de F-mesure, tandis que la classe 1 oscille entre 70% et 76%. Pour les classes 2 et 3, le modèle a plus de difficultés à les distinguer, les deux gravitent autour de 50%, même si la classe 3 obtient une meilleure f-mesure, montrant que le modèle parvient mieux à identifier le nombre de tours à venir, quand il reste entre sept et dix tours de parole plutôt que quand il en reste entre quatre et six, ces derniers correspondant à la fois au pic d'entropie et au pic d'erreurs...Il est possible qu'il existe une corrélation entre ces trois phénomènes.

— Dans le cas de la classe 4, où le modèle doit deviner qu'il reste au moins dix tours de parole, les résultats sont très bas dans les deux corpus, mais plus encore dans MultiWOZ. Cela peut s'expliquer par le fait que les dialogues y sont généralement plus courts et dépassent rarement les dix tours de parole. Par ailleurs, dans DSTC2, les longs dialogues sont généralement ceux contenant le plus d'incohérences, liées à des répétitions ou à des erreurs dans la reconnaissance vocale, ce qui peut entraîner des difficultés pour savoir combien de tours il reste.

— Les scores d'entropie ne sont pas déterminants. Ils nous ont permis de créer les tranches temporelles qui, elles, semblent très utiles, notamment pour MultiWOZ. Cependant, l'évolution de l'entropie par tour de parole n'aide pas vraiment le Bi-LSTM à s'améliorer.

— D'une façon générale, les deux corpus utilisés obtiennent des résultats d'exactitude assez proches. Il est très difficile pour l'apprentissage automatique de savoir s'il reste cinq ou neuf tours de parole avant la fin du dialogue. Ce phénomène n'est pas étonnant puisque nous sommes nous-mêmes incapables de répondre à cette question. Pour en être totalement sûr il faudrait faire une expérience où l'on demanderait à des personnes le nombre de tours de parole restants.

— Enfin, quelle que soit la configuration, sauf les tranches temporelles pour DSTC2, l'ajout des informations améliore à chaque fois le modèle. Cela montre l'intérêt pour la prédiction temporelle d'ajouter non seulement le tour de parole suivant, mais également que la connaissance sur la présence d'une incohérence améliore la prédiction sur le nombre de tours de parole restants.

5.5 Le Bi-WMM2Seq

L'objectif maintenant est d'intégrer ces différents apprentissages en tant que fonctionnalités (ou fonctions) du Bi-WMM2Seq, conformément à notre modèle à trois axes (4.3) . Nous venons de conclure que la prédiction de la progression dans le dialogue est facilitée par la détection d'une incohérence, dans la mesure

où cette dernière dépend autant de la prédiction du tour de parole suivant que de l'historique du dialogue. Il s'agit désormais d'observer si ces deux fonctionnalités améliorent la génération de l'énoncé du système à un tour de parole donné.

5.5.1 Modèles de base

Dans la mesure où notre modèle apprend la tâche de génération de réponse sur le corpus DSTC2, nous avons choisi comme modèles de référence des modèles que nous avons déjà utilisés. Le corpus DSTC2 modifié avec l'ajout des incohérences et du nombre de tours de parole restants, que l'on nomme DSTC2.1 est utilisé pour résoudre la tâche de génération de réponse.

1. **WMM2Seq** nous choisissons comme première référence ce modèle car c'est celui qui nous a permis de parvenir à proposer notre contribution. Il est vrai que les trois fonctionnalités que nous ajoutons restent adaptables à un autre modèle, mais l'élégance et la clarté du WMM2Seq nous ont fourni la possibilité de comprendre et de présenter une variante basée sur l'architecture multi-mémorielle du WMM2Seq original.
2. **Marco** nous utilisons également celui-ci qui nous a été utile pour constituer DIASER. Son architecture assez proche du WMM2Seq mais avec la particularité d'utiliser l'état de croyance comme indice nous intéresse, d'autant plus que celui-ci est prédit conjointement avec le tour de parole, ce qui fait son originalité.
3. **GPT-2** ce modèle de référence, un tranformeur ayant obtenu d'excellents résultats à la fois sur DIASER et sur la détection d'incohérences, nous semble indispensable comme comparateur.

5.5.2 Bi-WMM2Seq en mode oracle

L'objectif ici est de permettre au WMM2Seq d'utiliser les indices que nous cherchons à prédire les uns après les autres comme s'ils étaient acquis. Nous fournissons au WMM2Seq le tour de parole suivant de l'utilisateur dans sa mémoire épisodique (qui gère rappelons-le, entre autres, l'historique du dialogue). Nous injectons également, au moment où le WMM2Seq décode sa mémoire épisodique dans le but de générer à la fois un vocabulaire lié à l'énoncé de l'utilisateur et un pointeur vers l'historique de dialogue, les informations sur la présence potentielle d'une incohérence et le nombre de tours de parole restants. Nous simulons en quelque sorte la génération du tour de parole suivant par le WMM2Seq interne ainsi que la détection d'une incohérence, puis enfin la prédiction du nombre de tours de parole restants. Pour simplifier la notation, dans le tableau 5.11, nous appelons Bi-WMM2Seq l'imitation en mode oracle du WMM2SeqG qui aurait inféré

et prédit le tour de parole suivant de l'utilisateur grâce au WMM2SeqI, et l'aurait concaténé avec son historique de dialogue.
Nous y revenons en section 6.6, les résultats présentés ci-dessous prennent en compte les informations réelles et non générées. Nous n'avons pas pris le risque d'implémenter le Bi-WMM2Seq en mode génération, tel que nous l'avons défini en 4.3 car nous désirons optimiser l'architecture avec ces nouvelles informations, notamment au niveau de l'utilisation du tour de parole suivant, avant de risquer de déployer un modèle qui n'améliorerait pas forcément les performances du WMM2Seq original.

5.5.3 DSTC2.1

Il nous faut rappeler le fait que DSTC2 utilisé par les réseaux de mémoire ne contient ni actes de dialogue ni états de croyances. Les réseaux de mémoire ne cherchent à faire ni de la prédiction d'acte, ni du pistage d'état. D'autre part, nous avons ajouté trois informations à chaque tour de parole de DSTC2 : la présence d'une incohérence via un booléen (0 ou 1), le nombre de tours de parole restants via les cinq classes vues en 5.4.4 et l'énoncé de l'utilisateur à $t+1$. Ce dernier est concaténé au contexte (historique de dialogue) lors de l'encodage du WMM2Seq, chaque *token* étant étiqueté avec $\$n$ (n comme next), pour différencier les mots issus de l'énoncé de l'utilisateur annotés par $\$u$ et ceux de l'énoncé du système annotés par $\$s$).
Lors du décodage, à chaque saut de réflexion de la mémoire épisodique, avant la fonction d'activation $SOFTMAX$, nous concaténons la dernière couche de la mémoire avec les informations catégorielles sur l'incohérence et le nombre de tours de parole restants. Lors de nos tests d'ablation, nous enlevons le tour suivant au moment de l'encodage, ainsi que le vecteur catégoriel dans le décodage. Nous utilisons la fonction $TRANS$ avec Bi-GRU par défaut du WMM2Seq pour opérer la copie de l'historique de dialogue à chaque saut de réflexion.
Enfin, nous utilisons les hyperparamètres par défaut du WMM2Seq.

5.5.4 Métriques d'évaluation

Nous utilisons les mêmes métriques d'évaluations que dans l'article original du WMM2Seq :

- **La F-mesure des entités** : cela mesure la moyenne harmonique entre la précision et le rappel du système à l'heure de prédire les entités présentes dans l'énoncé de l'utilisateur.
- **L'exactitude jointe** : cela mesure combien de réponses identiques à la référence ont été générées par le système.
- **Le score BLEU** : cela mesure la propension des contextes locaux des énoncés produits à ressembler à ceux de l'énoncé attendu.

Modèles	exactitude jointe	F-1 d'entités	BLEU
WMM2Seq	0,375	0,639	**53,76**
MarCo	0,358	0,561	34,82
GPT-2	**0,540**	**0,840**	46,31
WMM2Seq+inc	0,362	0,580	52,71
WMM2Seq+tours	0,389	0,634	54,23
WMM2Seq+tours+inc	0,380	0,621	52,86
Bi-WMM2Seq	**0,422**	0,689	**55,47**
Bi-WMM2Seq+inc	0,414	0,676	53,62
Bi-WMM2Seq+tours	0,413	0,670	54,79
Bi-WMM2Seq+inc+tours	0,404	**0,703**	54,45

TABLE 5.11 – Résultats sur DSTC2,1

Nous évaluons notre système en terme d'exactitude jointe, de f-mesure sur les entités, et de score BLEU. Cependant, nous y revenons à la fin de ce chapitre, comme certaines réponses du système contiennent des incohérences, une prédiction d'une réponse semblable à celle attendue n'est pas souhaitée dans ce cas. Il est donc important de relativiser la pertinence voire l'utilité d'une mesure telle que BLEU pour évaluer la qualité d'une réponse du système.

5.5.5 Analyse des résultats

Les résultats des évaluations offrent un certain nombre d'informations intéressantes :

- Les résultats avec ce nouveau DSTC2.1 montrent que parvenir à générer des réponses correctes est plus difficile qu'avec DSTC2. En effet, dans des comparaisons récentes, l'exactitude jointe dépasse régulièrement les 0,45 et la f-mesure des entités les 0.80 avec DSTC2. Ce n'est pas le cas avec DSTC2.1.
- Seul GPT-2 semble être capable de gérer correctement et de profiter des informations liées au tour de parole utilisateur suivant, aux incohérences, et au nombre de tours restants au point d'obtenir des résultats considérés comme à l'état de l'art pour DSTC2, sur l'évaluation des entités.
- Comme nous le craignions, MarCo, système expert de MultiWOZ, montre une faible capacité à s'adapter non seulement à un nouveau corpus, mais aussi à de nouvelles informations.
- D'une façon générale, le WMM2Seq original est surpassé par l'ajout du tour de parole suivant. Les informations catégorielles ne semblent pas améliorer le modèle. Cela peut être un défaut d'implémentation (utiliser les informations après le dernier saut, et pas à chaque saut), ou tout simplement que la présence d'une incohérence perturbe le système. Si notre corpus était fiable, sans erreurs, nous serions inquiets. Cependant, DSTC2 comme DSTC2.1 ne peuvent pas être considérés comme des référenciels (aussi appelés *gold*)

vu le nombre d'incohérences qu'ils comportent. Par contre, le fait que les incohérences fassent baisser le score de F-mesure d'entités est plus gênant. Une étude approfondie de ce phénomène serait bienvenue dans des travaux futurs.

— En ce qui concerne la f-mesure d'entités, le Bi-WMM2Seq tire profit de l'ensemble des nouvelles informations combinées (0,70 de f-mesure) et améliore sensiblement le WMM2Seq original (0,64).

5.5.6 Conclusion

En définitive, nous avons entraîné une version "oracle" du Bi-WMM2Seq en ajoutant les trois fonctionnalités décrites dans notre modèle sous la forme d'annotations dans DSTC2, afin de créer DSTC2.1. Ces annotations, en particulier celles sur le tour de parole suivant, améliorent le WMM2Seq original. La seule mesure qui a une vraie signification est la f-mesure des entités, car c'est la seule qui ne soit pas directement liée à la génération d'une réponse, mais plus au pistage d'état et à la compréhension de l'énoncé de l'utilisateur. C'est d'ailleurs celle que nous améliorons le plus grâce aux trois fonctionnalités. Les scores BLEU et d'exactitude jointe ont été ajoutés car ils font partie des mesures utilisées fréquemment dans la littérature. Cependant, comme le corpus DSTC2.1 est plein d'incohérences, nous se considérons pas ces mesures comme produisant des résultats fondamentalement significatifs pour déterminer les performances d'un système de dialogue.
A l'avenir, il est indispensable d'entraîner les modèles les plus performants sur ce nouveau jeu de données sur, comme cela a été fait avec GPT-2. Une étude faite par [Weng et al., 2020] a montré que sur DSTC2, les meilleures performances de GPT-2 étaient de 0,813 pour la f-mesure des entités. Il est possible que nos fonctionnalités permettent à ce modèle de mieux inférer les entités présentes dans l'énoncé de l'utilisateur.
Il serait également important, en plus de pratiquer une personnalisation des hyperparamètres de Bi-WMM2Seq, de réfléchir à une optimisation de l'architecture, notamment sur l'endroit où placer les informations catégorielles, ainsi que sur l'intérêt d'encoder le tour de parole suivant. Peut-être vaudrait-il mieux ne se servir de ce dernier qu'au moment du décodage, comme faisant déjà partie de la mémoire épisodique, au lieu de l'encoder comme s'il faisait partie de l'historique de dialogue.

Chapitre 6

Discussion

> We cannot see beyond the choices
> we don't understand
>
> the Oracle

Dans ce dernier chapitre, nous discutons des travaux réalisés, des résultats obtenus et des pistes futures pour creuser la question.

6.1 Des modèles de l'interlocuteur et de la réflexivité des agents dialogiques

6.1.1 Une interaction humain-machine définie par l'information

A travers l'analyse de la croissance d'entropie et de la gestion des erreurs, nous avons démontré que l'information durant un dialogue ne s'échangeait ni au même rythme, ni au même moment, selon s'il s'agissait d'un dialogue entièrement humain ou un dialogue humain-machine. La courbe de croissance d'entropie, dans les dialogues humains, est sensiblement identique pour les énoncés de l'agent et ceux de l'utilisateur. Il y a fort à parier que cela soit lié au fait que les participants humains possèdent des capacités mnnémoniques non seulement similaires mais également complémentaires, qui leur permettent de prévoir la suite du dialogue, ce dernier étant moins incertain. Dans le cas du dialogue humain-machine, nous avons observé comment la croissance d'entropie des énoncés du système diverge de celle des énoncés utilisateur, en étant supérieure, montrant là l'imprévisibilité des réponses du système. Celle-ci est non seulement due aux nombreuses incohérences dans les énoncés du système que nous avons relevées et appris à prédire, mais également au fait que le système n'apprend pas explicitement la fonction lui permettant de se

mettre à la place de l'utilisateur. Notre Bi-WMM2Seq peut remplir cette fonction en fournissant au modèle les informations nécessaires pour explicitement apprendre à modéliser une représentation de l'utilisateur et de lui-même. La question que l'on peut se poser, dans la mesure où cette étude a été réalisée sur un corpus bien spécifique (DSTC2.1), si l'on créait un corpus à partir du Bi-WMM2Seq, observerait-on un changement dans cette courbe d'entropie ?

6.1.2 Une nouvelle mesure d'évaluation ?

Dans la mesure où le but du Bi-WMM2Seq n'est pas de devenir excellent en score BLEU ou en exactitude, mais de déterminer les passages "délicats" dans un dialogue, où le système apparaît le plus susceptible de générer des incohérences et de les corriger, comment mesurer l'impact positif du Bi-WMM2Seq dans un système de dialogue de façon automatique ?
En effet, dans le cas d'une évaluation humaine, il n'est pas déraisonnable d'imaginer que l'on pourrait vérifier la qualité du détecteur d'incohérences interne au Bi-WMM2Seq, l'approximation de la taille restante du dialogue, ainsi que la décision finale du système : générer une réponse, ou renvoyer un message signalant l'incapacité du système à répondre à la demande de l'utilisateur. L'échec de l'accession à la demande de l'utilisateur se définit à deux niveaux : soit localement (demande de reformulation), soit définitivement (fin du dialogue et redirection vers un agent humain).
En revanche, comment évaluer automatiquement un système qui ne sert pas tant à optimiser la génération d'une réponse qu'à tenter de comprendre à quel moment il peut en générer une ou pas ? Dans la mesure où certaines des réponses de référence du système sont truffées d'incohérences, un score BLEU élevé ne signifie pas autre chose que le fait que notre modèle est capable de générer une réponse similaire. Par contre, comme nous possédons un détecteur d'incohérences, nous pouvons évaluer la qualité de la réponse en fonction des trois informations que nous avons rajoutées : le modèle est-il capable de correctement prévoir le nombre de tours de parole, et est-ce que la réponse générée contient une incohérence ? Le modèle peut ensuite optimiser la fonction qui minimise le nombre de réponses incohérentes. Il est probable que l'on puisse proposer une version DSTC2.1 où le Bi-WMM2Seq a corrigé certaines incohérences, ou a remplacé la réponse incohérente initiale par "je ne peux pas répondre à votre demande". Le score BLEU, pour cette version améliorée du corpus de base, prendrait alors plus de sens.

6.1.3 Réponses alternatives ou multiples réponses ?

Nous venons d'évoquer le cas des réponses alternatives : soit une version corrigée d'une réponse à l'origine incohérente, soit une réponse par défaut signifiant l'incapacité du système à poursuivre localement ou globalement le dialogue. Il serait

à notre avis important d'apprendre à un système à différencier les erreurs causées par des incohérences qui sont fatales (fin du dialogue inévitable) de celles qui sont rattrapables (demande de reformulation ou passage à une autre requête de la part de l'utilisateur), un peu à l'image de ce qu'ont réalisé [Filisko and Seneff, 2004] pour le système **MERCURY** [Seneff and Polifroni, 2000]. D'autre part, comme mentionné en 4.5.1, les multiples réponses possibles à un même énoncé de l'utilisateur restent une piste intéressante dans laquelle le calcul de l'entropie prendrait encore plus de poids. En effet, si comme dans [Zhang et al., 2020b], nous exploitions plusieurs réponses possibles à un même énoncé, toutes valables, et que nous les combinions en une seule, nous serions capables de fournir un énoncé système plus riche en informations, et peut-être plus humain que le côté "une chose à la fois" des réponses dans les systèmes actuels. Cette question des multiples réponses est déjà partiellement traitée par le système de sauts dans les réseaux de mémoire, mais il n'y a pas une fonction de combinaison de ces réponses candidates, il existe juste une fonction $SOFTMAX$ qui détermine lequel des sauts fournit la réponse la plus acceptable.

6.1.4 De la prédiction de la taille restante du dialogue

L'impact de la taille restante du dialogue est moins déterminant que ce que nous avions pensé. Il est évident qu'il existe une corrélation entre le nombre de tours de parole restants, l'apparition des incohérences et le succès ou l'échec du dialogue. Cependant, nous nous en apercevons dans l'expérience où nous cherchons à prédire ce nombre, il est extrêmement difficile de savoir pour un modèle pourtant performant dans la tâche de classification de savoir combien de tours de parole reste-t-il au delà de quatre tours restants. La précision est excellente lorsqu'il reste un ou deux tours, correcte lorsqu'il en reste trois ou quatre, puis c'est une chance sur deux lorsqu'il y en a entre cinq et neuf, et enfin la précision chute totalement, et ce quel que soit le corpus sur lequel la tâche est réalisée. C'est tout à fait logique : nous, humains, serions parfaitement incapables de savoir combien de tours de parole il reste en plein milieu d'un dialogue. Si le tour de parole commençait par "Bonjour, bienvenue...", comment pourrions-nous déterminer sa fin ? Nous répondrions qu'il en reste beaucoup.
En revanche, si le tour de parole commençait par "avez-vous besoin de quelque chose d'autre ?" nous aurions tendance à affirmer qu'il reste quelques tours de parole (un, deux...trois peut-être si l'utilisateur a une dernière question ?), ce qui correspond aux scores obtenus par notre modèle Bi-LSTM.

6.1.5 De l'apport du double modèle

A l'inverse, l'impact du double modèle incarné par le Bi-WMM2Seq interne est indéniable. Que ce soit pour prédire la taille restante du dialogue ou pour générer

des réponses de meilleure qualité, l'information sur le tour de parole suivant de l'utilisateur est indispensable. Cela paraît évident pour l'humain : si l'agent connaît le tour de parole suivant de l'utilisateur, il est plus à même de fournir une réponse pertinente (corrigée) au tour de parole courant. Toutefois, pour la machine, nous doutions de la bonne utilisation de cette information. En effet, l'ajout du tour de parole suivant, une sorte de projection hypothétique sur l'avenir proche, aurait pu perturber l'apprentissage du modèle, en lui introduisant du bruit, et une incapacité à correctement exploiter les informationsqu'il possède. Il n'en est rien, la prédiction du tour de parole suivant de l'utilisateur aide le système. De la même façon que nous, humains, lorsque nous dialoguons, nous nous forgeons une image de l'autre, et une image de l'autre en train de nous représenter, le système utilise ce double modèle pour améliorer la qualité de ses réponses.

6.1.6 Hypothèses sur les meilleures performances

Les meilleures performances sont atteintes lorsque les trois informations sont ajoutées dans le modèle, au moment du décodage de chaque saut dans la mémoire épisodique. Dans la mesure où parmi ces informations nous trouvons la présence ou non d'une incohérence, il est normal que chaque saut y ait accès. Cependant, chaque saut fournissant une réponse-candidat, il n'est pas impossible que la présence ou non d'une incohérence doive être déterminée par autre chose qu'une valeur booléenne. En effet, un score flou, une probabilité, un taux, seraient plus pertinents, qu'un simple OUI ou NON. De fait, le modèle ne semble pas s'en contenter puisque l'ajout de cette fonctionnalité ne contribue pas en soi à rendre meilleures les réponses de notre système. D'autre part, nous n'utilisons pas cette information, ni dans l'encodage du contexte, ni dans la mémoire sémantique. Nous pouvons nous demander quel impact cela pourrait avoir sur l'apprentissage de l'un et de l'autre, est-ce que cela permettrait au système de mieux choisir dans la base de données quelles informations sont cohérentes ? Il existe nécessairement une corrélation entre la mémoire épisodique et la mémoire sémantique, bien que nous ne soyons pas parvenus à l'expliciter. Nous pouvons nous demander si ce lien peut être représenté par le partage de l'information sur la présence ou non d'une incohérence.

6.1.7 Vers un modèle réflexif autonome ?

Dans un modèle comme le Bi-Wmm2Seq, où la présence d'un WMM2Seq interne dont la sortie sert d'entrée pour le calcul de la cohérence du WMM2Seq global, nous pouvons nous poser la question de la réflexivité d'un tel modèle. En effet, pour des raisons de temps de calcul et de manque de recul sur l'efficacité, nous avons décidé de ne simuler qu'une seule couche de WMM2Seq imbriqué. Cependant, il est tout à fait envisageable d'intégrer une couche interne au WMM2Seq interne, qui serait la prédiction du tour de parole suivant du système. En fonction

du nombre de tours de parole restants, le système prédirait plus ou moins un tour de parole précis. La question est de savoir si ce nouveau WMM2Seq interne, une sorte de double-double modèle est suffisant. Jusqu'où devons nous implémenter des WMM2Seq ? Nous avons prouvé en créant le premier WMM2Seq interne, que le modèle original n'élabore pas ce travail d'anticipation du tour de parole suivant, et qu'il faut explicitement lui faire apprendre ce tour de parole suivant afin qu'il l'exploite. Cependant, nous ne savons pas précisément combien l'humain anticipe de tours de parole. Est-ce un seul, est-ce quatre ? Nous pouvons nous demander si la création d'un modèle dont le but est de découvrir ce nombre (combien de tours de parole faut-il anticiper afin de générer la meilleure réponse possible au tour de parole courant ?) ne permettrait pas non seulement d'améliorer les performances d'un système de dialogue mais aussi lui donnerait une autonomie dans la gestion de l'information.

6.2 De la mémoire de travail

La place de la mémoire de travail dans nos études a été essentielle. Elle l'a été tout d'abord pour comprendre le fonctionnement des mémoires humaines, et le rôle central de celle de travail dans leur assemblage. Puis, elle l'a été pour comprendre le modèle WMM2Seq, et l'interdépendance entre les mémoires épisodique, sémantique et à court-terme.
Dans les modèles cognitifs tels que présentés dans le chapitre 3, nous observions comment la mémoire de travail jouait un rôle à la fois de tampon entre la mémoire perceptive et les mémoires déclaratives à long-terme, et de système central exécutif, choisissant à tout moment à partir de quel type d'information (épisode ou savoir) devons-nous effectuer l'action correspondante à notre perception. Ce système central exécutif applique également une fonction d'oubli, pour ne conserver en mémoire tampon que les informations essentielles à l'optimisation de la perception - et donc du choix de l'action - suivante, et une fonction d'enregistrement, pour amorcer le processus de conservation en mémoire épisodique des couples perception-action les plus saillants et marquants, afin de se montrer plus performant lorsqu'une situation similaire se présente. Dans le cas de nos travaux, la mémoire de travail du WMM2Seq ne possède ni fonction d'oubli, ni fonction d'enregistrement en mémoire épisodique. Nous avons entrevu la question lorsque nous avons évoqué la fonction de détection d'incohérence, et de prédiction du tour de parole suivant, qui pourraient servir de fonction d'oubli : en effet, la détection d'incohérence permettrait de considérer que l'énoncé courant de l'utilisateur, et sa compréhension par le système (soit par le biais d'une prédiction des actes de dialogues et de l'état de croyance, soit par le biais d'un score d'approximation sémantique telle que la similarité cosinus) ne doivent pas être enregistrés et sont à supprimer de la mémoire tampon. Cependant, il serait judicieux d'envisager une

étude approfondie de la fonction cognitive d'oubli dans le dialogue, son apport dans une conversation orientée-tâche, sa matérialisation dans les corpus humains existants (tels que MultiWOZ) et l'impact de l'intégration d'une fonction d'oubli explicite dans un modèle appris sur ce corpus.
Par ailleurs, la notion temporelle, mise de côté dans le WMM2Seq, à la fois dans la gestion de l'historique de dialogue et dans l'anticipation de la fin du dialogue, nous semble un manque essentiel dans l'architecture d'un système de dialogue. En effet, lors de la gestion de l'historique de dialogue, il est important d'avoir en tête qu'un réseau de mémoire, à l'instar d'un réseau récurrent, retient moins bien les informations présentes en début de dialogue, en particulier lorsque l'historique commence à posséder une taille conséquente. Les transformeurs tels que GPT-2 semblent réussir à apprendre une fonction de filtrage efficace de l'historique de dialogue. Toutefois, on perdrait l'intérêt du WMM2Seq, à savoir la séparation de la gestion des informations en différentes mémoires, toutes liées mais différenciées dans l'architecture. En ce qui concerne l'anticipation de la fin du dialogue, il serait judicieux d'expérimenter si la connaissance ou au moins la prévision du nombre de tours de parole restants favorise la détection des incohérences. Nous avons observé que l'inverse est vrai. Cependant, si la réciproque l'est aussi, cela corrobore notre hypothèse que la dimension temporelle est primordiale pour l'optimisation d'un système de dialogue.

6.3 Considérations éthiques et légales

Nous abordons maintenant la problématique éthique de la conception d'un système de dialogue ainsi que les limites légales de l'apprentissage automatique de ce dernier. Nous considérons dans l'éthique à la fois la partie morale, mais également la partie matérielle, à savoir quel est le coût, financier et environnemental, de l'apprentissage d'un Bi-WMM2Seq [Jobin et al., 2019] . Au niveau légal, nous nous intéressons au déploiement industriel d'une telle solution et des conséquences que cela peut engendrer, aussi bien au niveau de la nécessité de l'anonymisation, mais également au niveau de la législation européenne régie par le récent RGPD (Réglement Général sur la Protection des Données).

6.3.1 Energie et consommation

Plusieurs études ont montré que le traitement automatique des langues, en particulier depuis l'avènement des modèles de langue basés sur les transformeurs, fait exploser tous les compteurs d'empreinte carbone. [Strubell et al., 2019] ont montré que les expérimentations et l'optimisation à partir d'un simple algorithme d'apprentissage automatique pour une tâche telle que la classification ou la traduction, possède une empreinte carbone deux fois plus élevée que le mode de vie

d'un Américain pendant un an. Ils ont également avancé que ces mêmes expérimentations et optimisations à partir d'un transformeur polluent six fois plus qu'une voiture durant toute son existence (essence incluse). Dans la mesure où la tâche de dialogue, par sa nature interactive, nécessite non seulement un apprentissage lourd, mais également des réapprentissages fréquents et une capacité computationnelle élevée en mode déploiement, nous pouvons nous demander jusqu'où peuvent s'élever ces chiffres. Il est vrai que depuis cette étude, certains modèles "éthiques" ont vu le jour. En ce qui concerne les transformeurs, la communauté a cherché à développer des architectures éco-responsables [Sanh et al., 2019, Choromanski et al., 2020]. Il faut entendre ici éco-responsable par nécessitant moins de temps de calcul donc moins de consommation d'énergie. En ce qui concerne le dialogue, la boîte à outils RASA a sorti récemment une nouvelle architecture appelée DIET [Bunk et al., 2020], censée consommer moins qu'un transformeur optimisé pour le dialogue tel que DialoBERT ou AuGPT. Dans la tâche de question-réponse, Enfin, dans la prédiction des actes de dialogue, [Kumar et al., 2020] ont développé une solution pour réaliser cette tâche en consommant moins que les modèles état-de-l'art malgré des performances similaires. Cependant, aucune de ces solutions n'a été testée et évaluée en mode déploiement.

6.3.2 RGPD et système de dialogue

Cette section est tirée d'un article publié dans l'atelier LEGAL'20 de la conférence LREC2020 [Schaub et al., 2020].
En France, la loi sur la protection des données personnelles n'est pas une idée nouvelle. En 1978, la loi Informatique et Libertés (LIL) a été votée. En même temps, la Commission nationale de l'informatique et des libertés (CNIL) a été créée. Une agence est rarement la solution ultime pour résoudre tous les problèmes à la fois, tout dépend de l'équilibre délicat entre le périmètre de sa mission et ses moyens d'action. Comme d'autres institutions de ce type, la CNIL a parfois été critiquée pour son manque de transparence ou sa mauvaise communication avec les autres organismes de sécurité publique ou d'application de la loi, surtout depuis les années 1990 et la directive européenne de 1995, qui a donné plus de pouvoir à la CNIL en réponse au développement d'Internet.
Cependant, au cours des quinze dernières années et de l'essor des réseaux sociaux, la technologie numérique a révolutionné à la fois la vie quotidienne des personnes [1] et les pratiques commerciales des entreprises [Bonchi et al., 2011]. C'est pourquoi en 2016, le Parlement européen a voté le RGPD afin de protéger la vie privée des individus et d'empêcher l'utilisation abusive des informations personnelles. Voici les principales évolutions apportées par cette nouvelle loi :

— La transparence devient une obligation [Goddard, 2017]
— Les responsabilités sont rééquilibrées [Lindqvist, 2017]

1. http ://www.comonsense.fr/influence-medias-sociaux-vie-quotidienne/

— De nouveaux concepts sont créés ou instanciés : profilage, droit à l'oubli,*Protection de la vie privée dès la conception*. [Spiekermann, 2012]

L'intelligence artificielle (IA) qui se cache derrière les techniques de fouille de texte est analytique : elle prend des données en entrée et, en fonction de tous les textes que l'IA a vus auparavant, elle applique un algorithme (classification, traduction, analyse syntaxique...) en fonction de la ou des tâches pour lesquelles elle a été créée. Cependant, nos études se concentrent sur une technologie qui utilise non seulement l'analyse, mais aussi l'interaction humain-machine (IHM) : les systèmes de dialogue et plus précisément les systèmes de dialogue orientés tâche (tods).
Le problème des systèmes de dialogue orientés tâche avec le RGPD et la gestion des données est l'interaction en temps réel. En effet, que la tâche de fouille de texte soit une analyse de sentiments, une analyse syntaxique de dépendances ou le calcul d'une réponse à une question, l'anonymisation des données personnelles n'est pas un problème de même nature que pour le dialogue en ce qui concerne l'obtention de bonnes performances, car pour ces tâches l'IA n'a pas besoin d'avoir une quelconque interaction avec l'utilisateur. La principale différence avec la tâche de dialogue est la nécessité pour le système de dialogue orienté tâche de connaître son utilisateur pour améliorer l'acceptation humaine. Par exemple, [Tahara et al., 2019] améliorent la satisfaction de l'utilisateur en temps réel en apprenant l'intégration des émotions pour avoir une meilleure compréhension humaine.

Alors que le RGPD a commencé à être appliqué en 2019, de nombreuses entreprises et même des laboratoires de recherche travaillant sur les données textuelles ont concentré leurs travaux sur la recherche de la meilleure façon d'anonymiser les documents. [Di Cerbo and Trabelsi, 2018] propose un aperçu des techniques classiques d'anonymisation des textes et une nouvelle approche basée sur des algorithmes d'apprentissage automatique de pointe. [Kleinberg et al., 2018] a développé un logiciel libre d'anonymisation des entités nommées appelé NETANOS. Plus récemment, [Kim et al., 2019] a introduit un protocole pour anonymiser correctement les données, afin d'être totalement conforme au RGPD en montrant des améliorations des techniques d'anonymisation. Cependant, comme l'explique [Bottis and Bouchagiar, 2018], il est très difficile, voire impossible, d'anonymiser parfaitement toutes les données personnelles en raison des améliorations constantes des techniques de ré-identification et donc de la nécessité de faire évoluer périodiquement l'anonymiseur [Hayes et al., 2017].
Une fois encore, supposons qu'il existe un anonymiseur parfait. En effet, un système de dialogue orienté tâche alimenté par des entrées anonymisées est conforme au RGPD, mais il perd alors la capacité de se souvenir d'informations cruciales au cours d'une conversation orientée objectif, comme l'identité de son interlocuteur.
Pour la phase d'apprentissage, où l'IA derrière le système de dialogue orienté tâche apprend des conversations passées, l'anonymisation n'est pas un problème, tant que la structure originale de la conversation est conservée, afin d'être similaire à la

conversation réelle à laquelle le système de dialogue orienté tâche devra faire face pendant la phase de déploiement/évaluation. L'anonymisation pourrait cependant être problématique pour les nouvelles conversations, car nous savons que l'une des principales conditions pour qu'une machine soit conviviale est de ressembler à l'humain [Ouali et al., 2019], et nous doutons que le fait d'avoir un système de dialogue orienté tâche amnésique soit un moyen de simuler la vraisemblance et l'empathie humaines. Comme nous l'avons expliqué précédemment, un bon employé doit faire preuve d'empathie pendant le dialogue afin d'augmenter la probabilité de satisfaction du client.
Supposons maintenant que le client ne se soucie pas, pendant la conversation, de savoir si le système de dialogue orienté tâche fait preuve d'empathie ou non. Il existe au moins deux scénarios dans lesquels une anonymisation complète des données reste un problème.

Rappel du client

Imaginons la situation où un client C, après avoir terminé une conversation avec un agent A, appelle quelque temps plus tard, pour une raison quelconque, le même service et que ce soit le même agent qui réponde à l'appel. Dans une situation normale, si les deux appels sont effectués dans la même heure, C s'attend à ce que A se souvienne de l'appel ou au moins d'une information qui s'y rapporte, telle que : la raison du premier appel, le nom de C, et éventuellement les problèmes les plus importants rencontrés. Dans la plupart des cas, la satisfaction de C sera corrélée à la capacité de A à se souvenir de lui. Même si un système de dialogue orienté tâche B est bien formé sur ce que nous avons appelé le premier appel, il peut rencontrer des difficultés pour satisfaire le client lors de la conversation du deuxième appel. Il y a un déséquilibre entre les attentes de C et les capacités de B. Même si lors du premier appel, C n'a pas eu besoin de signes d'empathie de la part de B, lors du deuxième appel, ce sera différent car un lien existe déjà entre C et B, du point de vue de C. Mais à cause de l'anonymisation, même si B comprend qu'il s'agit d'une situation de second appel, il ne sera jamais capable de reconnaître C comme l'auteur d'un appel précédent.

Récurrence des données personnelles

Cette deuxième situation n'est pas un handicap rédhibitoire comme la précédente mais la technologie du système de dialogue orienté tâche apporterait une amélioration si elle avait une solution à cette situation.

Imaginons la situation où un fournisseur de services Internet reçoive des milliers d'appels le même jour parce qu'il y a une énorme panne dans une zone spécifique.

Après plusieurs appels de clients frustrés, lorsqu'un nouveau client C appelle avec le même ton ou écrit un courriel avec une sémantique similaire à celle des précédents, un agent A sait sans même demander de quoi il s'agit : la panne, l'endroit où elle s'est produite, et même les plaintes de C. Cette capacité d'inférence permet à A d'être plus efficace lors de la nouvelle conversation et de fournir à C toutes les informations nécessaires. De plus, A sait comment calmer C après avoir expérimenté des techniques avec des clients mécontents précédents toute la journée. Maintenant, si l'agent est plutôt notre système de dialogue orienté tâche B, cette amélioration d'un jour seulement est impossible en raison de l'anonymisation : dans le RGPD il est explicitement dit que toute information qui peut identifier une personne doit être transformée. Cela inclut la localisation du client et son état émotionnel. Il n'y a donc aucun moyen pour B, même s'il avait une mémoire d'un jour, de relier les plaintes précédentes à celle de C. Dans la mémoire de B, ce sera une coïncidence étonnante que la même panne se produise tant de fois ce jour.
Cependant, il est utile de préciser que désormais le RGPD prend cela en compte en imposant des durées maximales de conservation des données en fonction de leur type, mais cela nécessite des réajustements fréquents pour suivre l'évolution de la technologie et des pratiques courantes. Une infrastructure légale a du mal à mettre en place dans une société dont les moyens de communication sont en constante et rapide évolution.

6.3.3 Protocole de système de dialogue conforme au RGPD (PCCR)

Afin de répondre au besoin de protection des données, nous avons mis en place un protocole industriel appelé PCCR (Protocole de *Chatbot* Conforme au RGPD) qui est conforme aux normes du RGPD. La figure 6.1 schématise un système de dialogue respectant les normes d'anonymisation et d'oubli des données dites personnelles.

Voici les principaux composants de l'architecture du système de dialogue orienté tâche :
a. Interface utilisateur-bot UBI : il peut s'agir d'une boîte de dialogue comme Messenger ou d'un outil intégré à la plateforme de gestion de relation client (CRM pour *customer relationship management*).
b. Anonymiseur : un module externe utilisé pour anonymiser l'entrée d'un utilisateur à chaque tour de parole. Il peut être entraîné ou utiliser des règles a priori.
c. User database and API : il s'agit de la base de données clients (CRM) utilisée pour récupérer les données du profil de l'utilisateur si le chatbot en a besoin.
d. Mémoire de dialogue : c'est le premier et le plus petit des trois modules internes du système de dialogue orienté tâche, son but est de prétraiter les entrées anonymes sous forme de dictionnaire. Il est effacé après chaque tour de parole.
d. Mémoire de travail : c'est le deuxième module du système de dialogue orienté

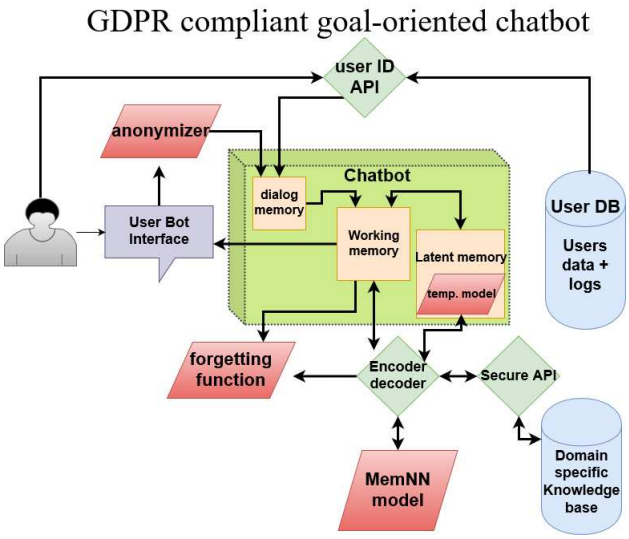

FIGURE 6.1 – *chatbot* conforme au RGPD. La figure est en anglais car le protocole a été écrit pour les normes européennes.

tâche et le cur du processus car c'est là que les souvenirs de la conversation sont enrichis par le Modèle de dialogue en réseau mémorisé (h.) et la base de données externe (i.) mais aussi par la mémoire latente (e.). Une fois la conversation terminée, la MM oublie (g.) les données personnelles récupérées au début de la conversation et encode la représentation de la conversation ainsi que le résultat du dialogue dans la mémoire latente. C'est également dans ce module que la sortie est générée et envoyée à l'UBI à chaque tour de parole. La mémoire latente est limitée et les anciennes conversations qu'elle contient sont effacées au bout d'un certain temps pour laisser place aux nouvelles.

e. Mémoire latente : c'est le troisième module du système de dialogue orienté tâche et il contient toutes les conversations que la mémoire de travail a effacées, mais sans les données personnelles. Sa capacité est également limitée mais est beaucoup plus importante que celle de la mémoire de travail. Cependant, elle n'a pas de fonction de traitement active. Lorsque sa taille atteint un certain seuil, elle apprend un modèle temporaire (neuronal) des conversations de la journée ainsi que le résultat du dialogue. Au cours d'une conversation, elle est utilisée par la mémoire de travail comme source concurrente de récupération d'informations du MNDM. Ce modèle temporaire est effacé à la fin de la journée.

f. Modèle temporaire C'est le modèle appris par la mémoire latente après qu'elle ait vu suffisamment de conversations pour le faire. À la fin de la journée, le modèle est encodé dans la MNDM pour l'améliorer, puis il est effacé.

g. Oubli : il s'agit d'une fonction apprise utilisée deux fois au cours du processus : d'abord à la fin d'une conversation pour purger toute donnée personnelle laissée dans la MM avant qu'elle ne se connecte à la mémoire latente, et ensuite à la fin

de la journée pour supprimer toute information non pertinente ou pour vérifier qu'il ne reste aucune donnée personnelle dans les données du modèle temporaire.
h. Modèle de dialogue en réseau mémorisé MNDM : c'est le modèle représentant toutes les conversations passées (corpus de dialogue) apprises. Il s'agit de la mémoire à long terme du système de dialogue orienté tâche. Dans l'architecture, il s'agit d'un modèle externe au cas où, pour une raison quelconque, le système de dialogue orienté tâche doive adopter un comportement différent de celui appris par le modèle. Elle s'inspire de [Zhang et al., 2019c]
. **i. Base de connaissances externe EKB** : c'est le système d'information fourni par un client du domaine tel que la liste des produits ou la documentation officielle. Elle représente la mémoire sémantique du système de dialogue orienté tâche et doit être déconnectée du MDNM parce que le même MDNM peut être utilisé pour différentes EKB et pour éviter que le système de dialogue orienté tâche devienne trop spécifique au domaine. Comme nous l'avons montré dans la section 2, en raison de l'opacité des modèles de pointe dans les systèmes de dialogue, il pourrait être difficile de construire une architecture entièrement de bout en bout, pour des raisons de sécurité, malgré leurs avantages tels que la vitesse d'apprentissage et la taille du modèle [Rajendran et al., 2018, Rajendran et al., 2019]. Cependant, ce que nous appelons la mémoire à long terme du système de dialogue orienté tâche, qui représente le modèle neuronal appris des conversations passées, peut être un système de bout en bout [Wu et al., 2018]. Dans notre architecture, le système de dialogue orienté tâche ne contient pas lui-même la mémoire à long terme, ni l'outil d'anonymisation, ni la base de connaissances externe spécifique au domaine.

6.3.4 Définir le PCCR

0. La première étape, et non la moindre, consiste à **demander aux utilisateurs s'ils acceptent** que les données recueillies au cours de la conversation puissent être utilisées par la suite pour améliorer le système de dialogue orienté tâche.
1. Comme nous l'avons vu dans la section 2, l'anonymisation est une étape nécessaire dans la conception du système de dialogue orienté tâche. Elle est appelée *Protection de la vie privée dès la conception*. L'anonymiseur doit être appelé pour chaque entrée de l'utilisateur.
2. Cependant, si la nature de la tâche nécessite certaines données personnelles telles qu'un courriel afin d'identifier le dossier ou le compte de l'utilisateur, ou une carte d'embarquement etc., le système de dialogue orienté tâche doit être capable de **récupérer ces informations dans la base de données de l'utilisateur**. Pour ce faire, l'architecture doit mettre en uvre une API avec un identifiant d'utilisateur temporaire pour fournir au système de dialogue orienté tâche toutes les informations dont il a besoin pour remplir son objectif.

3. **L'UID est stocké** dans la mémoire de dialogue du système de dialogue orienté tâche pendant le premier tour de parole et envoyé à la MM.
4. Les **données d'entrée anonymisées sont ensuite encodées dans le module MNDM** et **la mémoire latente.** Pendant le décodage, un appel API est effectué vers la connaissance externe au cas où cela serait nécessaire pour la conversation ou si un appel API devait être effectué. Le résultat du décodage est la sortie du tour de parole. Le **processus est répété tout au long de la conversation.**
5. A la fin de la conversation, **l'UID est conservé dans la mémoire de travail** pendant quelques minutes (jusqu'à une heure) au cas où le même utilisateur engage une nouvelle conversation pendant ce temps.
6. Après ce délai, **la fonction d'oubli est alors appelée** afin de supprimer de la mémoire de travail toutes les données personnelles. Celles-ci sont stockées dans la mémoire latente représentant les conversations quotidiennes.
7. Lorsque la mémoire latente commence à devenir plus grande, un modèle de conversation quotidienne est appris par le système de dialogue orienté tâche, afin de savoir s'il existe des tendances particulières ce jour-là qui devraient être saillantes pour la mémoire de travail du système de dialogue orienté tâche.
8. À la fin de la journée, le modèle du jour est encodé dans la mémoire à long terme et la fonction d'oubli est appelée, au cas où des données personnelles seraient malheureusement restées dans la mémoire latente.
9. Enfin, le **MNDM est réentraîné** avec les nouvelles conversations de la journée. En suivant ce protocole, le système de dialogue orienté tâche est à la fois conforme au RGPD et efficace pour n'importe quelle tâche.

6.3.5 Limites du PCCR

Bien que le PCCR semble obéir aux règles du RGPD, il existe plusieurs limites qu'il convient de mentionner. Comme nous l'avons dit dans la section 2, il n'existe pas d'anonymiseur parfait, et même si les nouveaux modèles deviennent très précis, il existe toujours la possibilité que certaines informations sensibles ne soient pas enlevées par le processus d'anonymisation. Deuxièmement, le modèle linguistique dans lequel les conversations précédentes sont stockées est également appris, et peut donc également contenir des données personnelles. Si, dans de nombreux cas, l'ajout de nouvelles conversations améliore l'efficacité du modèle, il peut également augmenter le risque que des données personnelles soient produites pendant l'inférence. Enfin, comme le système de dialogue orienté tâche est disponible en ligne, il peut y avoir un problème de sécurité lorsqu'il fait des appels API à la base de données de l'utilisateur. Une étude doit être réalisée afin de vérifier si le danger pour la sécurité est réel ou non.

6.4 Le problème de l'anonymisation

Le problème de l'anonymisation des données clients est plus ancien que le RGPD. [Zhong et al., 2005] montrent l'efficacité de la k-anonymisation pour la confidentialité des clients pendant un processus automatique. La k-anonymisation est définie comme suit : "Étant donné des données structurées par champs spécifiques à une personne, produire une libération des données avec des garanties scientifiques que les individus qui sont les sujets des données ne peuvent pas être ré-identifiés alors que les données restent pratiquement utiles" [Samarati and Sweeney, 1998]. Bien que [Nergiz and Clifton, 2007] ait surpassé la k-anonymisation avec des algorithmes basés sur le clustering, ces techniques n'étaient pas efficaces pour anonymiser les données non structurées comme le montre [Angiuli and Waldo, 2016] et le RGPD a introduit plusieurs changements dans la définition d'un texte anonymisé [Hintze and El Emam, 2018]. [Di Cerbo and Trabelsi, 2018] présentent un aperçu des techniques supervisées d'anonymisation.
Dans le domaine médical, les tâches TAL sont largement concernées par le RGPD : [Fraser et al., 2019, Kirinde Gamaarachchige and Inkpen, 2019, Berg and Dalianis, 2019, Dalianis, 2019]. [Chevrier et al., 2019] proposent une enquête sur les techniques et les problèmes spécifiques de l'anonymisation pour les ensembles de données médicales. [Goddard, 2017] proposent une approche de regroupement pour l'anonymisation des rapports médicaux afin de limiter la perte d'information et préserver l'utilité des données.
Dans le domaine didactique, [Megyesi et al., 2018] construit un corpus conforme au RGPD pour l'apprentissage des langues étrangères : leur méthode peut être partiellement réutilisée dans de nombreux domaines grâce à l'anonymisation complète des entités nommées qu'ils réalisent.
Plusieurs outils en code ouvert sont apparus récemment pour anonymiser les textes selon le RGPD [Adams et al., 2019, Kleinberg et al., 2017]. Néanmoins, à notre connaissance, il n'existe pas d'approche formelle de l'anonymisation des textes pour les tâches basées sur la fouille d'opinion dans le domaine de la gestion de la relation client.

6.4.1 Conditions requises pour des données anonymisées réutilisables

Dans le domaine de la relation-client, le sujet a été abordé par [Francopoulo and Schaub, 2020] dont nous nous inspirons pour la suite de cette section.

Les principales exigences fonctionnelles sont les suivantes :
— REQ#1 Éviter d'identifier les personnes mentionnées dans le texte,
— REQ#2 Permettre une analyse des données sémantiques en interne qui pourrait éventuellement être adaptée à un certain type d'entrée,

— REQ#3. Permettre l'utilisation d'outils TAL standard,
— REQ#4 Prouver qu'une anonymisation a été effectuée en cas de plainte d'une personne mentionnée dans un texte spécifique ou en cas de procès ou d'enquête de journalistes,
— REQ#5 Utilisable dans différentes langues européennes.

Ces exigences sont en quelque sorte contradictoires. Par exemple, dans le texte original :

Mon nom est Paul Smith, et j'ai déménagé de Leeds à Paris.

une anonymisation supprimera toutes les informations identifiables et produira :

Mon nom est X et j'ai déménagé de X à X.

Dans ce cas, les critères d'évaluation 1 et 4 sont remplis, mais le traitement sémantique des critères 2 et 3 sera profondément perturbé. Une autre option pourrait être de remplacer un nom par un nom aléatoire issu d'un dictionnaire tout en respectant le type du nom comme :

Mon nom est John Wilson et j'ai déménagé de Berlin à Madrid.

Dans ce cas, les substituts réalistes donnent l'impression que le texte est original mais la REQ#4 n'est pas remplie. Nous ne pouvons pas nous permettre une pseudonymisation globale car elle n'est pas vraiment une anonymisation sûre (comme mentionné dans l'introduction) mais la pseudonymisation locale semble un bon compromis remplissant quatre des cinq exigences donnant une phrase comme :

Mon nom est _Personne1 et j'ai déménagé de _Ville1 à _Ville2.

Le seul inconvénient de cette approche est que le texte ne peut pas être soumis à un processus de TAL qui n'est pas préparé à ce type d'altération, comme une traduction automatique, et donc la cible REQ#3 est manquée. En fait, **REQ#3 et REQ#4 sont contradictoires**. En réfléchissant à nouveau à ce problème, nous avons réalisé que certaines exigences ne doivent pas être satisfaites en toutes circonstances. La REQ#4 est importante lorsque l'on produit les données en dehors d'un périmètre sécurisé, tandis que la REQ#3 est importante lorsque l'on utilise des outils disponibles sur le marché en interne dans un périmètre sécurisé. Le dilemme peut être résolu en implémentant un paramètre booléen lors de l'exécution de l'anonymisation associée à la réalisation de la REQ#3 ou de la REQ#4. Ainsi, l'anonymisation est capable de produire :

Mon nom est _Personne1 et j'ai déménagé de _Ville1 à _Ville2.

quand il y a un besoin d'extérioriser, ainsi que :

Mon nom est John Wilson et j'ai déménagé de Berlin à Madrid.

en cas de traitement interne, selon l'option choisie. La satisfaction des besoins est résumée dans le tableau 6.1

Req.	substitution by X	global pseudo.	local pseudo.	aléatoire substitution
REQ#1	oui	non	oui	oui
REQ#2	non	oui	oui	oui
REQ#3	non	non	non	oui
REQ#4	oui	oui	oui	non
REQ#5	oui	oui	oui	oui

TABLE 6.1 – Conditions vs solutions

6.4.2 Implementation

L'idée est d'enchaîner trois processus : 1) une reconnaissance d'entité nommée, 2) un lieur d'entité, et 3) une substitution. Ces processus doivent s'exécuter dans un environnement sécurisé et ne doivent pas produire de traces d'exécution qui pourraient rompre l'anonymisation. C'est-à-dire que seul le résultat de la substitution est autorisé à être publié en dehors de l'environnement d'exécution.

La reconnaissance d'entités nommées (REN) est traitée par le détecteur d'entités nommées de la société Akio qui prend la sortie d'un analyseur syntaxique dont le nom est Tagparser [Francopoulo, 2008]. L'analyseur syntaxique combine l'induction statistique et des règles syntaxiques robustes. Le NER est implémenté par une cascade de règles de correspondance de patrons pour détecter les noms d'êtres humains, de lieux, d'entreprises, de marques, d'adresses électroniques et toutes sortes de formes numérales comme les dates, les montants d'argent, les numéros de vol, les IBAN, les numéros de téléphone, les numéros de passeport et les numéros de sécurité sociale[2]. Pour les noms propres, la REN utilise des indices locaux basés sur la langue, combinés à une liste de 1,2 million de noms propres extraits automatiquement de Wikidata. Il s'agit d'un détecteur industriel utilisé pour traiter actuellement une moyenne de 1M de textes chaque jour dans six langues (anglais, français, allemand, italien, espagnol, portugais). Il existe un analyseur spécifique pour chaque langue alors que la plupart des détections d'entités nommées sont neutres du point de vue linguistique, c'est-à-dire qu'elles sont les mêmes dans les six langues couvertes. En fait, seule une petite série de différences culturelles, comme l'identification des véhicules, sont différentes. Le système français n'est pas capable de reconnaître les plaques d'immatriculation allemandes, par exemple, mais les situations où cela est nécessaire sont extrêmement rares. Le programme comprend un correcteur orthographique spécifique pour traiter les entrées mal for-

2. Le lecteur peut reproduire notre travail en utilisant un autre NER à condition que toutes les formes précises et personnelles comme les identifiants de sécurité sociale soient correctement détectées. Évidemment, la qualité de l'ensemble du processus dépend fortement de celle du NER.

matées, basé sur une expérience de 10 ans de collecte d'entrées mal formatées.

Le but de l'outil de liaison des entités est de rassembler les entités nommées apparaissant à différents endroits du texte, éventuellement avec des variations encyclopédiques ou orthographiques. Par exemple, dans 'Nicolas Sarkozy said. . . Sarko replied. . . ' où 'Sarko' est un surnom de 'Nicolas Sarkozy', les deux noms doivent être liés. Un autre exemple est 'N Sarkozy' vs 'Nicolas Sarkozy' où 'N' n'est pas ambigu et doit être considéré comme un prénom. L'objectif est de relier ces énoncés dans une structure commune.

L'objectif de la substitution est de remplacer une sélection de types d'entités qui sont :

- **city** pour les noms de villes et d'agglomérations, comme 'Paris' (une ville) ou 'Cergy-Pontoise' qui n'est pas formellement une ville mais une agglomération.
- **contractNumber** pour la combinaison de chiffres et de lettres qui semble être autre chose qu'un mot ou un chiffre. Cette catégorie comprend certaines catégories personnelles spécifiques comme les IBAN (International Bank Account Number) et les BIC (Bank Identifier Code).
- **courrielAddress** pour les adresses électroniques.
- **personName** pour les noms des individus qui sont des êtres humains.
- **identificationNumber** pour l'identifiant d'un individu comme un numéro de sécurité sociale ou un numéro de passeport.
- **IPAddress** pour les adresses de protocole Internet.
- **phoneNumber** pour les différentes formes d'un numéro de téléphone.
- **vehicleIdentification** pour les plaques d'immatriculation des véhicules.
- **zipCode** pour les codes postaux.

Il convient de noter que la REN détecte d'autres types d'entités comme, par exemple, des pays, des régions, des organisations, des sommes d'argent ou des numéros de vol. Il est évidemment facile, d'un point de vue technique, de substituer ces entités, mais la question est la suivante : quelle est la raison de le faire ? Ces entités sont moins personnelles et, en l'absence d'indices personnels, il y a peu de danger à conserver la chaîne de caractères originale. Étant donné que la ville est remplacée, la localisation exacte ne peut être déterminée, il n'est donc pas nécessaire de substituer l'adresse dans son intégralité, outre le fait que la reconnaissance de la partie du texte indiquant la rue est en générale difficile en raison des nombreuses formes possibles.

6.4.3 Exemple

A partir de ce texte original (inventé) :

Cher Monsieur/Madame,

> Je vous écris aujourd'hui pour vous faire part du problème que je rencontre sur le site www.ameli.fr. J'aimerais créer un compte mais mon identifiant de sécurité sociale 200 11 99 109794 sur ma carte vitale n'est pas le même que celui de ma mutuelle 201 11 99 109794. Comment puis-je faire ?
> Bien à vous,
> Paul Watson,
> tél. 01 23 34 34 56 pwatson@aol.fr

Notez que la carte Vitale est la carte d'assurance maladie du système national de soins de santé en France. L'anonymisation produit le texte suivant, à condition que l'option de pseudonymisation soit sélectionnée :

> Madame, Monsieur,
> Je vous écris aujourd'hui pour vous faire part du problème que je rencontre sur le site www.ameli.fr. Je voudrais créer un compte mais mon identifiant de sécurité sociale _SSid1 sur ma carte vitale n'est pas le même que celui de ma mutuelle _SSid2 Comment faire ?
> Bien à vous,
> _Personne1,
> tel _Téléphone1 _Email1

En raison du fait que la numérotation des pseudonymes est réinitialisée pour chaque nouveau texte, il n'est pas possible d'induire des données personnelles à partir de ce texte ou de faire une quelconque corrélation avec un autre texte, donc le RGPD est respecté. Cependant, à condition que le programme d'analyse numérique soit spécialement adapté pour traiter orthographiquement les pseudonymes et pour interpréter le pseudonyme comme une valeur sémantique d'entité nommée, il est toujours possible de calculer que l'auteur a :

- Une plainte concernant un site web donné,
- Une plainte pour non-concordance entre différents identifiants de sécurité sociale,
- Une question.

Ce résultat est pleinement satisfaisant. C'est typiquement le genre de résultats qui fournit par le produit interne Akio Insight qui met en uvre l'ABSA (Aspect Based Sentiment Analysis) [Pontiki et al., 2014]. Notons que le domaine des systèmes d'anonymisation ou plus précisément de dé-idenification est en pleine expansion. Par exemple, le projet européen MAPA [3] a récemment produit à destination des administrations et du grand public un logiciel en accès libre et gratuit de déidentification pour toutes les langues officielles de l'Union Européenne [Gianola et al., 2020].

3. https://mapa-project.eu

6.4.4 Méthode utilisée pour la validation

La vérification manuelle d'un grand corpus de manière itérative avec des alternances de correction / vérification est une charge très lourde. Notre corpus de test est une collection de 18138 verbatim français issus du secteur d'activité juridique et administratif et nous ne pouvons pas vérifier l'ensemble du corpus après chaque amélioration du détecteur. Nous avons commencé par exclure de manière aléatoire 300 verbatims comme corpus de test, à utiliser par la suite.

L'objectif principal n'est pas d'éviter le bruit mais surtout d'éviter le silence, c'est-à-dire que nous considérons qu'il n'est pas très important qu'une chaîne de caractères soit sur-substituée. Au contraire, l'absence de substitution d'un nom de personne est une erreur grave. Nous utilisons le fait que le texte est transformé après pseudonymisation et si certains noms propres des neuf types restent, il y a de fortes chances qu'il y ait une erreur. Nous avons testé le système pour le français en suivant une triple approche itérative sur le corpus de développement contenant 18138-300=17838 textes :

— A l'étape nř1, le corpus est anonymisé avec l'option de pseudonymisation locale,

— A l'étape nř2 la détection d'entité nommées est à nouveau appliquée et le résultat est filtré pour retenir les entités nommées des neuf types qui ne commencent pas par le caractère underscore, ce caractère identifiant un pseudonyme. Lorsqu'il y a un résultat, il y a de fortes chances qu'il s'agisse d'une erreur.

— A l'étape nř3, les erreurs de reconnaissance d'entités nommées sont corrigées et le processus est appliqué à nouveau à l'étape nř1. On s'arrête quand on n'a pas trouvé d'erreur.

Les différentes phases de la validation sont présentées dans le tableau 6.2.

phases	nb de txt traité	nb d'erreurs
phase-1	17838	284
phase-2	284	53
phase-3	53	0

TABLE 6.2 – Résultats de validation

L'évaluation du corpus de test est présentée à 6.3 :

Nb de texte	Rappel	Precision	FMesure
300	100	99.5	99.7

TABLE 6.3 – Evaluation de la Qualité

La distribution totale sur le corpus est disponible Table 6.4

type	nb d'occ.	distrib.
city	6408	19%
contractNumber	17	0%
courrielAddress	2141	6%
personName	20835	63%
identificationNumber	146	0%
IPAddress	89	0%
phoneNumber	1721	5%
vehicleIdentification	97	0%
zipCode	1687	5%
total	33141	100%

TABLE 6.4 – Distribution des entités-type

6.5 Implémentation industrielle et évaluation humaine du Bi-WMM2Seq

L'implémentation d'un tel modèle dans un système de dialogue de type *chatbot* requiert une étude approfondie au niveau de l'apprentissage ainsi qu'au niveau du déploiement. Nous distinguons trois problématiques :

1. Quelle boîte à outil pour implémenter la solution ?
2. Quel format de données, quelles données ?
3. Comment intégrer le modèle à une plateforme de clavardage ? Quel est son coût ?

Nous n'abordons pas ici les problématiques d'ordre commercial et de gestion de projet qui se situent en dehors du cadre de notre travail.

6.5.1 Implémenter un Bi-WMM2Seq dans une société

Il est important de rappeler que, contrairement à ce qui se passe dans un environnement de recherche scientifique, les entreprises commerciales n'ont souvent pas le temps de créer une architecture de *chatbot* et des modèles personnalisés. Elles se servent en général de solutions pré-construites comme celles que nous avons présentées en 2. Si nous prenons l'exemple de RASA, il s'agit d'une solution gratuite et en libre accès, avec une licence permissive, qui a atteint une certain notoriété.

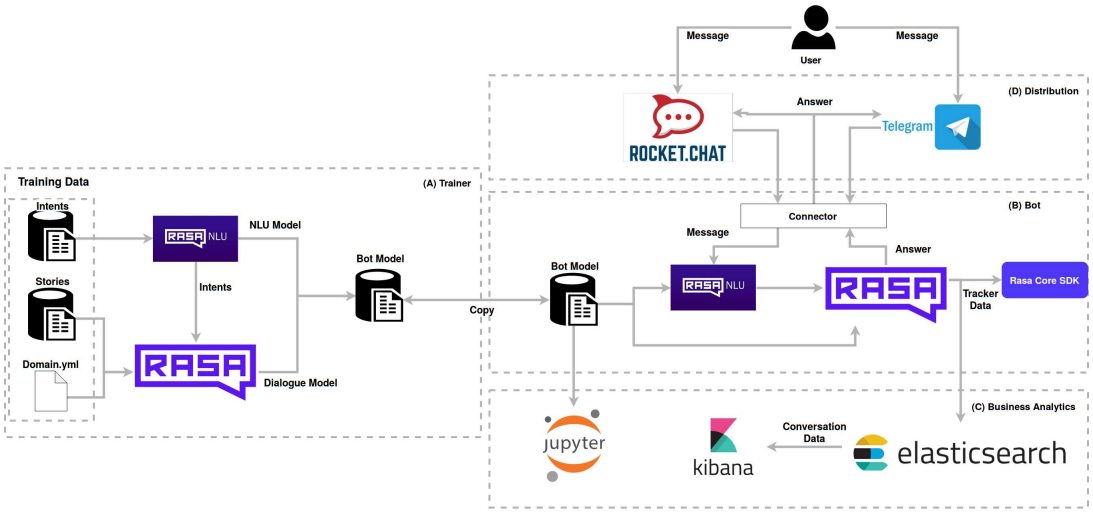

FIGURE 6.2 – Architecture de RASA (https ://github.com/lappis-unb/Bottis/blob/master/docs/README-en.md) d'après le LAPPIS de l'Universidade de Brasília

L'architecture permet soit de réaliser la tâche de prédiction d'actes de dialogue soit de pistage d'état. Il est également possible de lui ajouter un générateur de réponses en langage naturel et d'en faire un système de bout-en-bout (voir figure 6.2). Cependant, il n'est pas anodin de se lancer dans la conversion de l'architecture RASA pour qu'elle soit compatible avec notre modèle Bi-WMM2Seq [4]. Tout d'abord, cela signifie que l'on est dans l'obligation de créer deux instances de RASA, l'une pour le WMM2Seq global, l'autre pour le WMM2Seq interne. Ensuite, il faut créer un détecteur d'incohérences, si possible natif, donc inclus dans l'architecture RASA, pour ne pas complexifier l'environnement de travail. De plus, il faut y ajouter un classifieur pour prédire le nombre de tours de parole restants afin d'optimiser la réponse. Enfin, RASA travaille au niveau de la phrase, en générant une sortie qui est la réponse du système alors que l'architecture WMM2Seq travaille au niveau du mot. Cela implique une modification de l'architecture RASA au moment du décodage pour qu'il itère sur chaque mot de la réponse, au lieu de décoder la réponse d'un seul coup. Sachant qu'une seule instance de RASA demande un certain temps d'apprentissage pour l'optimisation [5], la conversion en Bi-WMM2Seq représente une quantité de temps conséquente et nécessite de disposer d'une puissance de calcul relativement importante.

4. https ://rasa.com/docs/rasa/custom-graph-components
5. https ://rasa.com/docs/rasa/tuning-your-model/

6.5.2 Le problème des données

L'autre problème en industrie concerne l'accessibilité et l'exploitation de données à la fois compatibles avec le système de dialogue et utiles à l'entreprise développant celui-ci. Dans la majorité des cas, les données ne sont pas disponibles et l'entreprise doit elle-même les créer, en inventant des possibles dialogues entre un client et un employé. En admettant que l'entreprise possède une expertise en systèmes de dialogue et d'annotations en état de croyance, en actes de dialogue... cela se traduit par un jeu de données réduit et peu ressemblant à la réalité de l'interaction utilisateur-*chatbot*. Dans le cas où l'entreprise dispose de données, comme c'est notre cas, il s'agit la plupart du temps de journaux de conversations entre un client final et un agent humain. Nous avons étudié dans ce chapitre les soucis posés par le RGPD et l'anonymisation de ces données. L'autre problème, au-delà du nettoyage et du formatage de données conversationnelles dans un format tel que DSTC2, bAbi ou MultiWOZ est celui de la *robotisabilité* des données.

Robotisabilité des données

Par ce terme barbare, nous entendons que certaines conversations ne sont pas exploitables par un système de dialogue. C'est une question déjà soulevée par [Kerr and Moloney, 2018, Dhoolia et al., 2021] mais cela dépasse le cadre de nos travaux. En somme cela signifie que, parmi les conversations entre un client et un agent disponibles, indépendamment du succès du dialogue de cette conversation, certaines sont considérées comme impossibles à faire apprendre à un système de dialogue. Que ce soit par leur taille, par leur difficulté (orthographe, nombre de sujets abordés par l'utilisateur, demande de conseils personnels) ou pour d'autres phénomènes comme le *troll* [Sheng et al., 2021] ou la faible capacité de communication de l'utilisateur, certaines conversations doivent être écartées. C'est ce genre de corpus qui nous a inspirés pour la détection des incohérences. En effet, DSTC2 présentait un certain type d'incohérences, surtout lié à des problèmes de compréhension du système ou des défauts dans la reconnaissance vocale. Cependant, dans des conversations humaines, le fait de trouver des incohérences également liées au comportement de l'utilisateur nous a permis d'établir une typologie plus riche et plus couvrante que celle uniquement basée sur les problèmes de DSTC2.

Un corpus ni avec compère ni humain-machine

Le résultat du filtrage de la robotisabilité d'un jeu de données résulte en un corpus de dialogue humain-humain, qui n'est ni du type compère (le client n'est pas convaincu de parler à un chatbot) ni du type humain-machine. De plus, contrairement aux corpus classiques disponibles dans la communauté, les conversations d'un tel corpus ne sont pas scriptées, il n'existe pas de but réellement prédéfini dans la conversation autre que celui de répondre au besoin du client dans le domaine

sélectionné. Cela rend chaque dialogue unique, avec une diversité linguistique plus importante que dans les corpus pré-existants. D'autre part, l'exercice d'annotation en actes de dialogue (intentions et entités notamment) demanderait une expertise forte à la fois dans les systèmes de dialogue et dans le domaine d'application.

Le problème du déploiement en temps réel

Le dernier problème que nous constatons est celui de la mise en pratique d'une telle architecture. En effet, dans le WMM2Seq, le temps de réponse est relativement rapide (moins d'une seconde entre la requête de l'utilisateur et la réponse générée). Cependant, dans notre architecture, comme nous y ajouterions un WMM2Seq interne et deux classifieurs, la latence entre la question de l'utilisateur et la réponse du système augmenterait. D'autre part, lors du déploiement, il faut également ajouter l'anonymiseur, puis la connexion à une plate-forme de messagerie instantanée et enfin la connexion à une base de données client, souvent sécurisée donc plus longue d'accès. Si l'on ajoute à cela les fois où le système fournit une réponse incohérente ou ne parvient pas à répondre à la requête de l'utilisateur, cela peut devenir long pour ce dernier. Sans compter que pour être rentable, l'entreprise doit pouvoir fournir un *chatbot* gérant plusieurs interactions simultanées, voire plusieurs instances de *chatbot* pour différents clients. Cela conduit à une surcharge des serveurs et des flux, et donc une nouvelle augmentation de la latence voire des possibles dysfonctionnements si les flux d'interrogation de la base de données sont saturés.

6.6 Travaux futurs

Malgré nos efforts, il reste une grande quantité de chantiers à explorer et à achever dans le domaine de la mémoire dans un système de dialogue orienté-tâche.

6.6.1 Données

Au niveau des données, nous n'avons exploité réellement qu'un seul corpus, DSTC2, qui bien qu'imparfait, est le seul corpus humain-machine avec annotation dans notre domaine. Une expérience à réaliser serait de faire apprendre au Bi-WMM2Seq le MultiWOZ, puis de lui faire générer un MultiWOZ humain-machine, avec des erreurs et des incohérences. Ce faisant, nous aurions une version du MultiWOZ bruitée par les défauts du modèle. Ainsi, nous pourrions apprendre au Bi-WMM2Seq à différencier les tours de parole corrects et ceux incohérents. Comme le MultiWOZ est un corpus plus imposant (trois fois plus de dialogues) et plus varié (plusieurs domaines au lieu d'un seul) que DSTC2, il est possible que la tâche pour le Bi-WMM2Seq soit encore plus difficile et que le résultat soit un système

plus robuste que celui entraîné sur DSTC2. Nous pourrions généraliser ce bruitage artificiel de corpus à DIASER.
D'autre part, il serait intéressant d'annoter le corpus industriel que nous avons anonymisé. C'est un corpus dans le domaine de la parfumerie entre des clients finaux et des employés, en clavardage instantané . Nous pourrions l'enrichir non seulement en annotations classiques (actes de dialogue, état de croyance, réussite de la tâche, but) mais également en termes de robotisabilité. Nous ne l'avons pas évoqué mais ce corpus présente aussi l'avantage de faire apparaître certains comportements de la part de l'utilisateur mais aussi de l'agent, comme l'exaspération, le désespoir, la reconnaissance, l'attrition... Autant d'éléments essentiels dans le monde de l'industrie mais peu présents dans les corpus orientés-tâche que nous exploitons. D'une façon générale, l'analyse de l'émotion dans le dialogue est une facette indispensable pour ce genre de corpus. En outre, il serait précieux de creuser le lien entre la rétention en mémoire des informations saillantes et les émotions ressenties. Il serait intéressant, grâce à ce corpus, de modéliser l'émotion chez l'utilisateur, mais aussi chez l'agent, cela rapprocherait encore plus le Bi-WMM2Seq des modèles cognitifs de la mémoire pendant le dialogue.

6.6.2 Architecture Bi-WMM2Seq

Plusieurs modules de notre architectures peuvent faire l'objet de travaux futurs. Pour commencer, le système encode l'historique de dialogue "tel quel" sans un filtrage ni mécanisme d'attention sur les informations les plus saillantes. Cette méthode montre ses limites dès que l'historique croit de manière importante ou que la quantité d'appels à la base de données est conséquent.

La gestion de l'historique de dialogue

Nous avons évoqué en 3.1 le fait que la mémoire de travail d'un humain est limitée en taille et en temps. Pendant un dialogue, plus celui-ci se prolonge, plus l'utilisateur filtre les informations qu'il a perçues pour ne conserver que celles essentielles à la fluidité conversationnelles. Lorsqu'il n'a plus de place dans sa mémoire de travail, il déclenche le sous-dialogue évoqué précédemment de confirmation afin de se rappeler les informations possiblement oubliées. Ainsi, la probabilité que les tours de parole les plus éloignées du tour de parole courant soient perdus augmente. En outre, l'utilisateur, devant également gérer les informations hypothétiques liées aux dimensions réflexive et temporelle du dialogue, il a de moins en moins de place en mémoire de travail pour continuer à stocker des perceptions au fur et à mesure que les tours de parole s'enchaînent. L'utilisateur opère alors un filtrage des informations et optimise la place restante dans sa mémoire de travail afin de pouvoir continuer le dialogue. Or, lors de l'apprentissage du modèle, le système a accès à l'intégralité de l'historique de dialogue, et ce quel que soit l'endroit où se situe le

tour de parole courant. [Gao et al., 2019] ont montré que plus l'historique de dialogue croît, et plus la probabilité que le système fournisse une réponse avec du bruit augmente, émettant l'hypothèse que les systèmes récents n'ont pas (encore) appris à optimiser le filtrage des informations lors de l'historique de dialogue. Certains travaux récents ont compris la nécessité de s'intéresser au filtrage de l'historique de dialogue afin d'améliorer la capacité de raisonnement du système. [Chen et al., 2020b] ont construit un modèle de pistage d'état de croyance avec une mémoire à taille figée, où le système ne peut raisonner que sur un nombre fixé à l'avance d'informations, limitant ainsi le nombre d'informations superflues et bruyantes que le système génère. [Liao et al., 2021] ont mis au point un système permettant au système de raisonner sur les informations du tour de parole courant, puis sur l'état de croyance précédent, au lieu de raisonner directement sur l'intégralité de l'historique de dialogue, ce qui a quelque peu augmenté la précision des informations apprises et générées par le système. Nous pourrions nous inspirer de ces deux approches pour filtrer l'historique de dialogue dans le Bi-WMM2Seq à chaque tour de parole, à la manière d'un mécanisme d'attention pour les tâches de résumé automatique [Zhang et al., 2018], de façon à n'encoder dans le contexte que les informations les plus saillantes. Cela rappelle les modèles cognitifs étudiés en 3.1 sur la mémoire sensorielle. Ce mécanisme limiterait le bruit généré par des historiques de dialogue trop longs et permettrait à la mémoire de travail de ne raisonner que sur une partie des informations présentes dans l'historique de travail.

La fonction de décision et d'oubli dans la mémoire de travail

Un autre chantier est celui de la fonction de décision définitive qui choisit à partir de quelle mémoire le mot suivant doit être tiré. Aujourd'hui il s'agit, dans notre architecture, de la fonction d'origine du WMM2Seq, à savoir une simple fonction de règle basée sur le pointeur le plus élevé entre les candidats des différentes mémoires. Il serait intéressant de réfléchir à une nouvelle fonction, plus basée sur les probabilités floues que sur une simple règle déterministe, afin de complexifier la mémoire de travail du Bi-WMM2Seq.
Par ailleurs, il serait intéressant d'expliciter une fonction d'oubli dans la mémoire de travail. Certes, dans la fonction $TRANS$ du WMM2seq (voir 4.2.3) dotée d'un bi-GRU présente dans la mémoire épisodique et dans le GRU de la mémoire de travail qui génère les pointeurs finaux, il existe une fonction d'oubli propre à l'architecture GRU. Il est vrai aussi qu'en appliquant un filtrage explicite sur l'historique de dialogue, on implémente une sorte de fonction d'oubli primaire. Cependant, nous pouvons nous demander si cela suffit à limiter le bruit et à ne garder dans la mémoire à court terme uniquement les informations utiles pour la suite du dialogue et pour mieux prédire l'état de croyance de l'utilisateur. Dans la mesure où cette fonction revêt une importance capitale dans les modèles cognitifs de mémoire de

travail, nous pensons qu'une piste future serait de réfléchir à comment insérer une fonction d'oubli par dessus le GRU final, afin d'optimiser la taille et le contenu du tampon de la mémoire de travail du Bi-WMM2Seq.

Quid des actes de dialogue et du pistage d'états ?

Le Bi-WMM2Seq est censé encoder une question de l'utilisateur avec un contexte et en déterminer un état de dialogue qui, combiné aux mémoires de décodage, lui permettent de générer une réponse. Cependant, nous ne possédons aucun moyen d'évaluer l'encodage de l'état. Il serait à notre avis indispensable pour notre modèle, mais aussi pour les réseaux de mémoire en général, de s'atteler à un apprentissage en deux phases, la compréhension du langage et le pistage de l'état du dialogue d'un côté, en exploitant les annotations sur le tour de parole, puis d'un autre côté la génération de la réponse. Sans doute pourrions-nous développer un double encodeur, l'un pour le contexte et les tours de parole, l'autre pour les actes de dialogue et l'état de croyance. Puis toutes ces informations seraient combinées dans une étape ultime de fusion.

6.6.3 Expérimentations et évaluations

Nous avons ajouté trois fonctionnalités, inspirées des modèles cognitifs de la mémoire, dans un WMM2Seq classique. Il serait intéressant à notre avis de creuser davantage chacune d'entre elles. Cependant, nous nous attardons davantage sur les deux fonctionnalités qui ont demandé le plus de travail, à savoir les incohérences et la temporalité.

Détection des incohérences

Pour le module de détection des incohérences, nous avons utilisé un modèle basé sur des règles afin d'annoter les tours de parole de DSTC2. Bien que ce modèle ait été évalué et vérifié par plusieurs linguistes, il reste non seulement perfectible mais aussi déterministe et spécialisé pour DSTC2. Il conviendrait de trouver un protocole d'annotation générique, agnostique des corpus, pour annoter des tours de parole en présence d'incohérence. D'autre part, nous avons utilisé une valeur booléenne pour le Bi-WMM2Seq (présence ou non d'incohérence). Cependant, nous possédons également les types des incohérences, il serait intéressant d'expérimenter une variante du détecteur qui renverrait un type d'incohérence plutôt qu'un booléen. Enfin, au lieu d'utiliser le résultat d'une classification binaire ou multi-classe, nous pourrions obtenir une estimation de la probabilité que ce tour de parole contienne une incohérence. Une valeur moins certaine pourrait permettre au Bi-WMM2Seq de raisonner différemment et d'obtenir de meilleures performances.

Prédiction du nombre de tours de parole restants

La prédiction de la taille restante du dialogue comme dimension temporelle reste un grand chantier, d'abord parce que les résultats de nos modèles ne sont pas assez bons (0.7 d'exactitude au mieux), ensuite parce que son apport dans le système reste très limité. Nous devrions songer à utiliser d'autres modèles que le Bi-LSTM. Cependant, des résultats avec BERT ont montré des scores similaires pour quelques combinaisons[6]. Peut-être le problème n'est pas de la classification mais bien de la régression. Le problème demeure dans le fait que si pour l'humain, il est presque impossible de déterminer s'il reste dix ou vingt tours de parole, nous ne voyons pas bien comment un modèle y parviendrait. Par ailleurs, l'information sur la présence d'une incohérence se montre déterminante pour améliorer cette prédiction. Si au lieu d'utiliser les booléens, nous avions utilisé les types, cela aurait-il pu augmenter les performances du modèle ?
Nous avons observé que le pic de croissance d'entropie se situe, en général, entre le deuxième et le troisième cinquième d'un dialogue. D'autre part, le Bi-LSTM se montre très précis pour déterminer s'il reste entre zéro et quatre tours de parole. Il serait intéressant d'ajouter une couche d'attention au modèle pour qu'il se concentre dès lors qu'il se trouve au delà de quatre tours de parole restants. [Li et al., 2019, Raheja and Tetreault, 2019] ont montré l'intérêt d'ajouter un mécanisme d'attention dans un réseau récurrent afin d'améliorer les résultats sur la tâche de classification d'actes de dialogue. [Hu et al., 2020] ont prouvé que pour la tâche de pistage d'état, la combinaison d'entités informatives et d'entités d'attention amélioraient la gestion des historiques de dialogue longs, un des problèmes encore présents dans le Bi-WMM2Seq. [Qiu et al., 2020] ont expérimenté avec succès l'utilisation de structures d'attention dans l'apprentissage non supervisé de dialogue, où la tâche consistait à regrouper les locutions de chaque participant (locuteur, écoutant) afin de modéliser des états latents du dialogue. Dans le domaine de la question-réponse, [Varanasi et al., 2020] ont utilisé un mécanisme de copie d'attention avec BERT, appelé CopyBERT, similaire au système de copie du WMM2Seq, pour améliorer la qualité des réponses. Bien que cela ne soit pas du dialogue, il nous faut rappeler que le modèle de réseau de mémoire originel est pensé pour la question réponse, ce n'est qu'ensuite qu'il a été adapté pour la tâche de dialogue (Mem2Seq). Il serait sans doute possible de transformer le CopyBERT pour la tâche de génération de réponse de dialogue, à l'instar de ReCoSa [Zhang et al., 2019a] et son système d'attention interne dans des encodeurs-décodeurs récurrents à hiérarchie.

6. résultats encore non vérifiés par validation croisée

6.6.4 Combiner GPT-2 et Bi-WMM2Seq

A la vue des résultats excellents de GPT-2 à la fois pour la détection des incohérences et pour la modélisation du dialogue, nous pensons qu'il serait intéressant de pouvoir combiner les réseaux de mémoires multiples avec les transformeurs. A notre connaissance, [Burtsev et al., 2020] sont les premiers à avoir proposé des tâches de traitement automatique des langues en combinant les deux architectures. Pour le dialogue, nous pourrions imaginer une architecture qui se serve à la fois de GPT-2 pour la compréhension de l'utilisateur et la gestion de l'historique de dialogue, ainsi que de la séparation des informations en différentes mémoires et le raisonnement sur plusieurs sauts de notre modèle Bi-WMM2Seq.

6.6.5 Un Bi-WMM2Seq en mode génération et déploiement ?

Enfin, le plus grand champ exploratoire restant à défricher nous paraît se trouver dans le déploiement et l'évaluation d'un Bi-WMM2Seq en situation réelle, c'est-à-dire où l'on prédit l'un après l'autre le tour de parole suivant, la présence d'une incohérence et le nombre de tours restants. Après plusieurs boucles entre les trois fonctionnalités, le système peut décider de générer la meilleure réponse possible, ou un message d'excuse si le détecteur d'incohérence en trouve une. Jusqu'à présent, nous avons implémenté un Bi-WMM2Seq en mode oracle, avec les informations supposées connues. Est-ce que le mode génération sera capable de se montrer aussi performant que le mode oracle et plus efficace qu'un WMM2Seq standard ?
De plus, il serait pertinent de réfléchir à une évaluation automatique d'un tel système. Comment vérifier que le système a raison d'envoyer un message d'erreur, ou qu'il a raison de corriger et modifier la réponse qu'il allait initialement produire ? Nous émettons l'hypothèse qu'un calcul de similarité cosinus à partir de vecteurs sémantiques de BERT entre l'énoncé de celui de l'utilisateur et celui du système pourrait fournir un score de pertinence de la réponse du système. D'autre part, il serait nécessaire de pénaliser un système qui génère trop souvent (plusieurs fois par dialogue) des messages d'excuse. Enfin, dans le cas où le Bi-WMM2Seq est déployé dans un *chatbot* de type RASA, il serait important d'évaluer à quel point la fonction de décision qui choisit soit de générer une réponse soit un message d'erreur est acceptable pour l'humain. Autrement dit, comment évaluer la patience d'un utilisateur face à un *chatbot* qui demande des reformulations ou abrège le dialogue fréquemment ? Comment évaluer quels types d'erreurs ou d'incohérences sont les plus propices à provoquer un rejet de l'utilisateur et donc de l'attrition ? Dans quelle mesure un analyseur d'émotions est nécessaire si le dessein du système est d'éviter le rejet de la part de l'utilisateur ?

6.7 Regrets

Nous avouons posséder également quelques regrets vis-à-vis de notre travail, certaines approches ou directions que nous aimerions avoir prises plus tôt ou mieux explorées. Certaines autres vers lesquelles nous aurions préféré de pas nous être dirigés.

Tout d'abord, nous regrettons de ne pas avoir découvert cette incroyable architecture neuronale qu'est le WMM2Seq plus tôt. Le modèle a été publié en 2019 et pourtant nous nous y sommes intéressés en 2020. Peut-être n'avions nous pas encore le recul nécessaire au moment de sa sortie pour comprendre sa pertinence et son utilité. Néanmoins, nous aurions eu un an de plus pour expérimenter et améliorer la solution actuelle.

Ensuite, nous n'avons pas été assez rapides dans l'implémentation de nos fonctionnalités. Les résultats montrent certes un progrès par rapport au modèle de référence. Cependant, ces progrès doivent être relativisés dans la mesure où le corpus de référence n'est pas sans défauts et donc ne possède pas une fiabilité à toute épreuve. Avec plus de temps pour optimiser les trois fonctionnalités, et l'évaluation par similarité sémantique, nos travaux auraient gagné en impact.

Puis, nous pouvons regretter de ne pas avoir exploré la temporalité dans les modèles de mémoire dans le dialogue que nous avons étudiés. En effet, nous nous sommes concentrés sur l'aspect dual du dialogue et de la représentation de soi vu à travers l'image que l'on possède de l'autre. Nous avons quelque peu délaissé la partie temporelle, en particulier le lien entre le passé (historique de dialogue), le présent (incohérence) et l'anticipation (tour de parole suivant et nombre de tours restants). Non seulement cela nous aurait permis de présenter un modèle de prédiction des tours restants plus robuste, mais aussi d'explorer la piste de l'anticipation qui influence le présent mais aussi l'interprétation du passé (mise à jour rétroactive de l'état de croyance).

Sûrement lié au premier point, nous ne nous sommes pas assez concentrés sur les transformeurs, cette architecture qui a été une révolution pour la modélisation du langage depuis 2018. Nous avons déjà cité la littérature qui utilise les transformeurs dans plusieurs tâches de dialogue. RASA a déployé une architecture basée sur deux transformeurs, l'un pour la compréhension de l'énoncé de l'utilisateur, et l'autre pour le pistage d'états. Nous avons la sensation d'avoir eu à choisir entre les réseaux de mémoire et les transformeurs, sans avoir trouvé le temps de comparer ou d'associer les deux.

Ce temps que nous n'avons pas consacré aux transformeurs, peut-être l'avons-nous utilisé pour créer et publier le corpus DIASER, qui nous semble important pour la communauté de dialogue. Cependant, le temps passé sur ce corpus (dix mois entre le début du projet et la publication de l'article) n'a pas pu être dédié à découvrir une autre architecture plus évoluée que le Bi-WMM2Seq et à optimiser les fonctionnalités. Toutefois, sans DIASER nous n'aurions pas découvert l'importance du

calcul de l'entropie pour comprendre les variations dans l'information au long d'un dialogue et les disparités entre les dialogue humains et humain-machine.
Enfin, cela est notre plus grand regret, nous ne sommes pas parvenus à effectuer le transfert entre la science et l'industrie. Nous aurions voulu intégrer pleinement le bi-wmm2seq aux solutions de l'entreprise, afin de pouvoir montrer son apport comparé à une boîte à outil en libre-accès et laisser une contribution pérenne. Nous aurions pu ainsi évaluer notre modèle avec des vrais usagers et vérifier nos hypothèses quant à l'intérêt de nos trois fonctionnalités et de notre évaluation en similarité sémantique.
Dans le même ordre d'idées, nous regrettons de ne pas avoir eu le temps de finir et de publier le corpus de journal de clavardage dans le domaine de la parfumerie, annoté dans le style MultiWOZ avec des indications sur la robotisabilité et des annotations en émotion. Ce corpus aurait pu être exploité et ajouté à DIASER.

Conclusion

And now, the conclusion

Teal'c

En définitive, la problématique principale de nos travaux était celle de la différence de gestion de la mémoire entre l'humain et la machine durant l'interaction dialogique écrite. Nous avons exploré plusieurs pistes, inspiré par la littérature et certaines études sur les modèles de mémoire cognitive, qui pourraient expliquer cette différence. Pour y parvenir, nous avons dans un premier temps proposé une typologie des systèmes de dialogue et un résumé de l'état de l'art à la fois pour les modèles de bout-en-bout et pour les modèles spécialisés dans les trois principales tâches (compréhension du langage, pistage de l'état du dialogue, génération des énoncés), ce faisant, nous avons découvert les réseaux de mémoire.
Ceux-ci, inspirés des recherches en science cognitive sur le fonctionnement des mémoires humaines et de leur utilité lors d'un dialogue, apportent un élément de réponse pour la simulation l'interaction entre les différents types de mémoires dans optimisation la compréhension de l'utilisateur et la qualité de la réponse de la part du système. En effet, dans la mesure où l'humain possède un système mnémonique composé de mémoires allant du très court terme (sensorielle), au très long terme (procédurale), en passant par les mémoires déclaratives à long terme (épisodique et sémantique) qui interagissent avec la mémoire à court terme (de travail), cette dernière servant à la fois de stockage et de processeur (fonction d'oubli, fonction de décision), il a été possible d'imiter ce système, grâce aux réseaux de neurones, dans un modèle appelé WMM2Seq
Ces modèles nous ont aidé à exhiber les différences fondamentales entre une interaction humaine et humain-machine. La première fonctionnant d'après le modèle COSMO comme un miroir, où chaque participant se représente non seulement une image de l'autre mais aussi une image de la représentation que se fait l'autre de soi. C'est ce qui rend les dialogues entre les personnes fluides et naturels.
Nous avons émis l'hypothèse qu'un des éléments pouvant expliquer les performances et surtout les limites des systèmes de dialogue orientés-tâche actuels réside dans l'absence de modélisation de certaines caractéristiques cognitives. Ces limites se manifestent en particulier par des incohérences dans le dialogue et les énoncés.

Nous avons alors exploité un jeu de données, appelé DSTC2, représentant un corpus de dialogues humain-machine, contenant des incohérences.
Pour prouver cette hypothèse, nous sommes partis de l'architecture qui s'accordait le plus avec notre problématique. Cette architecture, appelée WMM2Seq, intègre dans son fonctionnement des principes issus des neurosciences, de la psychologie et de la sémantique comme : la séparation des informations biographiques, la gestion des perceptions à court terme et le principe de décision de la prochaine action conditionné par le contenu des différents types de mémoires.
Nous avons développé une fonctionnalité supplémentaire au WMM2Seq : la détection des incohérences afin de faire apprendre au système quand il se trompait. Les meilleurs résultats obtenus l'ont été avec le transformeur GPT-2 , suivi du WMM2Seq.
Afin d'émuler la réflexivité du dialogue humain, nous avons incorporé un WMM2Seq capable de prédire le tour de parole suivant de l'utilisateur. Nous avons prouvé que cela améliorait la détection des incohérences du système, notre modèle est nommé Bi-WMM2Seq.
Cela nous a permis de comprendre que l'autre élément manquant à la machine était une anticipation de la prochaine étape du dialogue, comme le ferait un humain : nous avons explicité dans le Bi-WMM2Seq une fonction de prédiction du nombre de tours de parole restants, afin que le système sache non seulement depuis quand le dialogue a commencé, mais aussi ait une indication sur le moment où il devrait se finir.
Toutes ces découvertes et l'implémentation de ces fonctionnalités ont été permises grâce à la création du corpus DIASER à laquelle nous avons apporté une contribution conséquente dans le cadre d'une collaboration internationale avec des collègues de l'Université de Prague. Ce corpus combine et unifie différents dialogues annotés disponibles librement afin de créer un jeu de données plus riche. Il nous a permis de faire des mesures de l'entropie des tours de parole de dialogues issus de plusieurs domaines . Nous avons également montré que pour certaines tâches, l'utilisation du corpus DIASER rendait un modèle plus robuste, montrant les similitudes et la complémentarité des différentes sources qui le constituent.
Les différentes mesures d'entropie et d'analyse des erreurs ont montré des similitudes pour ce qui concerne le pic où se produisent les erreurs et où l'entropie est la plus élevée. L'exploitation de ces informations permet de donner au système de dialogue un ancrage temporel plus fort.
Ainsi, nous avons effectué des expérimentations afin de prédire la temporalité dans un corpus. Les meilleurs résultats ont été obtenus sur le corpus MultiWOZ où l'ajout du tour de parole suivant de l'utilisateur ainsi que le découpage du dialogue en tranches temporelles, déduites grâce aux mesures d'entropie, se sont révélés cruciaux. Le modèle était en général capable de prédire précisément le nombre de tours de parole restants lorsqu'il en reste moins de cinq.
En ayant ainsi une idée non seulement du début du dialogue mais aussi de sa fin,

et en étant aussi capable de détecter des incohérences et de prédire le tour de parole suivant, nous avons exploré la capacité du modèle Bi-WMM2Seq que nous proposons, à corriger et fournir des réponses plus pertinentes.
Les résultats montrent que les trois fonctionnalités, en mode oracle, fournissent un apport suffisant au modèle pour surpasser le WMM2Seq de référence, sur le nouveau corpus DSTC2.1, une version enrichie du DSTC2, où nous avons ajouté les informations sur la présence ou non d'une incohérence et le nombre de tours de parole restants à un tour de parole donné. Nous avons montré que les trois fonctionnalités sont utiles, la combinaison des trois offrant les meilleurs résultats en termes de f-mesure des entités (+ 0.07).
Cependant, lorsque nous avons comparé notre modèle avec GPT-2, sur les mêmes données de DSTC2.1, le transformeur s'est montré plus performant pour cette même mesure (+ 0.20). Sa performance dépasse même celle du modèle entraîné sur DSTC2 (+0.17). La combinaison de la séparation des informations en mémoire avec des sauts de réflexion comme l'effectue un WMM2Seq, ainsi que la modélisation du langage à la manière de GPT-2 pourrait-elle devenir une nouvelle référence pour la tâche de génération de réponse ?
Dans tous les cas, ces travaux ont permis de prouver que le rapprochement vers des modèles de mémoires inspirés des modèles cognitifs, pour la modélisation du dialogue améliore les résultats lorsqu'on fournit au système des fonctionnalités lui permettant de détecter ses propres incohérences, et d'avoir un meilleur ancrage temporel sur la progression du dialogue.

Bibliographie

[Acedo et al., 2019] Acedo, J., Fernández-Sellers, M., and Lozano-Tello, A. (2019). Detection model of similar datasets. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.

[Acedo et al., 2018] Acedo, J., Lozano-Tello, A., and Fernandez-Sellers, M. (2018). Model of datasets unification from different open data portals. In *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.

[Adams et al., 2019] Adams, A., Aili, E., Aioanei, D., Jonsson, R., Mickelsson, L., Mikmekova, D., Roberts, F., Valencia, J. F., and Wechsler, R. (2019). AnonyMate : A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, Turku, Finland. Linköping Electronic Press.

[Allwood et al., 1992] Allwood, J., Nivre, J., and Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1) :1–26.

[ALLWOOD et al., 2000] ALLWOOD, J., TRAUM, D., and JOKINEN, K. (2000). Cooperation, dialogue and ethics. *International Journal of Human-Computer Studies*, 53(6) :871–914.

[André, 2010] André, V. (2010). Analyse pragmalinguistique de la co-construction du discours : le cas des énonciations conjointes et des reprises. working paper or preprint.

[Angiuli and Waldo, 2016] Angiuli, O. and Waldo, J. (2016). Statistical tradeoffs between generalization and suppression in the de-identification of large-scale data sets. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 589–593.

[Aslot et al., 2001] Aslot, V., Domeika, M., Eigenmann, R., Gaertner, G., Jones, W. B., and Parady, B. (2001). Specomp : A new benchmark suite for measuring parallel computer performance. In Eigenmann, R. and Voss, M. J., editors, *OpenMP Shared Memory Parallel Programming*, pages 1–10, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Atkinson and Shiffrin, 1968] Atkinson, R. and Shiffrin, R. (1968). Human memory. volume 2 of *Psychology of Learning and Motivation*, pages 89 – 195. Academic Press.

[Auguste et al., 2019] Auguste, J., Béchet, F., Damnati, G., and Charlet, D. (2019). Skip Act Vectors : integrating dialogue context into sentence embeddings. In *Tenth International Workshop on Spoken Dialogue Systems Technology*, Syracuse, Italy.

[Baddeley, 1996] Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology Section A*, 49(1) :5–28.

[Baddeley, 2010] Baddeley, A. (2010). Working memory. *Current Biology*, 20(4) :R136 – R140.

[Baddeley et al., 1998] Baddeley, A., Gathercole, S., and Papagno, C. (1998). The phonological loop as a language learning device. *Psychological review*, 105 :158–73.

[Baddeley, 1995] Baddeley, A. D. (1995). *The psychology of memory*, chapter The psychology of memory., pages 3–25. John Wiley and Sons, Oxford, England.

[Baddeley and Dale, 1966] Baddeley, A. D. and Dale, Harold, C. (1966). The effect of semantic similarity on retroactive interference in long-and short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 5(5) :417–420.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date : 07-05-2015 Through 09-05-2015.

[Bailey et al., 2008] Bailey, H., Dunlosky, J., and Kane, M. J. (2008). Why does working memory span predict complex cognition ? testing the strategy affordance hypothesis. *Memory & Cognition*, 36(8) :1383–1390.

[Bailly et al., 2008] Bailly, G., Elisei, F., and Raidt, S. (2008). Boucles de perception-action et interaction face-à-face. *Revue française de linguistique appliquée*, 13(2) :121–131.

[Balaraman et al., 2021] Balaraman, V., Sheikhalishahi, S., and Magnini, B. (2021). Recent neural methods on dialogue state tracking for task-oriented dialogue systems : A survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 239–251, Singapore and Online. Association for Computational Linguistics.

[Bangerter, 2004] Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological science*, 15 6 :415–9.

[Barnaud et al., 2018] Barnaud, M.-L., Bessière, P., Diard, J., and Schwartz, J.-L. (2018). Reanalyzing neurocognitive data on the role of the motor system in speech perception within cosmo, a bayesian perceptuo-motor model of speech communication. *Brain and language*, 187 :19–32.

[Becerra et al., 2018] Becerra, A., De La Rosa, J. I., and González, E. (2018). Speech recognition in a dialog system : from conventional to deep processing. *Multimedia Tools and Applications*, 77(12) :15875–15911.

[Bennacef et al., 1996] Bennacef, S., Devillers, L., Rosset, S., and Lamel, L. (1996). Dialog in the railtel telephone-based system. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 1, pages 550–553 vol.1.

[Bentz et al., 2017] Bentz, C., Alikaniotis, D., Cysouw, M., and Ferrer-i Cancho, R. (2017). The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, 19(6).

[Berg and Dalianis, 2019] Berg, H. and Dalianis, H. (2019). Augmenting a de-identification system for Swedish clinical text using open resources and deep learning. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 8–15, Turku, Finland. Linköping Electronic Press.

[Bernsen et al., 1996] Bernsen, N. O., Dybkjær, H., and Dybkjær, L. (1996). Co-operativity in humanmachine and humanhuman spoken dialogue. *Discourse Processes*, 21(2) :213–236.

[Bertin and Masson, 2020] Bertin, T. and Masson, C. (2020). Rôle du dialogue et de la co-construction du discours dans lacquisition de la morphosyntaxe : un processus interactionnel et dialogique. *Bakhtiniana : Revista de Estudos do Discurso*, 16 :88–113.

[Bibauw et al., 2015] Bibauw, S., François, T., and Desmet, P. (2015). Dialogue-based call : an overview of existing research. In *Critical CALL–Proceedings of the 2015 EUROCALL Conference, Padova, Italy*, pages 57–64. Research-publishing. net Dublin.

[Blanchet, 2015] Blanchet, P. (2015). *Guide pour la recherche en didactique des langues et des cultures : approches contextualisées*. Archives contemporaines.

[Blum-Kulka et al., 1989] Blum-Kulka, S., House, J., and Kasper, G. (1989). Investigating cross-cultural pragmatics : An introductory overview. *Cross-cultural pragmatics : Requests and apologies*, 31 :1–34.

[Bobrow et al., 1977] Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. (1977). Gus, a frame-driven dialog system. *Artificial Intelligence*, 8(2) :155 – 173.

[Bocklisch et al., 2017a] Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. (2017a). Rasa : Open source language understanding and dialogue management. *arXiv preprint arXiv :1712.05181*.

[Bocklisch et al., 2017b] Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. (2017b). Rasa : Open source language understanding and dialogue management. *CoRR*, abs/1712.05181.

[Bodner and Lindsay, 2003] Bodner, G. E. and Lindsay, D. S. (2003). Remembering and knowing in context. *Journal of Memory and Language*, 48(3) :563 – 580.

[Boland, 2019] Boland, J. (2019). Conversation transition times : Working memory & conversational alignment. In *CogSci*, volume 19, pages 159–165.

[Bonchi et al., 2011] Bonchi, F., Castillo, C., Gionis, A., and Jaimes, A. (2011). Social network analysis and mining for business applications. *ACM Trans. Intell. Syst. Technol.*, 2(3) :22 :1–22 :37.

[Bordes et al., 2017] Bordes, A., Boureau, Y.-L., and Weston, J. (2017). Learning end-to-end goal-oriented dialog.

[Bostan and Klinger, 2018] Bostan, L. A. M. and Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119. Association for Computational Linguistics.

[Bottis and Bouchagiar, 2018] Bottis, M. and Bouchagiar, G. (2018). Personal data v. big data in the eu : Control lost, discrimination found. *Open Journal of Philosophy*, 08 :192–205.

[Braun et al., 2017] Braun, D., Hernandez Mendez, A., Matthes, F., and Langen, M. (2017). Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany. Association for Computational Linguistics.

[Brod et al., 2016] Brod, G., Lindenberger, U., Wagner, A. D., and Shing, Y. L. (2016). Knowledge acquisition during exam preparation improves memory and modulates memory formation. *Journal of Neuroscience*, 36(31) :8103–8111.

[Budzianowski et al., 2018a] Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018a). MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

[Budzianowski et al., 2018b] Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018b). Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv :1810.00278*.

[Bunk et al., 2020] Bunk, T., Varshneya, D., Vlasov, V., and Nichol, A. (2020). DIET : lightweight language understanding for dialogue systems. *CoRR*, abs/2004.09936.

[Bunt et al., 2010] Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Chengyu Fang, A., Hasida, K. . ., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., and Traum, D. (2010). Towards an ISO Standard for Dialogue Act Annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*, La Valette, Malta.

[Bunt et al., 2020] Bunt, H., Petukhova, V., Gilmartin, E., Pelachaud, C., Fang, A., Keizer, S., and Prévot, L. (2020). The ISO Standard for Dialogue Act Annotation, Second Edition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

[Burtsev et al., 2018] Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., Litinsky, A., Logacheva, V., Lymar, A., Malykh, V., Petrov, M., Polulyakh, V., Pugachev, L., Sorokin, A., Vikhreva, M., and Zaynutdinov, M. (2018). Deeppavlov : Open-source library for dialogue systems.

[Burtsev et al., 2020] Burtsev, M. S., Kuratov, Y., Peganov, A., and Sapunov, G. V. (2020). Memory transformer. *arXiv preprint arXiv :2006.11527*.

[Byrne et al., 2019] Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Duckworth, D., Yavuz, S., Goodrich, B., Dubey, A., Kim, K.-Y., and Cedilnik, A. (2019). Taskmaster-1 : Toward a realistic and diverse dialog dataset.

[Cao et al., 2020] Cao, W., Mirjalili, V., and Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140 :325–331.

[Cassell et al., 1994] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation : rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420.

[Cattan, 2020] Cattan, O. (2020). L'adaptabilité comme compétence pour les systèmes de dialogue orientés tâche (adaptability as a skill for goal-oriented dialog systems). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL*, pages 85–95, Nancy, France. ATALA et AFCP.

[Chen et al., 2017] Chen, H., Liu, X., Yin, D., and Tang, J. (2017). A survey on dialogue systems : Recent advances and new frontiers. *CoRR*, abs/1711.01731.

[Chen et al., 2018] Chen, H., Ren, Z., Tang, J., Zhao, Y. E., and Yin, D. (2018). Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1653-1662, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Chen et al., 2020a] Chen, J., Agbodike, O., and Wang, L. (2020a). Memory-based deep neural attention (mdna) for cognitive multi-turn response retrieval in task-oriented chatbots. *Applied Sciences*, 10(17) :5819.

[Chen et al., 2009] Chen, J., Huang, H., Tian, S., and Qu, Y. (2009). Feature selection for text classification with naïve bayes. *Expert Systems with Applications*, 36(3) :5432–5435.

[Chen et al., 2020b] Chen, L., Lv, B., Wang, C., Zhu, S., Tan, B., and Yu, K. (2020b). Schema-guided multi-domain dialogue state tracking with graph attention neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05) :7521–7528.

[Chen and Kan, 2013] Chen, T. and Kan, M.-Y. (2013). Creating a live, public short message service corpus : the nus sms corpus. *Language Resources and Evaluation*, 47(2) :299–335.

[Chen et al., 2019a] Chen, W., Chen, J., Qin, P., Yan, X., and Wang, W. Y. (2019a). Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709, Florence, Italy. Association for Computational Linguistics.

[Chen et al., 2019b] Chen, X., Xu, J., and Xu, B. (2019b). A working memory model for task-oriented dialog response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2687–2693, Florence, Italy. Association for Computational Linguistics.

[Chen et al., 2019c] Chen, X., Xu, J., and Xu, B. (2019c). A Working Memory Model for Task-oriented Dialog Response Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2687–2693, Florence, Italy. Association for Computational Linguistics.

[Chernyshova, 2018] Chernyshova, E. (2018). *Expliciter et inférer dans la conversation : modélisation de la séquence d'explicitation dans l'interaction*. Theses, Université de Lyon.

[Chevrier et al., 2019] Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., and Lovis, C. (2019). Use and understanding of anonymization and de-identification in the biomedical literature : Scoping review. *J Med Internet Res*, 21(5) :e13484.

[Chomsky, 1965] Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

[Choromanski et al., 2020] Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al. (2020). Rethinking attention with performers. *arXiv preprint arXiv :2009.14794*.

[Ciaramelli et al., 2013] Ciaramelli, E., Bernardi, F., and Moscovitch, M. (2013). Individualized theory of mind (itom) : When memory modulates empathy. *Frontiers in Psychology*, 4 :4.

[Cintrón-Valentín and Ellis, 2016] Cintrón-Valentín, M. C. and Ellis, N. C. (2016). Salience in second language acquisition : Physical form, learner attention, and instructional focus. *Front Psychol*, 7 :1284–1284. 27621715[pmid].

[Clark and Marshall, 1981] Clark, H. H. and Marshall, C. R. (1981). Definite knowledge and mutual knowledge. In Joshi, A. K., Webber, B. L., and Sag, I. A., editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge, UK : Cambridge University Press.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1) :37–46.

[Colby, 1974] Colby, K. M. (1974). Ten criticisms of parry. *ACM SIGART Bulletin*, (48) :5–9.

[Coucke et al., 2018] Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., Primet, M., and Dureau, J. (2018). Snips voice platform : an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

[Cowan, 2008] Cowan, N. (2008). What are the differences between long-term, short-term, and working memory ? *Progress in brain research*, 169 :323–338.

[Creissels, 1995] Creissels, D. (DL1995). *Éléments de syntaxe générale / Denis Creissels*. Linguistique nouvelle. Presses universitaires de France, Paris.

[Croce et al., 2020] Croce, D., Castellucci, G., and Basili, R. (2020). GAN-BERT : Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.

[Dalianis, 2019] Dalianis, H. (2019). Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

[Denny et al., 2014] Denny, C. A., Kheirbek, M. A., Alba, E. L., Tanaka, K. F., Brachman, R. A., Laughman, K. B., Tomm, N. K., Turi, G. F., Losonczy, A., and Hen, R. (2014). Hippocampal memory traces are differentially modulated by experience, time, and adult neurogenesis. *Neuron*, 83(1) :189–201. 24991962[pmid].

[Deppermann, 2018] Deppermann, A. (2018). Inferential practices in social interaction : A conversation-analytic account. *Open Linguistics*, 4(1) :35 – 55.

[D'Esposito, 2007] D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 362(1481) :761–772.

[DeVault et al., 2011] DeVault, D., Sagae, K., and Traum, D. (2011). Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, 2(1) :143–170.

[Devillers et al., 2004a] Devillers, L., Bonneau-Maynard, H., Rosset, S., Paroubek, P., McTait, K., Mostefa, D., Choukri, K., Charnay, L., Bousquet-Vernhettes, C., Vigouroux, N., Béchet, F., Romary, L., Antoine, J.-Y., Villaneau, J., Vergnes, M., and Goulian, J. (2004a). The french media/evalda project : the evaluation of the understanding capability of spoken language dialogue systems. In *LREC*.

[Devillers et al., 2003] Devillers, L., Maynard, H., Paroubek, P., and Rosset, S. (2003). The peace slds understanding evaluation paradigm of the french media campaign. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing : are evaluation methods, metrics and resources reusable ?*, Budapest, Hungary.

[Devillers et al., 2004b] Devillers, L., Maynard, H., Rosset, S., Paroubek, P., McTait, K., Mostefa, D., Choukri, K., Charnay, L., Bousquet, C., Vigouroux, N., et al. (2004b). The french media/evalda project : the evaluation of the understanding capability of spoken language dialogue systems. In *LREC*. Citeseer.

[Devillers et al., 2002] Devillers, L., Rosset, S., Bonneau-Maynard, H., and Lamel, L. (2002). Annotations for dynamic diagnosis of the dialog state. In *LREC*. Citeseer.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Dhoolia et al., 2021] Dhoolia, P., Kumar, V., Contractor, D., and Joshi, S. (2021). Bootstrapping dialog models from human to human conversation logs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 16024–16025.

[Di Cerbo and Trabelsi, 2018] Di Cerbo, F. and Trabelsi, S. (2018). Towards personal data identification and anonymization using machine learning techniques. In Benczúr, A., Thalheim, B., Horváth, T., Chiusano, S., Cerquitelli, T., Sidló, C., and Revesz, P. Z., editors, *New Trends in Databases and Information Systems*, pages 118–126, Cham. Springer International Publishing.

[Dong and Wyer, 2014] Dong, P. and Wyer, R. S. (2014). How time flies : The effects of conversation characteristics and partner attractiveness on duration judgments in a social interaction. *Journal of Experimental Social Psychology*, 50 :1–14.

[Dušek et al., 2020] Dušek, O., Novikova, J., and Rieser, V. (2020). Evaluating the State-of-the-Art of End-to-End Natural Language Generation : The E2E NLG Challenge. *Computer Speech & Language*, 59 :123–156.

[Dzikovska et al., 2011] Dzikovska, M. O., Moore, J. D., Steinhauser, N., and Campbell, G. (2011). Exploring user satisfaction in a tutorial dialogue system.

In *Proceedings of the SIGDIAL 2011 Conference*, pages 162–172, Portland, Oregon. Association for Computational Linguistics.

[Dziri et al., 2019] Dziri, N., Kamalloo, E., Mathewson, K., and Zaiane, O. (2019). Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.

[E. Bratman, 1987] E. Bratman, M. (1987). Intention, plans and practical reason. *Bibliovault OAI Repository, the University of Chicago Press*, 100.

[Ebbinghaus, 2013] Ebbinghaus, H. (2013). Memory : a contribution to experimental psychology. *Ann Neurosci*, 20(4) :155–156. 25206041[pmid].

[El Asri et al., 2017] El Asri, L., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., and Suleman, K. (2017). Frames : a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

[Ellenbogen et al., 2006] Ellenbogen, J. M., Payne, J. D., and Stickgold, R. (2006). The role of sleep in declarative memory consolidation : passive, permissive, active or none ? *Current opinion in neurobiology*, 16(6) :716–722.

[Engelbrecht and Möller, 2010] Engelbrecht, K.-P. and Möller, S. (2010). Sequential classifiers for the prediction of user judgments about spoken dialog systems. *Speech Communication*, 52(10) :816 – 833.

[Engle, 2002] Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in psychological science*, 11(1) :19–23.

[Eric et al., 2019] Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., and Hakkani-Tur, D. (2019). Multiwoz 2.1 : Multi-domain dialogue state corrections and state tracking baselines. *arXiv preprint arXiv :1907.01669*.

[Favre et al., 2017] Favre, B., Bechet, F., Damnati, G., and Charlet, D. (2017). Apprentissage d'agents conversationnels pour la gestion de relations clients (training chatbots for customer relation management). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 3 - Démonstrations*, pages 17–18, Orléans, France. ATALA.

[Fernandez et al., 2006] Fernandez, R., Lucht, T., Rodriguez, K., and Schlangen, D. (2006). Interaction in task-oriented human-human dialogue : the effects of different turn-taking policies. In *2006 IEEE Spoken Language Technology Workshop*, pages 206–209.

[Fernández-Sellers et al., 2019] Fernández-Sellers, M., Acedo, J., and Lozano-Tello, A. (2019). Identification of representative terms of datasets. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.

[Filisko and Seneff, 2004] Filisko, E. and Seneff, S. (2004). Error detection and recovery in spoken dialogue systems. In *Proceedings of the HLT-NAACL 2004 Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing*, pages 31–38, Boston, Massachusetts, USA. Association for Computational Linguistics.

[Fisher, 1936] Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, 1(3923) :554–554. PMC2458144[pmcid].

[Forbes-Riley and Litman, 2006] Forbes-Riley, K. and Litman, D. (2006). Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 264–271, New York City, USA. Association for Computational Linguistics.

[Fortuna et al., 2018] Fortuna, P., Ferreira, J., Pires, L., Routar, G., and Nunes, S. (2018). Merging datasets for aggressive text identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 128–139, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[Francopoulo, 2008] Francopoulo, G. (2008). Tagparser : well on the way to ISO-TC37 conformance. In *ICGL (International Conference on Global Interoperabilty for Language Resources)*, Hong Kong.

[Francopoulo and Schaub, 2020] Francopoulo, G. and Schaub, L.-P. (2020). Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. In *workshop on Legal and Ethical Issues (Legal2020)*, pages 9–14, Marseille, France. LREC2020, ELRA.

[Fraser et al., 2019] Fraser, K. C., Linz, N., Lindsay, H., and König, A. (2019). The importance of sharing patient-generated clinical speech and language data. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 55–61, Minneapolis, Minnesota. Association for Computational Linguistics.

[Fryer and Carpenter, 2006] Fryer, L. and Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, 10(3) :8–14.

[Fuster, 2004] Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends in cognitive sciences*, 8(4) :143–145.

[Galibert et al., 2005] Galibert, O., Illouz, G., and Rosset, S. (2005). Ritel : un système de dialogue homme-machine à domaine ouvert. In *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 439–444, Dourdan, France. ATALA.

[Gao et al., 2018] Gao, J., Galley, M., and Li, L. (2018). Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374.

[Gao and Zhang, 2005] Gao, J. and Zhang, J. (2005). Clustered svd strategies in latent semantic indexing. *Information Processing & Management*, 41(5) :1051–1063.

[Gao et al., 2019] Gao, S., Sethi, A., Agarwal, S., Chung, T., and Hakkani-Tur, D. (2019). Dialog state tracking : A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–273, Stockholm, Sweden. Association for Computational Linguistics.

[García, 1993] García, C. (1993). Making a request and responding to it : A case study of peruvian spanish speakers. *Journal of Pragmatics*, 19(2) :127–152.

[Gasic et al., 2010] Gasic, M., Jurcicek, F., Keizer, S., Mairesse, F., Thomson, B., Yu, K., and Young, S. (2010). Gaussian processes for fast policy optimisation of pomdp-based dialogue managers. In *Proceedings of the SIGDIAL 2010 Conference*, pages 201–204.

[Gatt and Krahmer, 2018] Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation : Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61 :65–170.

[Gauvain et al., 1997] Gauvain, J.-L., Bennacef, S., Devillers, L., Lamel, L., and Rosset, S. (1997). The spoken language component of the mask kiosk.

[Gianola et al., 2020] Gianola, L., riks Ajausks, Arranz, V., Bendahman, C., Bié, L., Borg, C., Cerdà, A., Choukri, K., Cuadros, M., de Gibert, O., Degroote, H., Edelman, E., Etchegoyhen, T., Ángela Franco Torres, Gatt, M. G. H. A. G. P. A., Grouin, C., Herranz, M., Kohan, A. A., Lavergne, T., Melero, M., Paroubek, P., Rigault, M., Rosner, M., Rozis, R., van der Plas, L., Vksna, R., and Zweigenbaum, P. (2020). *Legal Knowledge and Information Systems*, chapter Automatic Removal of Identifying Information in Official EU Languages for Public Administrations : The MAPA Project, pages 223–226. IOS Press, s. villata et al. (eds.) edition. doi :10.3233/FAIA200869.

[Gleason, 2005] Gleason, J. B. (2005). The development of language, 6/e.

[Goddard, 2017] Goddard, M. (2017). The eu general data protection regulation (gdpr) : European regulation that has a global impact. *International Journal of Market Research*, 59(6) :703–705.

[Godfrey et al., 1992] Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard : Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

[Goel et al., 2018] Goel, R., Paul, S., Chung, T., Lecomte, J., Mandal, A., and Hakkani-Tur, D. (2018). Flexible and scalable state tracking framework for goal-oriented dialogue systems. *arXiv preprint arXiv :1811.12891*.

[Goel et al., 2019] Goel, R., Paul, S., and Hakkani-Tür, D. (2019). Hyst : A hybrid approach for flexible and accurate dialogue state tracking. *arXiv preprint arXiv :1907.00883*.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

[Goodfellow et al., 2020] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11) :139–144.

[Gordon et al., 2001] Gordon, P. C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *Journal of experimental psychology : learning, memory, and cognition*, 27(6) :1411.

[Greenberg and Verfaellie, 2010] Greenberg, D. L. and Verfaellie, M. (2010). Interdependence of episodic and semantic memory : evidence from neuropsychology. *J Int Neuropsychol Soc*, 16(5) :748–753. 20561378[pmid].

[Grouin et al., 2014] Grouin, C., Lavergne, T., and Névéol, A. (2014). Optimizing annotation efforts to build reliable annotated corpora for training statistical models. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 54–58, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

[Gunasekara et al., 2020] Gunasekara, R. C., Kim, S., D'Haro, L. F., Rastogi, A., Chen, Y., Eric, M., Hedayatnia, B., Gopalakrishnan, K., Liu, Y., Huang, C., Hakkani-Tür, D., Li, J., Zhu, Q., Luo, L., Liden, L., Huang, K., Shayandeh, S., Liang, R., Peng, B., Zhang, Z., Shukla, S., Huang, M., Gao, J., Mehri, S., Feng, Y., Gordon, C., Alavi, S. H., Traum, D. R., Eskénazi, M., Beirami, A., Cho, E., Crook, P. A., De, A., Geramifard, A., Kottur, S., Moon, S., Poddar, S., and Subba, R. (2020). Overview of the ninth dialog system technology challenge : DSTC9. *CoRR*, abs/2011.06486.

[Hahn et al., 2011] Hahn, S., Dinarelli, M., Raymond, C., Lefevre, F., Lehnen, P., De Mori, R., Moschitti, A., Ney, H., and Riccardi, G. (2011). Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6) :1569–1583.

[Halliday et al., 1964] Halliday, M. A. K. et al. (1964). The linguistic sciences and language teaching.

[Han et al., 2020a] Han, T., Liu, X., Takanobu, R., Lian, Y., Huang, C., Peng, W., and Huang, M. (2020a). Multiwoz 2.3 : A multi-domain task-oriented dataset enhanced with annotation corrections and co-reference annotation. *CoRR*, abs/2010.05594.

[Han et al., 2020b] Han, T., Liu, X., Takanobu, R., Lian, Y., Huang, C., Peng, W., and Huang, M. (2020b). Multiwoz 2.3 : A multi-domain task-oriented

dataset enhanced with annotation corrections and co-reference annotation. *arXiv preprint arXiv :2010.05594*.

[Hancock et al., 2019] Hancock, B., Bordes, A., Mazare, P.-E., and Weston, J. (2019). Learning from dialogue after deployment : Feed yourself, chatbot ! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.

[Hardy et al., 2004] Hardy, H., Strzalkowski, T., Wu, M., Ursu, C., Webb, N., Biermann, A., Inouye, R. B., and McKenzie, A. (2004). Data-driven strategies for an automated dialogue system. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Hayes et al., 2017] Hayes, J., Melis, L., Danezis, G., and Cristofaro, E. D. (2017). LOGAN : evaluating privacy leakage of generative models using generative adversarial networks. *CoRR*, abs/1705.07663.

[Hemphill et al., 1990] Hemphill, C. T., Godfrey, J. J., and Doddington, G. R. (1990). The ATIS spoken language systems pilot corpus. In *Speech and Natural Language : Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

[Henderson et al., 2014a] Henderson, M., Thomson, B., and Williams, J. D. (2014a). The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

[Henderson et al., 2014b] Henderson, M., Thomson, B., and Williams, J. D. (2014b). The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

[Henderson et al., 2014c] Henderson, M., Thomson, B., and Williams, J. D. (2014c). The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329.

[Higashinaka et al., 2015] Higashinaka, R., Mizukami, M., Funakoshi, K., Araki, M., Tsukahara, H., and Kobayashi, Y. (2015). Fatal or not ? finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248.

[Hildebrand et al., 2005] Hildebrand, A. S., Eck, M., Vogel, S., and Waibel, A. (2005). Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the 10th EAMT Conference : Practical applications of machine translation*.

[Hintze and El Emam, 2018] Hintze, M. and El Emam, K. (2018). Comparing the benefits of pseudonymisation and anonymisation under the GDPR. *Journal of Data Protection & Privacy*, 2(2) :145–158.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8) :1735-1780.

[Hofstadter, 1996] Hofstadter, D. R. (1996). *Fluid Concepts and Creative Analogies : Computer Models of the Fundamental Mechanisms of Thought*. Basic Books, Inc., USA.

[Hori et al., 2019] Hori, C., Perez, J., Higashinaka, R., Hori, T., Boureau, Y.-L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., and Kim, S. (2019). Overview of the sixth dialog system technology challenge : DSTC6. *Computer Speech & Language*, 55 :1–25.

[Hovy, 1990] Hovy, E. H. (1990). Pragmatics and natural language generation. *Artificial Intelligence*, 43(2) :153–197.

[Hu et al., 2020] Hu, J., Yang, Y., Chen, C., He, L., and Yu, Z. (2020). SAS : Dialogue state tracking via slot attention and slot information sharing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6366–6375, Online. Association for Computational Linguistics.

[Huang et al., 2018] Huang, X., Liu, L., Carey, K., Woolley, J., Scherer, S., and Borsari, B. (2018). Modeling temporality of human intentions by domain adaptation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 696–701, Brussels, Belgium. Association for Computational Linguistics.

[Hussain et al., 2019] Hussain, S., Ameri Sianaki, O., and Ababneh, N. (2019). A survey on conversational agents/chatbots classification and design techniques. In Barolli, L., Takizawa, M., Xhafa, F., and Enokido, T., editors, *Web, Artificial Intelligence and Network Applications*, pages 946–956, Cham. Springer International Publishing.

[Io and Lee, 2017] Io, H. and Lee, C. (2017). Chatbots and conversational agents : A bibliometric analysis. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 215–219. IEEE.

[Issarny et al., 2005] Issarny, V., Sacchetti, D., Tartanoglu, F., Sailhan, F., Chibout, R., Lévy, N., and Talamona, A. (2005). Developing Ambient Intelligence Systems : A Solution based on Web Services. *Automated Software Engineering*, 12(1) :101–137.

[Jagfeld et al., 2018] Jagfeld, G., Jenne, S., and Vu, N. T. (2018). Sequence-to-sequence models for data-to-text natural language generation : word- vs. character-based processing and output diversity. *arXiv preprint arXiv :1810.04864*.

[Jakobson, 1960] Jakobson, R. (1960). Linguistics and poetics. In *Style in language*, pages 350–377. MA : MIT Press.

[Janarthanam, 2017] Janarthanam, S. (2017). *Hands-On Chatbots and Conversational UI Development : Build Chatbots and Voice User Interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*. Packt Publishing.

[Jang et al., 2020] Jang, B., Kim, M., Harerimana, G., Kang, S.-u., and Kim, J. W. (2020). Bi-lstm model to increase accuracy in text classification : Combining word2vec cnn and attention mechanism. *Applied Sciences*, 10(17).

[Jansen et al., 2013] Jansen, S., Wesselmeier, H., and Müller, H. M. (2013). Eeg responses to turn-taking anticipation in communication. *Aphasiology*, 12 :1007–1031.

[Jeon and Lee, 2021] Jeon, H. and Lee, G. G. (2021). Domain state tracking for a simplified dialogue system. *arXiv preprint arXiv :2103.06648*.

[Jobin et al., 2019] Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9) :389–399.

[Jokinen, 2018] Jokinen, K. (2018). *Dialogue Models for Socially Intelligent Robots : 10th International Conference, ICSR 2018, Qingdao, China, November 28 - 30, 2018, Proceedings*, pages 127–138.

[Jokinen and McTear, 2009] Jokinen, K. and McTear, M. (2009). Spoken dialogue systems. *Synthesis Lectures on Human Language Technologies*, 2(1) :1–151.

[Joulin et al., 2016] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759.

[Jurafsky and Martin, 2017] Jurafsky, D. and Martin, J. H. (2017). Dialog systems and chatbots. *Speech and language processing*, 3.

[Kelley, 1983] Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '83, pages 193–196, New York, NY, USA. ACM.

[Kerr and Moloney, 2018] Kerr, W. R. and Moloney, E. (2018). Vodafone : Managing advanced technologies and artificial intelligence.

[Kim et al., 2019] Kim, B., Chung, K., Lee, J., Seo, J., and Koo, M.-W. (2019). A bi-lstm memory network for end-to-end goal-oriented dialog learning. *Computer Speech & Language*, 53 :217–230.

[Kim et al., 2017a] Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., and Henderson, M. (2017a). *The Fourth Dialog State Tracking Challenge*, pages 435–449. Springer Singapore, Singapore.

[Kim et al., 2016] Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., Henderson, M., and Yoshino, K. (2016). The fifth dialog state tracking challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 511–517.

[Kim et al., 2020] Kim, S., Yang, S., Kim, G., and Lee, S.-W. (2020). Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 567–582, Online. Association for Computational Linguistics.

[Kim et al., 2017b] Kim, Y.-B., Lee, S., and Stratos, K. (2017b). Onenet : Joint domain, intent, slot prediction for spoken language understanding. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 547–553.

[Kingma and Welling, 2014] Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

[Kirinde Gamaarachchige and Inkpen, 2019] Kirinde Gamaarachchige, P. and Inkpen, D. (2019). Multi-task, multi-channel, multi-input learning for mental illness detection using social media text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 54–64, Hong Kong. Association for Computational Linguistics.

[Klatzky, 1980] Klatzky, R. (1980). Human memory : Structures and processes. *The American Journal of Psychology*, 93.

[Kleinberg et al., 2017] Kleinberg, B., Mozes, M., Toolen, Y., and Verschuere, B. (2017). Netanos - named entity-based text anonymization for open science.

[Kleinberg et al., 2018] Kleinberg, B., Mozes, M., van der Toolen, Y., and Verschuere, B. (2018). Netanos - named entity-based text anonymization for open science. *OSF*.

[Kuckartz, 2014] Kuckartz, U. (2014). *Qualitative text analysis : A guide to methods, practice and using software*. Sage.

[Kuhlen et al., 2017] Kuhlen, A. K., Bogler, C., Brennan, S. E., and Haynes, J.-D. (2017). Brains in dialogue : decoding neural preparation of speaking to a conversational partner. *Social Cognitive and Affective Neuroscience*, 12(6) :871–880.

[Kulhánek et al., 2021] Kulhánek, J., Hudecek, V., Nekvinda, T., and Dusek, O. (2021). Augpt : Dialogue with pre-trained language models and data augmentation. *CoRR*, abs/2102.05126.

[Kumar et al., 2020] Kumar, A., Di Eugenio, B., Aurisano, J., and Johnson, A. (2020). Augmenting small data to classify contextualized dialogue acts for exploratory visualization. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 590–599, Marseille, France. European Language Resources Association.

[Kumar, 2010] Kumar, V. (2010). Customer relationship management. *Wiley international encyclopedia of marketing*.

[la Brecque, 1972] la Brecque, M. (1972). Photographic memory. *Leonardo*, 5(4) :347–349.

[Labov, 1973] Labov, W. (1973). *Sociolinguistic patterns*. Number 4. University of Pennsylvania press.

[Lamel et al., 1998] Lamel, L., Rosset, S., Gauvain, J., Bennacef, S., Garnier-Rizet, M., and Prouts, B. (1998). The limsi arise system [rail travel information system]. pages 209 – 214.

[Landragin, 2008] Landragin, F. (2008). Vers l'identification et le traitement des actes de dialogue composites. In *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 151–160, Avignon, France. ATALA.

[Landragin, 2013] Landragin, F. (2013). *Dialogue homme-machine*. Hermès Science-Lavoisier.

[Larson, 2003] Larson, J. A. (2003). Voicexml and the w3c speech interface framework. *IEEE MultiMedia*, 10(4) :91–93.

[Larsson, 2017] Larsson, S. (2017). User-initiated sub-dialogues in state-of-the-art dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 17–22, Saarbrücken, Germany. Association for Computational Linguistics.

[Laurent, 2014] Laurent, R. (2014). *COSMO : un modèle bayésien des interactions sensori-motrices dans la perception de la parole*. Theses, Université de Grenoble.

[Lavie and Agarwal, 2007] Lavie, A. and Agarwal, A. (2007). Meteor : An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lecun, 1985] Lecun, Y. (1985). Une procedure d'apprentissage pour reseau a seuil asymmetrique (a learning scheme for asymmetric threshold networks). In *Proceedings of Cognitiva 85, Paris, France*, pages 599–604.

[Lee et al., 2009] Lee, C., Jung, S., Kim, S., and Lee, G. G. (2009). Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication*, 51(5) :466–484.

[Lemon and Pietquin, 2012] Lemon, O. and Pietquin, O. (2012). *Data-driven methods for adaptive spoken dialogue systems : Computational learning for conversational interfaces*. Springer Science & Business Media.

[Lewis et al., 2017] Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., and Batra, D. (2017). Deal or no deal ? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

[L'Haridon and Paraschiv, 2009] L'Haridon, O. and Paraschiv, C. (2009). Choix individuel et décision fondée sur l'expérience. une étude expérimentale. *Revue économique*, 60(4) :949–978.

[Li, 2020] Li, J. (2020). Teaching machines to converse. *CoRR*, abs/2001.11701.

[Li et al., 2016] Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., and Dolan, B. (2016). A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

[Li et al., 2021a] Li, J., Liu, C., Tao, C., Chan, Z., Zhao, D., Zhang, M., and Yan, R. (2021a). Dialogue history matters ! personalized response selection in multi-turn retrieval-based chatbots. *ACM Trans. Inf. Syst.*, 39(4).

[Li et al., 2021b] Li, L., Zhang, Y., and Chen, L. (2021b). Personalized transformer for explainable recommendation. In *ACL*.

[Li et al., 2020a] Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y.-L., Cho, K., and Weston, J. (2020a). Don't say that ! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

[Li et al., 2020b] Li, M., Xiong, H., and Cao, Y. (2020b). The SPPD system for schema guided dialogue state tracking challenge. *CoRR*, abs/2006.09035.

[Li et al., 2019] Li, R., Lin, C., Collinson, M., Li, X., and Chen, G. (2019). A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392, Hong Kong, China. Association for Computational Linguistics.

[Liao et al., 2021] Liao, L., Long, L. H., Ma, Y., Lei, W., and Chua, T.-S. (2021). Dialogue State Tracking with Incremental Reasoning. *Transactions of the Association for Computational Linguistics*, 9 :557–569.

[Lin et al., 2020] Lin, Z., Kang, X., Li, G., Ji, F., Chen, H., and Zhang, Y. (2020). "wait, i'm still talking !" predicting the dialogue interaction behavior using imagine-then-arbitrate model.

[Lindqvist, 2017] Lindqvist, J. (2017). New challenges to personal data processing agreements : is the GDPR fit to deal with contract, accountability and liability in a world of the Internet of Things ? *International Journal of Law and Information Technology*, 26(1) :45–63.

[Lison and Kennington, 2016] Lison, P. and Kennington, C. R. (2016). Opendial : A toolkit for developing spoken dialogue systems with probabilistic rules. In *ACL*.

[Litman et al., 2000] Litman, D., Kearns, M. S., Singh, S., and Walker, M. (2000). Automatic optimization of dialogue management. In *COLING 2000 Volume 1 : The 18th International Conference on Computational Linguistics*.

[Litman et al., 2006] Litman, D., Swerts, M., and Hirschberg, J. (2006). Characterizing and predicting corrections in spoken dialogue systems. *Computational Linguistics*, 32(3) :417–438.

[Liu and Lane, 2018a] Liu, B. and Lane, I. (2018a). Adversarial learning of task-oriented neural dialog models. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 350–359, Melbourne, Australia. Association for Computational Linguistics.

[Liu and Lane, 2018b] Liu, B. and Lane, I. (2018b). End-to-end learning of task-oriented dialogs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*, pages 67–73, New Orleans, Louisiana, USA. Association for Computational Linguistics.

[Liu and Mazumder, 2021] Liu, B. and Mazumder, S. (2021). Lifelong and continual learning dialogue systems : Learning during conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17) :15058–15063.

[Liu et al., 2016a] Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016a). How NOT to evaluate your dialogue system : An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

[Liu and Perez, 2017] Liu, F. and Perez, J. (2017). Gated end-to-end memory networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

[Liu et al., 2016b] Liu, G.-Z., Lin, V., Kou, X., and Wang, H.-Y. (2016b). Best practices in l2 english source use pedagogy : A thematic review and synthesis of empirical studies. *Educational Research Review*, 19 :36–57.

[Lopez, 1999] Lopez, P. (1999). *Analyse d'énoncés oraux pour le dialogue homme-machine à l'aide de grammaires lexicalisées d'arbres*. PhD thesis, Université Henri Poincaré-Nancy 1.

[Luria, 1987] Luria, A. R. (1987). *The Mind of a Mnemonist : A Little Book about a Vast Memory, With a New Foreword by Jerome S. Bruner*. Harvard University Press.

[Lykartsis and Kotti, 2019] Lykartsis, A. and Kotti, M. (2019). Prediction of user emotion and dialogue success using audio spectrograms and convolutional neural networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and*

Dialogue, pages 336–344, Stockholm, Sweden. Association for Computational Linguistics.

[Ma et al., 2019] Ma, M., Huang, L., Xiong, H., Zheng, R., Liu, K., Zheng, B., Zhang, C., He, Z., Liu, H., Li, X., Wu, H., and Wang, H. (2019). Stacl : Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

[Macken et al., 2014] Macken, B., Taylor, J. C., and Jones, D. M. (2014). Language and short-term memory : the role of perceptual-motor affordance. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 40(5) :1257.

[Madotto et al., 2018] Madotto, A., Wu, C.-S., and Fung, P. (2018). Mem2Seq : Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1468–1478, Melbourne, Australia. Association for Computational Linguistics.

[Maes, 1987] Maes, P. (1987). Concepts and experiments in computational reflection. *ACM Sigplan Notices*, 22(12) :147–155.

[Mandel et al., 2016] Mandel, A., Bourguignon, M., Parkkonen, L., and Hari, R. (2016). Sensorimotor activation related to speaker vs. listener role during natural conversation. *Neuroscience letters*, 614 :99–104.

[Manning and Schutze, 1999] Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

[Manzoor and Jannach, 2021] Manzoor, A. and Jannach, D. (2021). Generation-based vs. retrieval-based conversational recommendation : A user-centric comparison. In *Fifteenth ACM Conference on Recommender Systems*, pages 515–520.

[Mast et al., 1996] Mast, M., Kompe, R., Harbeck, S., Kiessling, A., Niemann, H., Noth, E., Schukat-Talamazzini, E., and Warnke, V. (1996). Dialog act classification with the help of prosody. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1732–1735 vol.3.

[Campillos Llanos et al., 2017] Campillos Llanos, L., Rosset, S., and Zweigenbaum, P. (2017). Automatic classification of doctor-patient questions for a virtual patient record query task. In ACL, editor, *BioNLP Shared-Task Workshop*, Vancouver, Canada. ACL.

[McCabe et al., 2010] McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., and Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning : evidence for a common executive attention construct. *Neuropsychology*, 24(2) :222–243. 20230116[pmid].

[McTear, 2020] McTear, M. (2020). Conversational ai : Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, 13(3) :1–251.

[Mecklinger et al., 2004] Mecklinger, A., Gruenewald, C., Weiskopf, N., and Doeller, C. F. (2004). Motor affordance and its role for visual working memory : Evidence from fmri studies. *Experimental Psychology*, 51(4) :258–269. PMID : 15620227.

[Megyesi et al., 2018] Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M., and Volodina, E. (2018). Learner corpus anonymization in the age of GDPR : Insights from the creation of a learner corpus of Swedish. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 47–56, Stockholm, Sweden. LiU Electronic Press.

[Mezza et al., 2018] Mezza, S., Cervone, A., Stepanov, E., Tortoreto, G., and Riccardi, G. (2018). ISO-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3539–3551, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[Mi et al., 2021] Mi, F., Zhou, W., Cai, F., Kong, L., Huang, M., and Faltings, B. (2021). Self-training improves pre-training for few-shot learning in task-oriented dialog systems.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

[Miller, 1962] Miller, B. (1962). "physiologie de l'hippocampe". Colloques internationaux du Centre national de la recherche scientifique.

[Montemurro and Zanette, 2011] Montemurro, M. A. and Zanette, D. H. (2011). Universal entropy of word ordering across linguistic families. *PLOS ONE*, 6(5) :1–9.

[Morgan-Short et al., 2014] Morgan-Short, K., Faretta-Stutenberg, M., Brill-Schuetz, K. A., Carpenter, H., and Wong, P. C. (2014). Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism : Language and Cognition*, 17(1) :56–72.

[Morik, 1985] Morik, K. (1985). User modelling, dialog structure, and dialog strategy in HAM-ANS. In *Second Conference of the European Chapter of the Association for Computational Linguistics*, Geneva, Switzerland. Association for Computational Linguistics.

[Nergiz and Clifton, 2007] Nergiz, M. E. and Clifton, C. (2007). Thoughts on k-anonymization. *Data & Knowledge Engineering*, 63(3) :622 – 645. 25th International Conference on Conceptual Modeling (ER 2006).

[Nguyen et al., 2019] Nguyen, T. T., Nguyen, C. M., Tien Nguyen, D., Thanh Nguyen, D., and Nahavandi, S. (2019). Deep learning for deepfakes creation and detection : A survey. *arXiv e-prints*, pages arXiv–1909.

[Nie et al., 2019] Nie, W., Narodytska, N., and Patel, A. (2019). RelGAN : Relational generative adversarial networks for text generation. In *International Conference on Learning Representations*.

[Nivre et al., 2016] Nivre, J., Marneffe, M.-C. d., Ginter, F., Goldberg, Y., Haji, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1 : A Multilingual Treebank Collection. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portoro, Slovenia. European Language Resources Association (ELRA).

[Noh et al., 2015] Noh, T.-G., Padó, S., Shwartz, V., Dagan, I., Nastase, V., Eichler, K., Kotlerman, L., and Adler, M. (2015). Multi-level alignments as an extensible representation basis for textual entailment algorithms. In *Proceedings of the fourth joint conference on lexical and computational semantics*, pages 193–198.

[Olabiyi et al., 2019] Olabiyi, O., Salimov, A. O., Khazane, A., and Mueller, E. (2019). Multi-turn dialogue response generation in an adversarial learning framework. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 121–132, Florence, Italy. Association for Computational Linguistics.

[Ouali et al., 2019] Ouali, L. O., Sabouret, N., and Rich, C. (2019). Guess my power : A computational model to simulate a partner's behavior in the context of collaborative negotiation. In Arai, K., Kapoor, S., and Bhatia, R., editors, *Intelligent Systems and Applications*, pages 1317–1337, Cham. Springer International Publishing.

[Owen et al., 1992] Owen, A., James, M., Leigh, P., Summers, B., Marsden, C., Quinn, N. a., Lange, K. W., and Robbins, T. (1992). Fronto-striatal cognitive deficits at different stages of parkinson's disease. *Brain*, 115(6) :1727–1751.

[Paek, 2006] Paek, T. (2006). Reinforcement learning for spoken dialogue systems : Comparing strengths and weaknesses for practical deployment. In *Proc. Dialog-on-Dialog Workshop, Interspeech*. Citeseer.

[Papangelis et al., 2017a] Papangelis, A., Kotti, M., and Stylianou, Y. (2017a). Predicting dialogue success, naturalness, and length with acoustic features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5010–5014. IEEE.

[Papangelis et al., 2017b] Papangelis, A., Kotti, M., and Stylianou, Y. (2017b). Predicting dialogue success, naturalness, and length with acoustic features. In

2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5010–5014.

[Papineni et al., 2002a] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002a). Bleu : A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Papineni et al., 2002b] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002b). Bleu : A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311-318, USA. Association for Computational Linguistics.

[Pasupat et al., 2019] Pasupat, P., Gupta, S., Mandyam, K., Shah, R., Lewis, M., and Zettlemoyer, L. (2019). Span-based hierarchical semantic parsing for task-oriented dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1520–1526, Hong Kong, China. Association for Computational Linguistics.

[Payne, 1987] Payne, D. G. (1987). Hypermnesia and reminiscence in recall : A historical and empirical review. *Psychological Bulletin*, 101(1) :5.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn : Machine learning in python. *the Journal of machine Learning research*, 12 :2825–2830.

[Pei et al., 2021] Pei, J., Ren, P., and de Rijke, M. (2021). A cooperative memory network for personalized task-oriented dialogue systems with incomplete user profiles. In *Proceedings of the Web Conference 2021*, WWW '21, page 1552-1561, New York, NY, USA. Association for Computing Machinery.

[Peng et al., 2020] Peng, B., Li, C., Li, J., Shayandeh, S., Liden, L., and Gao, J. (2020). Soloist : Building task bots at scale with transfer learning and machine teaching. *arXiv preprint arXiv :2005.05298*.

[Peng et al., 2021] Peng, B., Li, C., Li, J., Shayandeh, S., Liden, L., and Gao, J. (2021). Soloist : BuildingTask Bots at Scale with Transfer Learning and Machine Teaching. *Transactions of the Association for Computational Linguistics*, 9 :807–824.

[Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

[Pinar Saygin et al., 2000] Pinar Saygin, A., Cicekli, I., and Akman, V. (2000). Turing test : 50 years later. *Minds and Machines*, 10(4) :463–518.

[Pontiki et al., 2014] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4 : Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

[Poudade, 2006] Poudade, J. (2006). *Emergence d'un lexique dans une population d'agents autonomes par l'action, la perception et la contingence sensorimotrice*. PhD thesis, Paris-Sud. Thèse de doctorat dirigée par Jardino, Michèle Informatique Paris 11 2006.

[Qin et al., 2021] Qin, L., Gupta, A., Upadhyay, S., He, L., Choi, Y., and Faruqui, M. (2021). TIMEDIAL : Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 7066–7076, Online. Association for Computational Linguistics.

[Qiu et al., 2020] Qiu, L., Zhao, Y., Shi, W., Liang, Y., Shi, F., Yuan, T., Yu, Z., and Zhu, S.-C. (2020). Structured attention for unsupervised dialogue structure induction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1889–1899, Online. Association for Computational Linguistics.

[Quan and Xiong, 2020] Quan, J. and Xiong, D. (2020). Modeling long context for task-oriented dialogue state generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7119–7124, Online. Association for Computational Linguistics.

[Qun et al., 2020] Qun, H., Wenjing, L., and Zhangli, C. (2020). B&anet : Combining bidirectional lstm and self-attention for end-to-end learning of task-oriented dialogue system. *Speech Communication*, 125 :15–23.

[Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

[Radziwill and Benton, 2017] Radziwill, N. M. and Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv :1704.04579*.

[Raheja and Tetreault, 2019] Raheja, V. and Tetreault, J. (2019). Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.

[Rajendran et al., 2019] Rajendran, J., Ganhotra, J., and Polymenakos, L. (2019). Learning end-to-end goal-oriented dialog with maximal user task success and minimal human agent use. *CoRR*, abs/1907.07638.

[Rajendran et al., 2018] Rajendran, J., Ganhotra, J., Singh, S., and Polymenakos, L. (2018). Learning end-to-end goal-oriented dialog with multiple answers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3834–3843, Brussels, Belgium. Association for Computational Linguistics.

[Rame et al., 2018] Rame, A., Garreau, E., Ben-Younes, H., and Ollion, C. (2018). OMNIA Faster R-CNN : Detection in the wild through dataset merging and soft distillation. *arXiv e-prints*, page arXiv :1812.02611.

[Rao et al., 2021] Rao, S., Li, Y., Ramakrishnan, R., Hassaine, A., Canoy, D., Cleland, J., Lukasiewicz, T., Salimi-Khorshidi, G., and Rahimi, K. (2021). An explainable transformer-based deep learning model for the prediction of incident heart failure. *arXiv preprint arXiv :2101.11359*.

[Rastogi et al., 2017] Rastogi, A., Hakkani-Tur, D., and Heck, L. (2017). Scalable multi-domain dialogue state tracking. In *Proceedings of IEEE ASRU*.

[Rastogi et al., 2020] Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. (2020). Schema-guided dialogue state tracking task at dstc8. *arXiv preprint arXiv :2002.01359*.

[Raubal and Moratz, 2008] Raubal, M. and Moratz, R. (2008). A functional model for affordance-based agents. In Rome, E., Hertzberg, J., and Dorffner, G., editors, *Towards Affordance-Based Robot Control*, pages 91–105, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Ren et al., 2018] Ren, L., Xie, K., Chen, L., and Yu, K. (2018). Towards universal dialogue state tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786, Brussels, Belgium. Association for Computational Linguistics.

[Rescigno et al., 2020] Rescigno, A. A., Monti, J., Way, A., and Vanmassenhove, E. (2020). A case study of natural gender phenomena in translation : A comparison of Google Translate, bing Microsoft translator and DeepL for English to Italian, French and Spanish. In *Workshop on the Impact of Machine Translation (iMpacT 2020)*, pages 62–90, Virtual. Association for Machine Translation in the Americas.

[Roediger, 1990] Roediger, H. (1990). Implicit memory : Retention without remembering. *The American psychologist*, 45 :1043–56.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6) :386.

[Rosset, 2018] Rosset, S. (2018). Introduction 1 aux systèmes de dialogue.

[Roy et al., 2000] Roy, N., Pineau, J., and Thrun, S. (2000). Spoken dialogue management using probabilistic reasoning. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 93–100.

[Rubin, 1978] Rubin, A. D. (1978). A theoretical taxonomy of the differences between oral and written language. *Center for the Study of Reading Technical Report ; no. 035*.

[Rubin, 1975] Rubin, J. (1975). What the" good language learner" can teach us. *TESOL quarterly*, pages 41–51.

[Rudner and Rönnberg, 2008] Rudner, M. and Rönnberg, J. (2008). The role of the episodic buffer in working memory for language processing. *Cognitive processing*, 9 :19–28.

[Ruede et al., 2017] Ruede, R., Müller, M., Stüker, S., and Waibel, A. (2017). Enhancing backchannel prediction using word embeddings. In *Interspeech*, pages 879–883.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088) :533–536.

[S. Wallace, 2009] S. Wallace, R. (2009). *The anatomy of A.L.I.C.E*, pages 181–210.

[Sadoun et al., 2016] Sadoun, D., Mkhitaryan, S., Nouvel, D., and Valette, M. (2016). The MultiTal NLP tool infrastructure. In *Language Technology Resources and Tools for Digital Humanities*, pages 156 – 163, Osaka, Japan.

[Salmon Alt, 2001] Salmon Alt, S. (2001). *Référence et dialogue finalisé : de la linguistique à un modèle opérationnel*. Theses, Université Henri Poincaré - Nancy 1.

[Samarati and Sweeney, 1998] Samarati, P. and Sweeney, L. (1998). Protecting privacy when disclosing information : k-anonymity and its enforcement through generalization and suppression. Technical report.

[Sammut and Webb, 2010] Sammut, C. and Webb, G. I., editors (2010). *POMDPs*, pages 776–776. Springer US, Boston, MA.

[Sams et al., 1993] Sams, M., Hari, R., Rif, J., and Knuutila, J. (1993). The human auditory sensory memory trace persists about 10 sec : neuromagnetic evidence. *Journal of cognitive neuroscience*, 5(3) :363–370.

[Sanh et al., 2019] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert : smaller, faster, cheaper and lighter. *arXiv preprint arXiv :1910.01108*.

[Sankar et al., 2019] Sankar, C., Subramanian, S., Pal, C., Chandar, S., and Bengio, Y. (2019). Do neural dialog systems use the conversation history effectively ?

an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics.

[Schaub et al., 2020] Schaub, L.-P., Bruzaud, C., and Paroubek, P. (2020). GDPR Compliance for task-oriented dialog systems conception. In *workshop on Legal and Ethical Issues (Legal2020)*, volume 1, Marseille, France. LREC, European Language Resources Association (ELRA).

[Schaub et al., 2021] Schaub, L.-P., Hudecek, V., Stancl, D., Dusek, O., and Paroubek, P. (2021). Définition et détection des incohérences du système dans les dialogues orientés tâche. (we present experiments on automatically detecting inconsistent behavior of task-oriented dialogue systems from the context). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 142–152, Lille, France. ATALA.

[Schaub and Vaudapiviz, 2019a] Schaub, L.-P. and Vaudapiviz, C. (2019a). Goal-oriented dialog systems and Memory : an overview. In *9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland. Zygmunt Vetulani & Patrick Paroubek.

[Schaub and Vaudapiviz, 2019b] Schaub, L.-P. and Vaudapiviz, C. (2019b). Les systèmes de dialogue orientés-but : état de l'art et perspectives d'amélioration (goal-oriented dialog systems : a recent overview and research prospects). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume III : RECITAL*, pages 541–562, Toulouse, France. ATALA.

[Schegloff et al., 1977] Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2) :361–382.

[Schlangen and Skantze, 2009] Schlangen, D. and Skantze, G. (2009). A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 710–718, Athens, Greece. Association for Computational Linguistics.

[Seneff and Polifroni, 2000] Seneff, S. and Polifroni, J. (2000). Dialogue management in the mercury flight reservation system. In *ANLP-NAACL 2000 Workshop : Conversational Systems*.

[Serban et al., 2016] Serban, I., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

[Serban et al., 2018] Serban, I. V., Lowe, R., Charlin, L., and Pineau, J. (2018). A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *Dialogue & Discourse*, 9(1) :1–49. arXiv : 1512.05742.

[Serban et al., 2015] Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2015). A survey of available corpora for building data-driven dialogue systems. *CoRR*, abs/1512.05742.

[Shalyminov et al., 2020] Shalyminov, I., Sordoni, A., Atkinson, A., and Schulz, H. (2020). Fast domain adaptation for goal-oriented dialogue using a hybrid generative-retrieval transformer. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8039–8043.

[Shang et al., 2020] Shang, G., Tixier, A. J.-P., Vazirgiannis, M., and Lorré, J.-P. (2020). Speaker-change aware crf for dialogue act classification.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27 :379–423.

[Shawar and Atwell, 2003] Shawar, B. A. and Atwell, E. (2003). Using dialogue corpora to train a chatbot. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 681–690.

[Sheng et al., 2021] Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2021). "nice try, kiddo" : Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 750–767, Online. Association for Computational Linguistics.

[Shriberg et al., 2004] Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

[Shriberg et al., 1998] Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., Taylor, P., Ries, K., Martin, R., and van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech ? *Language and Speech*, 41(3-4) :443–492. PMID : 10746366.

[Shum et al., 2018] Shum, H.-y., He, X.-d., and Li, D. (2018). From eliza to xiaoice : challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1) :10–26.

[Singh et al., 2002] Singh, S., Litman, D., Kearns, M., and Walker, M. (2002). Optimizing dialogue management with reinforcement learning : Experiments with the njfun system. *Journal of Artificial Intelligence Research*, 16 :105–133.

[Skantze, 2017] Skantze, G. (2017). Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, Saarbrücken, Germany. Association for Computational Linguistics.

[Small, 2002] Small, G. W. (2002). What we need to know about age related memory loss. *Bmj*, 324(7352) :1502–1505.

[Sperling, 1967] Sperling, G. (1967). Successive approximations to a model for short term memory. *Acta Psychologica*, 27 :285 – 292.

[Spiekermann, 2012] Spiekermann, S. (2012). The challenges of privacy by design. *Commun. ACM*, 55(7) :38–40.

[Squire and Zola, 1996] Squire, L. and Zola, S. (1996). Structure and function of declarative and nondeclarative memory. *Proceedings of the National Academy of Sciences*, 93.

[Sreelakshmi et al., 2018] Sreelakshmi, K., Rafeeque, P., Sreetha, S., and Gayathri, E. (2018). Deep bi-directional lstm network for query intent detection. *Procedia computer science*, 143 :939–946.

[Srivastava et al., 2015] Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Training very deep networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

[Steels, 2000] Steels, L. (2000). The emergence of grammar in communicating autonomous robotic agents. In Horn, H., editor, *Proceedings of European Conference on Artificial Intelligence, ECAI*, pages 764–769, Amsterdam. IOS Press.

[Stolcke et al., 1998] Stolcke, A., Shriberg, E., Bates, R., Coccaro, N., Jurafsky, D., Martin, R., Meteer, M., Ries, K., Taylor, P., Van Ess-Dykema, C., et al. (1998). Dialog act modeling for conversational speech. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 98–105.

[Stolz, 2010] Stolz, T. (2010). Denis creissels, syntaxe générale. une introduction typologique. 1 : Catégories et constructions. 2 : La phrase, (collection langues et syntaxe), paris : Hermes sciences Ű lavoisier, 2006. vol. 1 : xviii + 412 pp.,vol. 2 : 334 pp. 63(02) :165–167.

[Strubell et al., 2019] Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

[Su et al., 2018a] Su, H., Shen, X., Hu, P., Li, W., and Chen, Y. (2018a). Dialogue generation with gan. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Su et al., 2019] Su, M.-H., Wu, C.-H., and Chen, L.-Y. (2019). Attention-based response generation using parallel double q-learning for dialog policy decision in a conversational system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :131–143.

[Su et al., 2017] Su, M.-H., Wu, C.-H., Huang, K.-Y., Hong, Q.-B., and Wang, H.-M. (2017). A chatbot using lstm-based multi-layer embedding for elderly

care. In *2017 International Conference on Orange Technologies (ICOT)*, pages 70–74. IEEE.

[Su et al., 2018b] Su, P.-H., Gai, M., and Young, S. (2018b). Reward estimation for dialogue policy optimisation. *Computer Speech and Language*, 51 :24 – 43.

[Su et al., 2021] Su, Y., Shu, L., Mansimov, E., Gupta, A., Cai, D., Lai, Y.-A., and Zhang, Y. (2021). Multi-task pre-training for plug-and-play task-oriented dialogue system.

[Sukhbaatar et al., 2015] Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. (2015). End-to-end memory networks. *Advances in Neural Information Processing Systems*, 2015 :2440–2448.

[Surendran and Levow, 2006] Surendran, D. and Levow, G.-A. (2006). Dialog act tagging with support vector machines and hidden markov models. In *Interspeech*. Citeseer.

[Swanborn and de Glopper, 1999] Swanborn, M. and de Glopper, K. (1999). Incidental word learning while reading : A meta-analysis. *Review of Educational Research*, 69(3) :261–285.

[Sykes, 2005] Sykes, J. M. (2005). Synchronous cmc and pragmatic development : Effects of oral and written chat. *CALICO Journal*, 22(3) :399–431.

[Szpunar et al., 2008] Szpunar, K. K., McDermott, K. B., and Roediger III, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 34(6) :1392.

[Tahara et al., 2019] Tahara, S., Ikeda, K., and Hoashi, K. (2019). Empathic dialogue system based on emotions extracted from tweets. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 52–56, New York, NY, USA. ACM.

[Tang, 2019] Tang, X. (2019). The effects of task modality on l2 chinese learners pragmatic development : Computer-mediated written chat vs. face-to-face oral chat. *System*, 80 :48–59.

[Tang, 2020] Tang, X. (2020). Task-based interactional sequences in different modalities : A comparison between computer-mediated written chat and face-to-face oral chat. *Applied Pragmatics*, 2(2) :174–198.

[Testoni and Bernardi, 2021] Testoni, A. and Bernardi, R. (2021). The interplay of task success and dialogue quality : An in-depth evaluation in task-oriented visual dialogues. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 2071–2082, Online. Association for Computational Linguistics.

[Tiedemann and Nygaard, 2004] Tiedemann, J. and Nygaard, L. (2004). The OPUS corpus - parallel and free : http://logos.uio.no/opus. In *Proceedings of the Fourth International Conference on Language Resources and Eva-*

luation (LREC'04), Lisbon, Portugal. European Language Resources Association (ELRA).

[Tjandra et al., 2017] Tjandra, A., Sakti, S., and Nakamura, S. (2017). Listening while speaking : Speech chain by deep learning. *CoRR*, abs/1707.04879.

[Tseng et al., 2021] Tseng, B.-H., Dai, Y., Kreyssig, F., and Byrne, B. (2021). Transferable dialogue systems and user simulators. *arXiv preprint arXiv :2107.11904*.

[Tulving, 1985] Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26(1) :1–12.

[TURING, 1950] TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236) :433–460.

[Udagawa and Aizawa, 2020] Udagawa, T. and Aizawa, A. (2020). An annotated corpus of reference resolution for interpreting common grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05) :9081–9089.

[Ultes et al., 2017a] Ultes, S., Rojas-Barahona, L. M., Su, P.-H., Vandyke, D., Kim, D., Casanueva, I., Budzianowski, P., Mrkšić, N., Wen, T.-H., Gašić, M., and Young, S. (2017a). PyDial : A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, Vancouver, Canada. Association for Computational Linguistics.

[Ultes et al., 2017b] Ultes, S., Rojas Barahona, L. M., Su, P.-H., Vandyke, D., Kim, D., Casanueva, I. n., Budzianowski, P., Mrkšić, N., Wen, T.-H., Gasic, M., and Young, S. (2017b). PyDial : A Multi-domain Statistical Dialogue System Toolkit. pages 73–78.

[van Aken et al., 2019] van Aken, B., Winter, B., Löser, A., and Gers, F. A. (2019). How does bert answer questions ? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1823–1832.

[Varanasi et al., 2020] Varanasi, S., Amin, S., and Neumann, G. (2020). CopyBERT : A unified approach to question generation with self-attention. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 25–31, Online. Association for Computational Linguistics.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, Long Beach, CA, USA.

[Veron, 2019] Veron, M. (2019). Lifelong learning et systèmes de dialogue : définition et perspectives. In *Rencontres des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, Toulouse, France.

[Visser et al., 2015] Visser, S., Thangarajah, J., Harland, J., and Dignum, F. (2015). Preference-based reasoning in bdi agent systems. *Autonomous Agents and Multi-Agent Systems*, 30.

[Vlasov et al., 2019] Vlasov, V., Mosig, J. E., and Nichol, A. (2019). Dialogue transformers. *arXiv preprint arXiv :1910.00486*.

[Wachowski et al., 1999] Wachowski, A., Wachowski, L., Reeves, K., Fishburne, L., Moss, C.-A., Weaving, H., Foster, G., Pantoliano, J., and Staenberg, Z. (1999). *Matrix*. Warner Home Video Burbank.

[Wadsley and Ryan, 2013] Wadsley, T. and Ryan, M. (2013). A belief-desire-intention model for narrative generation. In *Intelligent Narrative Technologies - Papers from the 2013 AIIDE Workshop, Technical Report*, volume WS-13-21, pages 105–108, United States. AI Access Foundation.

[Wahlster and Kobsa, 1989] Wahlster, W. and Kobsa, A. (1989). User models in dialog systems. *User models in dialog systems*, pages 4–34.

[Walker et al., 2001] Walker, M., Aberdeen, J., Boland, J., Bratt, E., Garofolo, J., Hirschman, L., Le, A., Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Sanders, G., Seneff, S., Stallard, D., and Whittaker, S. (2001). Darpa communicator dialog travel planning systems : The june 2000 data collection. In *In proceedings of the 7th European Conference on Speech Processing (EUROSPEECH)*, pages 1371–1374, Aalborg, Denmark.

[Walker et al., 2000] Walker, M., Langkilde, I., Wright, J., Gorin, A., and Litman, D. (2000). Learning to predict problematic situations in a spoken dialogue system : Experiments with How May I Help You ? In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

[Walker, 1996] Walker, M. A. (1996). The effect of resource limits and task complexity on collaborative planning in dialogue. *Artificial Intelligence*, 85(1) :181–243.

[Walker et al., 1997a] Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997a). PARADISE : A framework for evaluating spoken dialogue agents. *CoRR*, cmp-lg/9704004.

[Walker et al., 1997b] Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997b). PARADISE : A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

[Wallace, 2004] Wallace, R. (2004). The elements of AIML style. ALICE AI Foundation.

[Wallace, 2009] Wallace, R. S. (2009). *The Anatomy of A.L.I.C.E.*, pages 181–210. Springer Netherlands, Dordrecht.

[Wang et al., 2020a] Wang, J., Liu, J., Bi, W., Liu, X., He, K., Xu, R., and Yang, M. (2020a). Dual dynamic memory network for end-to-end multi-turn task-oriented dialog systems. In *Proceedings of the 28th International Conference on*

Computational Linguistics, pages 4100–4110, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[Wang et al., 2020b] Wang, J., Liu, J., Bi, W., Liu, X., He, K., Xu, R., and Yang, M. (2020b). Dual dynamic memory network for end-to-end multi-turn task-oriented dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4100–4110.

[Wang et al., 2020c] Wang, K., Tian, J., Wang, R., Quan, X., and Yu, J. (2020c). Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7125–7134, Online. Association for Computational Linguistics.

[Weiss, 2020] Weiss, B. (2020). Talker quality in interactive scenarios. In *Talker Quality in Human and Machine Interaction*, pages 67–106. Springer.

[Weisser, 2016] Weisser, M. (2016). Dart–the dialogue annotation and research tool. *Corpus Linguistics and Linguistic Theory*, 12(2) :355–388.

[Weizenbaum, 1966] Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1) :36-45.

[Weld et al., 2021] Weld, H., Huang, X., Long, S., Poon, J., and Han, S. C. (2021). A survey of joint intent detection and slot-filling models in natural language understanding. *arXiv preprint arXiv :2101.08091*.

[Welleck et al., 2019] Welleck, S., Weston, J., Szlam, A., and Cho, K. (2019). Dialogue Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

[Wen et al., 2016a] Wen, T.-H., Gasic, M., Mrkšić, N., Barahona, L. M. R., Su, P.-H., Ultes, S., Vandyke, D., and Young, S. (2016a). Conditional generation and snapshot learning in neural dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2153–2162. Association for Computational Linguistics.

[Wen et al., 2016b] Wen, T.-H., Gašić, M., Mrkšić, N., M. Rojas-Barahona, L., Su, P.-H., Ultes, S., Vandyke, D., and Young, S. (2016b). Conditional generation and snapshot learning in neural dialogue systems. *arXiv preprint : 1606.03352*.

[Wen et al., 2016c] Wen, T.-H., Vandyke, D., Mrkšić, N., Gašić, M., M. Rojas-Barahona, L., Su, P.-H., Ultes, S., and Young, S. (2016c). A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint : 1604.04562*.

[Weng et al., 2020] Weng, Y., Miryala, S. S., Khatri, C., Wang, R., Zheng, H., Molino, P., Namazifar, M., Papangelis, A., Williams, H., Bell, F., and Tür, G. (2020). Joint contextual modeling for ASR correction and language understanding. *CoRR*, abs/2002.00750.

[Westerlund, 2019] Westerlund, M. (2019). The emergence of deepfake technology : A review. *Technology Innovation Management Review*, 9(11).

[Weston et al., 2015] Weston, J., Chopra, S., and Bordes, A. (2015). Memory networks. In *3rd International Conference on Learning Representations (ICLR2015)*, San Diego, CA, USA.

[Whiteside, 1993] Whiteside, S. P. (1993). Peter b. denes and elliot n. pinson the speech chain : The physics and biology of spoken language, 2nd edition. oxford : W.h. freeman and company, 1993. pp. 246 pb. us$14.95. isbn 0-7167-2344-1. *Journal of the International Phonetic Association*, 23(2) :98 - 101.

[Whitney et al., 2017] Whitney, D., Rosen, E., MacGlashan, J., Wong, L. L. S., and Tellex, S. (2017). Reducing errors in object-fetching interactions through social feedback. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1006–1013.

[Wik and Hjalmarsson, 2009] Wik, P. and Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech communication*, 51(10) :1024–1037.

[Williams, 2008] Williams, J. (2008). The best of both worlds : Unifying conventional dialog systems and pomdps. pages 1173–1176.

[Wobcke et al., 2005] Wobcke, W., Ho, V., Nguyen, A., and Krzywicki, A. (2005). A bdi agent architecture for dialogue modelling and coordination in a smart personal assistant. In *IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 323–329.

[Wu et al., 2018] Wu, C., Madotto, A., Winata, G. I., and Fung, P. (2018). End-to-end dynamic query memory network for entity-value independent task-oriented dialog. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6154–6158.

[Wu et al., 2020a] Wu, C.-S., Hoi, S. C., Socher, R., and Xiong, C. (2020a). TOD-BERT : Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

[Wu et al., 2019] Wu, C.-S., Madotto, A., Hosseini-Asl, E., Xiong, C., Socher, R., and Fung, P. (2019). Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

[Wu et al., 2020b] Wu, P., Zou, B., Jiang, R., and Aw, A. (2020b). Gcdst : A graph-based and copy-augmented multi-domain dialogue state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : Findings*, pages 1063–1073.

[Wu, 2012] Wu, X. (2012). Calculation of the minimum computational complexity based on information entropy. *arXiv preprint arXiv :1203.1792*.

[Xiong et al., 2016] Xiong, C., Merity, S., and Socher, R. (2016). Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406. PMLR.

[Xu and Rudnicky, 2000] Xu, W. and Rudnicky, A. I. (2000). Task-based dialog management using an agenda. In *ANLP-NAACL 2000 Workshop : Conversational Systems*.

[Xu and Reitter, 2018] Xu, Y. and Reitter, D. (2018). Information density converges in dialogue : Towards an information-theoretic model. *Cognition*, 170 :147–163.

[Xu et al., 2020] Xu, Z., Chen, Z., Chen, L., Zhu, S., and Yu, K. (2020). Memory attention neural network forămulti-domain dialogue state tracking. In Zhu, X., Zhang, M., Hong, Y., and He, R., editors, *Natural Language Processing and Chinese Computing*, pages 41–52, Cham. Springer International Publishing.

[Yamaguchi et al., 2021] Yamaguchi, A., Iwasa, K., and Fujita, K. (2021). Dialogue act-based breakdown detection in negotiation dialogues. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 745–757, Online. Association for Computational Linguistics.

[Yang et al., 2018] Yang, L., Liang, X., Wang, T., and Xing, E. (2018). Real-to-virtual domain unification for end-to-end autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[Yang et al., 2020] Yang, Y., Li, Y., and Quan, X. (2020). UBAR : towards fully end-to-end task-oriented dialog systems with GPT-2. *CoRR*, abs/2012.03539.

[Ye et al., 2021a] Ye, F., Manotumruksa, J., and Yilmaz, E. (2021a). Multiwoz 2.4 : A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *CoRR*, abs/2104.00773.

[Ye et al., 2021b] Ye, F., Manotumruksa, J., and Yilmaz, E. (2021b). Multiwoz 2.4 : A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv :2104.00773*.

[yew Lin, 2004] yew Lin, C. (2004). Rouge : a package for automatic evaluation of summaries. pages 25–26.

[Yong Nahm and J. Mooney, 2002] Yong Nahm, U. and J. Mooney, R. (2002). Text mining with information extraction.

[Young et al., 2013] Young, S., Gašić, M., Thomson, B., and Williams, J. D. (2013). Pomdp-based statistical spoken dialog systems : A review. *Proceedings of the IEEE*, 101(5) :1160–1179.

[Young et al., 2013] Young, S., Gai, M., Thomson, B., and Williams, J. D. (2013). POMDP-based statistical spoken dialog systems : A review. *Proceedings of the IEEE*, 101(5) :1160–1179.

[Yu and Yu, 2021] Yu, D. and Yu, Z. (2021). MIDAS : A dialog act annotation scheme for open domain HumanMachine spoken conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 1103–1120, Online. Association for Computational Linguistics.

[Zang et al., 2020] Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., and Chen, J. (2020). Multiwoz 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, pages 109–117.

[Zhai and Lafferty, 2017] Zhai, C. and Lafferty, J. (2017). A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, volume 51, pages 268–276. ACM New York, NY, USA.

[Zhang et al., 2019a] Zhang, H., Lan, Y., Pang, L., Guo, J., and Cheng, X. (2019a). ReCoSa : Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730, Florence, Italy. Association for Computational Linguistics.

[Zhang et al., 2020a] Zhang, J., Hashimoto, K., Wu, C.-S., Wang, Y., Yu, P., Socher, R., and Xiong, C. (2020a). Find or classify ? dual strategy for slot-value predictions on multi-domain dialog state tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 154–167, Barcelona, Spain (Online). Association for Computational Linguistics.

[Zhang et al., 2018] Zhang, Q., Liang, S., and Yilmaz, E. (2018). Variational self-attention model for sentence representation. *arXiv preprint arXiv :1812.11559*.

[Zhang et al., 2019b] Zhang, W., Feng, Y., Meng, F., You, D., and Liu, Q. (2019b). Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.

[Zhang et al., 2020b] Zhang, Y., Ou, Z., and Yu, Z. (2020b). Task-oriented dialog systems that consider multiple appropriate responses under the same context. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05) :9604–9611.

[Zhang et al., 2020c] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020c). DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics :*

System Demonstrations, pages 270–278, Online. Association for Computational Linguistics.

[Zhang et al., 2019c] Zhang, Z., Huang, M., Zhao, Z., Ji, F., Chen, H., and Zhu, X. (2019c). Memory-augmented dialogue management for task-oriented dialogue systems. *ACM Trans. Inf. Syst.*, 37(3) :34 :1–34 :30.

[Zhao et al., 2021] Zhao, J., Mahdieh, M., Zhang, Y., Cao, Y., and Wu, Y. (2021). Effective sequence-to-sequence dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7486–7493, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Zhong et al., 2005] Zhong, S., Yang, Z., and Wright, R. N. (2005). Privacy-enhancing k-anonymization of customer data. In *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pages 139–147, New York, NY, USA. ACM.

[Zhu et al., 2020a] Zhu, Q., Huang, K., Zhang, Z., Zhu, X., and Huang, M. (2020a). CrossWOZ : A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8 :281–295.

[Zhu et al., 2020b] Zhu, Q., Zhang, Z., Fang, Y., Li, X., Takanobu, R., Li, J., Peng, B., Gao, J., Zhu, X., and Huang, M. (2020b). ConvLab-2 : An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 142–149, Online. Association for Computational Linguistics.

[Zilka and Jurcicek, 2015] Zilka, L. and Jurcicek, F. (2015). Incremental lstm-based dialog state tracker. In *2015 Ieee Workshop on Automatic Speech Recognition and Understanding (Asru)*, pages 757–762. IEEE.