Modèle de traduction statistique à fragments enrichi par la syntaxe

Vassilina Nikoulina

Université de Grenoble

Ecole Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique

THÈSE

Presentée par Vassilina NIKOULINA

Pour l'obtention du grade de Docteur de l'Université de Grenoble

Spécialité : Informatique

Modèle de traduction statistique à fragments enrichi par la syntaxe

Thèse soutenue le 19 mars 2010.

Composition du jury :

Laurent BESACIER (président)
Igor BOGUSLAVSKY (rapporteur)
Yves LEPAGE (rapporteur)
Holger SCHWENK (rapporteur)
François YVON (examinateur)
Christian BOITET (directeur)
Marc DYMETMAN (co-directeur)

Thèse préparée au sein de l'équipe GETALP du Laboratoire d'Informatique de Grenoble (LIG) et de l'équipe CLT du Xerox Research Centre Europe (XRCE)

Résumé

Les modèles de traduction automatique probabiliste traditionnels ignorent la structure syntaxique des phrases source et cible. Le choix des unités lexicales cibles et leur ordre sont contrôlés uniquement par des statistiques de surface sur le corpus d'entraînement. La connaissance de la structure linguistique peut être bénéfique car elle fournit des informations génériques compensant la pauvreté des données directement observables.
Nos travaux ont pour but d'étudier l'impact des informations syntaxiques sur un modèle de traduction probabiliste de base, fondé sur des fragments [phrase-based], dans le cadre d'un analyseur dépendanciel particulier, XIP (Xerox Incremental Parser), dont la performance est bien adaptée à nos besoins.
Dans un premier temps, nous étudions l'intégration des informations syntaxiques dans un but de reclassement des traductions proposées par le modèle de base. Nous définissons un ensemble de traits mesurant la similarité entre les structures de dépendance source et cible pour contrôler l'adéquation de la traduction, et des traits de cohérence linguistique (basés uniquement sur l'analyse cible) pour contrôler la fluidité. L'apprentissage automatique des poids de ces traits permet de détecter l'importance de chaque trait.
L'évaluation manuelle des différents modèles de reclassement nous a permis de montrer le potentiel de ces traits à améliorer la qualité des traductions proposées par le modèle de base, ainsi que la déficience des mesures automatiques d'évaluation de la traduction (BLEU et NIST).
Dans la suite de nos travaux, nous avons proposé un modèle pour réduire la taille du graphe des hypothèses exploré par le modèle de base à l'aide de connaissances sur la structure syntaxique source. Nous avons également proposé une procédure de décomposition d'une phrase source initiale en sous-phrases pour simplifier la tâche de traduction. Les évaluations initiales de ces modèles se sont montrées prometteuses .

Abstract

Traditional Statistical Machine Translation models are not aware of linguistic structure. Thus, target lexical choices and word order are controlled only by surface-based statistics learned from the training corpus. Knowledge of linguistic structure can be beneficial since it provides generic information compensating data sparsity.
The purpose of our work is to study the impact of syntactic information while preserving the general framework of Phrase-Based SMT.
First, we study the integration of syntactic information using a reranking approach. We define features measuring the similarity between the dependency structures of source and target sentences, as well as features of linguistic coherence of the target sentence. The importance of each feature is assessed by learning the weights through a Structured Perceptron algorithm.
The evaluation of several reranking models shows that these features often improve the quality of translations produced by the basic model, in terms of manual evaluations, as opposed to automatic measures.
Then, we propose different models in order to increase the quality and diversity of the search graph produced by the decoder, through filtering out uninteresting hypotheses based on the source syntactic structure. This is done either by learning limits on the phrase reordering, or by decomposing the source sentence in order to simplify the translation process. The initial evaluations of these models look promising.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'abondance des ressources multilingues, le développement du Web multilingue et la nécessité de maintenir la documentation technique à jour dans plusieurs langues ont été, et sont toujours, des facteurs importants contribuant au développement de la traduction automatique. L'abondance des données a favorisé l'agrandissement de ce domaine et plusieurs lignes de recherche se développent activement, de nos jours, pour réexploiter les données multilingues afin de minimiser le travail humain lors de la conception du système de traduction et d'automatiser le processus de traduction autant que possible.

L'idée de la traduction automatique empirique "pure" est de réexploiter un grand volume de textes traduits précédemment et de créer des traductions de nouveaux textes automatiquement, sans intervention d'un expert humain à aucune étape. Ainsi, les modèles de traduction empirique pure n'exploitent aucune information linguistique et se basent uniquement sur des traductions brutes. Aussi, ces modèles ont besoin d'un volume important de textes traduits pour obtenir des traductions de qualité acceptable.

Les connaissances syntaxiques donnent des informations générales qui manquent au modèles probabilistes et peuvent améliorer les traductions, même en l'absence de grands volumes de textes parallèles. Dans nos travaux, nous nous sommes intéressée aux méthodes dites "hybrides", qui exploitent à la fois des connaissances expertes linguistiques, et des méthodes empiriques de traduction.

Nos travaux s'inscrivent dans le cadre des modèles de traduction à fragments, des modèles considérés comme à l'état de l'art parmi les modèles de TA probabiliste. Ainsi, nous proposons d'enrichir ces modèles avec des informations syntaxiques, tout en restant dans le paradigme de la traduction à fragments. Nous accordons une attention particulière à l'analyse humaine des traductions et des erreurs typiques faites par ces differents modèles lors de l'évaluation.

Cette thèse est organisée comme suit :

- Dans la première partie, nous présentons d'abord les modèles existants de la traduction automatique et l'état de l'art des modèles de la traduction dits "hybrides". Nous présentons ensuite les outils et les données que nous avons utilisés dans nos travaux. Enfin, nous décrivons les directions de recherche que nous avons suivies.
- Dans les deuxième et troisième parties, nous proposons différentes façons d'étendre un modèle de traduction probabiliste à fragments avec des connaissances syntaxiques. D'abord nous introduisons des fonctions de traits dépendant des analyses syntaxiques source et cible dans le cadre du reclassement des traductions générées par un modèle de traduction probabiliste. Nous évaluons ce modèle par des mesures automatiques, et en faisant une évaluation manuelle.
- Dans la troisième partie, nous proposons une autre façon d'étendre le modèle probabiliste à fragments, qui consiste à utiliser la structure syntaxique source pour simplifier la tâche de traduction : soit en guidant la traduction avec des contraintes syntaxiquement motivées, soit en décomposant une phrase source en plusieurs fragments qui peuvent être traduits de façon autonome.

Première partie

Positionnement du problème

Chapitre 1

État de l'art

Dans ce chapitre, nous présentons les travaux existants qui sont pertinents pour cette thèse. Nous allons d'abord résumer la situation de la Traduction Automatique en général. Nous présenterons ensuite les différents modèles de la TA probabiliste, qui est au centre de notre travail. Enfin, nous décrivons différentes approches d'hybridation existant dans la littérature : quelques exemples d'intégration des méthodes empiriques dans les systèmes de TA experte seront présentés brièvement, puis l'introduction de connaissances linguistiques dans un système de TA probabiliste sera étudiée plus en détail.

1.1 Introduction à la Traduction Automatique

Les premières idées sur la Traduction Automatique (TA) datent de l'année 1933, quand deux brevets furent déposés : l'un en France par Georges Arstrouni, et l'autre en URSS par Piotr Trojanski. Le brevet d'Astrouni proposait un système de traduction générique, fonctionnant comme un dictionnaire mécanique. Trojanski est allé plus loin en proposant un dictionnaire mécanique avec l'encodage et les interprétations des rôles grammaticaux en utilisant des symboles universels basés sur l'espéranto.

Les tentatives d'utilisation des premiers ordinateurs pour l'automatisation de la traduction ont été faites par Andrew Booth et Warren Weaver. En juin 1952, lors de la première conférence sur la traduction automatique organisée par Y. Bar-Hillel, il était déjà reconnu que la traduction entièrement automatisée de haute qualité pour la langue générale (FAHQMT) n'était pas réalisable et beaucoup d'accent a été mis sur la nécessité de la pré- et de post- édition.

La première démonstration d'un système de TA, comprenant un dictionnaire de 250 mots et 6 règles, a eu lieu le 7 janvier 1954 dans le cadre du « Georgetown Experiment ». Cette démonstration, où 49 phrases soigneusement choisies ont été traduites du russe vers l'anglais, incita de nombreuses équipes partout dans le monde (notamment en URSS et aux États-Unis) à entreprendre des recherches dans le domaine de la TA.

1.1.1 Architectures linguistiques

Trois types de base d'architectures linguistiques existent. Ces architectures correspondent à des types différents de chemins dans le fameux triangle de Vauquois (Figure 1.1). Nous allons brièvement présenter les principales architectures linguistiques . Plus de détails sur ces architectures avec des exemples plus complets peuvent être trouvés dans [Boitet, 2008].

1.1.1.1 Systèmes directs

Le modèle de TA *direct* opère directement au niveau du texte d'entrée (phrase source) et du texte de sortie (phrase cible), sans utilisation de représentations intermédiaires. Quand il existe une phase de segmentation et/ou d'analyse/génération morphologique, on parlera de traduction *semi-directe*. Les systèmes de TA probabiliste classique utilisent l'architecture directe (ou semi-directe) de traduction, ainsi que les

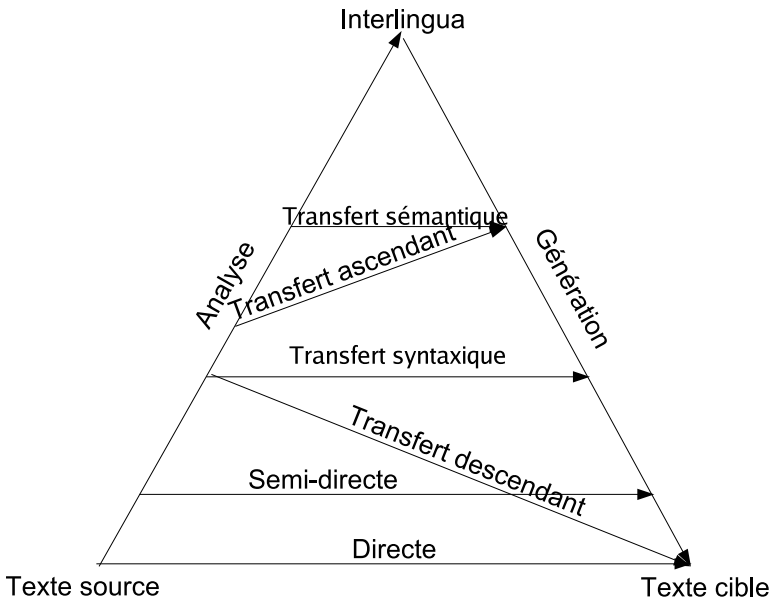


FIG. 1.1 – Triangle de Vauquois ([Vauquois and Boitet, 1985]). Représentation des différentes architectures linguistiques.

systèmes de « première génération » (russe-anglais et anglais-russe) des années 1950.

1.1.1.2 Systèmes à transfert

Les systèmes *à transfert* utilisent des représentations intermédiaires pour passer d'une phrase source à sa traduction, basées sur l'analyse syntaxique ou sémantique. Le transfert entre la langue source et cible peut être effectué au même niveau, en passant, par exemple, de l'analyse syntaxique source à l'analyse syntaxique cible. Il est possible d'effectuer un transfert de nature descendante ou ascendante, opérant à deux niveaux différents de représentation entre les deux langues. Le transfert descendant passe d'une structure source intermédiaire (eg. l'arbre syntaxique source) à un niveau moins abstrait : au niveau des unités morphologiques, ou directement au niveau des formes de surface. Le transfert ascendant, au contraire, passe à un niveau plus abstrait d'une représentation cible.

Le système de traduction SYSTRAN qui est un des systèmes les plus utilisés aujourd'hui est un exemple de système à transfert descendant.

1.1.1.3 Interlingua

L'approche de type Interlingua est basée sur une représentation abstraite (pivot), indépendante de la langue. Ainsi, le problème de traduction est vu comme la combinaison de deux problèmes : transformation d'une phrase source en une représentation pivot, et génération d'une phrase cible à partir de la représentation pivot. Chaque module est réutilisable pour la création d'un système de TA pour une nouvelle paire de langues, ce qui rend cette approche très attirante.
UNL [Uchida et al., 2005] est un exemple de langage pivot utilisé pour la construction d'un système de traduction de type Interlingua et dans les autres problèmes de TALN.

1.1.2 Architectures computationnelles

Le processus automatisé de traduction est en général un programme, qui produit une phrase cible à partir d'une phrase source.

La nature de ce type de processus automatique peut être *experte* ou *empirique*.

1.1.2.1 TA experte

Le processus automatisé expert est un programme créé par un expert :
– par une programmation directe dans un langage algorithmique classique ;
– dans un langage de haut niveau,
– dans un langage spécialisé pour la programmation linguistique d'automates,
– dans des formalismes de grammaires déclaratives.

Ce programme est souvent spécifique à une paire de langues. Ainsi, pour créer un système de traduction pour une nouvelle paire de langues, il est nécessaire de créer un nouveau programme (bien que dans certaines architectures la réutilisation de certains éléments génériques soit possible).

L'architecture linguistique à transfert a été introduite à l'origine dans le cadre de la TA experte. Aujourd'hui, de plus en plus de systèmes de TA probabiliste (ou hybride) se tournent vers ce type d'architecture (plutôt que vers l'architecture directe).

La majorité des systèmes à interlingua sont des systèmes de TA experte : les règles d'analyse et de génération en passant par le langage pivot sont définies par des experts.[1]

1.1.2.2 TA empirique

On distingue deux groupes principaux d'approches à la TA empirique : TA fondée sur les exemples (TAFE)[2] et TA probabiliste (SMT, Statistical Machine Translation).

Les méthodes empiriques se basent sur l'idée suivante : il existe un programme qui est capable d'apprendre à faire la traduction à partir de textes déjà traduits. Ce programme s'appelle un *module d'apprentissage*.

Les textes déjà traduits forment une base d'apprentissage. On appelle *corpus parallèle* un ensemble de couples de textes tels qu'un des textes est la traduction de l'autre. La nature des textes parallèles peut être différente et dépend de l'architecture du système de TA.

Le processus automatisé de traduction utilise soit le corpus parallèle prétraité (traduction fondée sur des exemples), soit un modèle de traduction extrait par un module d'apprentissage, afin de produire une nouvelle traduction. En théorie, les deux programmes, d'apprentissage et de traduction, sont indépendants de la paire de langues. Ainsi, il suffit de disposer d'un corpus d'une taille suffisante pour une nouvelle paire de langues pour construire un nouveau modèle de traduction.

Notons qu'en pratique il y a une certaine dépendance à la paire des langues. Ainsi, un prétraitement spécifique à la langue est souvent nécessaire : tokenisation spécifique à la langue, ou segmentation pour certaines langues (chinois ou japonais).

1.1.2.2.1 TA probabiliste. La TA probabiliste extrait un modèle probabiliste de traduction à partir d'une paire de textes où chaque phrase du premier texte est une traduction d'une phrase correspondante du deuxième (*corpus parallèle aligné*). Ce modèle est utilisé pour la génération d'une nouvelle traduction ; une fois que le modèle est appris, le corpus parallèle n'est plus réutilisé.

Le premier modèle de TA probabiliste a été inspiré par le succès des techniques stochastiques dans le domaine de la reconnaissance vocale, et a été proposé par IBM en 1988. Ce modèle a été testé sur les transcriptions des débats du parlement canadien (Canadian Hansards). Il n'utilisait aucune règle de nature linguistique, mais s'appuyait uniquement sur des statistiques extraites du corpus. Les résultats de ce premier système étant de qualité surprenante, ont incité nombreuses équipes de recherche à suivre cette direction.

Nous allons présenter différents modèles de TA probabiliste plus en détail dans la section 1.2.

[1] Mais il en existe qui sont empiriques comme celui de Gu and Gao [2004].
[2] Example-Based Machine Translation (EBMT)

1.1.2.2.2 TA par des exemples. Le trait le plus important distinguant la TAFE de la TA probabiliste est une exploitation du corpus parallèle au moment de l'exécution. L'idée de base de la TAFE (introduite par Nagao [1984]) peut être vue comme une application du principe de "raisonnement par cas" (*case-based reasoning*) où la solution du problème consiste à trouver un problème similaire résolu précédemment dans une base de problèmes résolus.

Ainsi, la TAFE se résume à retrouver des phrases similaires traduites précédemment dans le corpus parallèle aligné (par des techniques statistiques ou basées sur des règles) et (partie difficile) à extraire les parties cibles pour construire une phrase cible qui, on l'espère, est une traduction de la phrase source. Il existe des critères de similarité différents se basant sur des réseaux sémantiques, des thésaurus ou des informations statistiques.

La réutilisation du corpus parallèle à l'exécution permet à la TA par des exemples d'extraire de grands fragments de texte déjà traduits, ce qui lui donne un avantage par rapport à la plupart des systèmes de TA probabiliste qui sont limités aux fragments de textes extraits au moment de l'entraînement du système.

En TAFE par analogie de Lepage and Denoual [2005], il n'y a pas de décomposition en fragments explicite, ni d'étape de prétraitement ou d'alignement sous-phrastique du corpus des exemples. La traduction est faite en utilisant l'opération d'analogie proportionnelle. Si S est un segment à traduire, on cherche les "rectangles analogiques" $S_1 : S_2 :: S_3 : S$ (en langue source), tels qu'on dispose d'exemples de traduction T_1, T_2, T_3 pour chacun des fragments S_1, S_2, S_3. La traduction T du segment S est une solution de l'équation analogique (en langue cible) $T_1 : T_2 :: T_3 : T$.

L'exploitation de plus en plus importante des techniques probabilistes dans la TAFE, et la réexploitation de fragments de plus en plus longs par les modèles de la TA probabiliste rendent la frontière entre ces deux modèles de la TA empirique de plus en plus floue.

1.1.3 TA hybride

La création automatique d'un système de TA sans intervention humaine et la possibilité de réexploiter des textes déjà traduits sont des côtés attirants des méthodes empiriques. L'apprentissage à partir des traductions produites par des traducteurs humains permet d'acquérir une terminologie adaptée au domaine spécifique.

Par contre, la majorité des méthodes empiriques classiques sont basées sur l'architecture linguistique directe. Donc, ils sont souvent limités à des sous-fragments de corpus parallèles bruts et leur capacité de généralisation est faible.

La nécessité de généralisation devient plus évidente quand il s'agit de langues plus éloignées (anglais - tchèque, anglais - japonais) ou à morphologie plus riche (traduction de/vers l'allemand, le turc, le finnois, ou le hongrois). De plus, des corpus parallèles de taille importante n'existent pas pour de nombreuses paires de langues, et le problème de rareté des données devient plus important pour des corpus plus petits. La connaissance de la structure linguistique permet de compenser la rareté des données et d'introduire des règles plus génériques.

L'introduction des structures intermédiaires dans le processus de la traduction (et par conséquent le changement de l'architecture linguistique) donne plus de capacité de généralisation. La transformation vers les structures intermédiaires peut être faite soit par des méthodes empiriques, soit par des méthodes expertes. Les formalismes expressifs riches linguistiquement ont la capacité de décrire des correspondances complexes entre la langue source et la langue cible. Les règles introduites par des experts sont génériques et ne sont pas limitées au corpus d'apprentissage. Les règles génériques rendent l'approche experte plus robuste par rapport aux approches empiriques.

Les dictionnaires créés par des experts couvrent bien certaines paires de langues spécifiques et ont l'avantage par rapport aux dictionnaires créés empiriquement d'une connaissance morphologique inaccessible aux méthodes purement empiriques.

De l'autre côté, quand il s'agit d'un système de TA expert, l'adaptation au nouveau domaine peut être difficile et peut demander non seulement l'adaptation du dictionnaire, mais aussi la création de nouvelles règles pour l'analyse/transfert/génération. Certains formalismes peuvent être trop rigides et ne pas tenir compte des exceptions ou de l'évolution de la langue.

Les méthodes hybrides ont pour but de combiner l'approche experte et l'approche empirique, en passant éventuellement, par un changement d'architecture linguistique. Deux directions d'hybridation sont possibles :

- certains composantes du système de TA expert peuvent être remplacées par des composantes empiriques. Dans ce contexte, l'architecture linguistique initiale ne change pas, mais l'architecture computationelle change.
- les connaissances linguistiques expertes peuvent être introduites dans le modèle de TA empirique, ce qui peut changer l'architecture linguistique, et éventuellement l'architecture computationnelle.

Nous présentons quelques approches hybrides plus en détail dans la section 1.3.

1.2 TA probabiliste

L'idée générale de la TA probabiliste est d'apprendre un modèle de traduction à partir d'un corpus parallèle. Ce modèle est utilisé ensuite pour produire de nouvelles traductions. L'apprentissage du modèle a pour but d'établir les correspondances entre des mots ou des fragments source et cible. Le processus de traduction consiste à trouver la suite de fragments cible la plus probable selon le modèle appris précédemment.
Une vue d'ensemble des différents modèles de TA probabiliste est présentée dans [Lopez, 2008]. Dans cette section nous allons nous concentrer sur quelques modèles de base, en décrivant plus en détail les étapes les plus pertinentes par rapport aux travaux que nous avons effectués.

1.2.1 Premiers modèles de TA probabiliste : modèles lexicaux

Le premier modèle probabiliste pour la TA a été introduit par IBM [Brown and Della Pietra, 1993]. C'est un modèle génératif qui modélise la probabilité $P(e|f)$ à partir du corpus parallèle, où e est une phrase dans la langue cible, et f une phrase dans la langue source.

La traduction optimale $\hat{e}$ d'une phrase source f sera définie comme suit :

$$\hat{e} = \arg\max_{e} P(e|f) \tag{1.1}$$

En pratique, l'espace de recherche des phrases cible e est limité aux phrases qui peuvent être générées à partir d'une phrase source f, en appliquant une suite d'opérations d applicables à f. Le problème de trouver la traduction optimale $\hat{e}$ revient alors à trouver une paire $(\hat{e}, \hat{d})$, où $\hat{d}$ est un alignement[3] (ou plus généralement une suite de décisions ou opérations créée lors de la génération de la traduction) tel qu'en l'appliquant à une phrase d'entrée f, une traduction $\hat{e}$ est obtenue. Le problème de la traduction est alors reformulé comme suit :

$$\hat{e}, \hat{d} = \arg\max_{e,d} P(e, d|f) \tag{1.2}$$

La décomposition de $P(e, d|f)$ en appliquant le théorème de Bayes permet de reformuler le problème de traduction de la façon suivante :

$$\hat{e}, \hat{d} = \arg\max_{e,d} P(e, d|f) = \arg\max_{e,d} \frac{P(f, d|e)P(e)}{P(f)} = \arg\max_{e,d} P(f, d|e)P(e) \tag{1.3}$$

On parle du *modèle de canal bruité*[4]. L'intérêt d'une telle décomposition est le suivant : il est difficile d'apprendre directement un modèle $P(e, d|f)$ qui pourrait bien décrire les données, tandis que l'apprentissage des deux modèles indépendants $P(f, d|e)$ (*modèle de traduction*) et $P(e)$ (*modèle de langage de*

[3] Correspondance entre les mots source et cible.

[4] *noisy channel model* - terme venant du domaine de la reconnaissance de la parole.

la langue cible) donne un double contrôle sur la qualité de la traduction : les erreurs d'un des modèles peuvent être compensées par l'autre, ce qui rend le modèle final plus robuste.

1.2.1.1 Modèle de langage

La probabilité $P(e^I)$ de produire une phrase $e^I = e_1 \cdots e_I$ (e^I est une phrase de I mots : e_1, ..., e_I) dans la langue cible peut être décomposée :

$$P(e^I) = P(e_1 \cdots e_I) = \Pi_{i=1}^{I} P(e_i|e_1 ... e_{i-1}) \tag{1.4}$$

Afin de simplifier ce problème de modélisation, on fait l'hypothèse que le mot e_i de e^I n'est dépendant que des $n-1$ mots le précédant. Cette hypothèse permet de réécrire la probabilité $P(e^I)$ comme suit :

$$P(e^I) = P(e_1 \cdots e_I) = \Pi_{i=1}^{I} P(e_i|e_{i-(n-1)} ... e_{i-1}) \tag{1.5}$$

On appellera ce modèle de langage *modèle n-grammes* par la suite.

1.2.1.2 Modèle de traduction

Le modèle de traduction $P(f^J, d|e^I)$ modélise le processus de génération d'une phrase source $f^J = f_1 \cdots f_J$ à partir d'une phrase cible $e^I = e_1 \cdots e_I$ (suite à l'application du modèle de canal bruité). Le processus de génération de la phrase e^I à partir de la phrase f^J est divisé en plusieurs modèles plus simples.

Prenons l'exemple du modèle IBM 4 [Brown and Della Pietra, 1993]. Le modèle de génération de f^J à partir de $e^I = e_1 \cdots e_I$ par le modèle IBM 4 est décomposé en 3 modèles successifs :

- modèle du mot vide p_0 - la probabilité d'insérer le mot vide e_0 dans une phrase e^I ; le mot vide e_0 peut générer de faux mots f_j, ce qui revient à dire (en revenant au problème initial de la traduction de f^J) qu'un mot source f_j peut être "effacé" dans la cible (f_j est généré à partir de e_0).
- *le modèle de fertilité* $n(\phi_i|e_i)$ modélise le nombre ϕ_i des mots qu'il faut générer pour chaque mot e_i ; chaque mot e_i est dupliqué ϕ_i fois dans une phrase : $\underbrace{e_0..e_0}_{\phi_0} ... \underbrace{e_i..e_i}_{\phi_i} ... \underbrace{e_I..e_I}_{\phi_I}$.
- *le modèle de traduction lexicale* $t(f_{\tau_{j,k}}|e_i)$, $i = 0...I$ choisit le mot $f_{\tau_{j,k}}$ correspondant au doublon k du mot e_i ; chaque doublon de chaque mot cible produit un mot source unique.
- *le modèle de distorsion* $d(\delta_i|i)$ modélise la position δ_i du mot f_{δ_i} généré par le mot e_i, en autorisant ainsi un déplacement du mot e_i par rapport à sa position de départ ; à cette étape les mots f_j générés précédemment sont permutés suivant le modèle de distorsion.

La décomposition du modèle de traduction IBM-4 est alors décrite par l'équation suivante[5] :

$$\begin{aligned} P(f^J|e^I) \quad = \quad & \left(\prod_{i=1}^{I} n(\phi_i|e_i) \left(\prod_{k=1}^{\phi_i} t(f_{\tau_{i,k}}|e_i) \cdot d(\delta_{\tau_{i,k}}|i) \right) \right) \times \\ & \prod_{l=1}^{\phi_0} t(f_{\tau_{0,k}}|e_0) \times \binom{J-\phi_0}{\phi_0} p_0^{J-2\phi_0} (1-p_0)^{\phi_0} \frac{1}{\phi_0!} \end{aligned}$$

La première ligne de cette décomposition modélise le processus de transformation de chaque mot e_i de la phrase e, tandis que la deuxième ligne décrit les transformations du mot vide e_0.
Plus de détails et d'explications sur les modèles IBM se trouvent dans [Brown and Della Pietra, 1993].

1.2.1.3 Décodage

Le processus de traduction (transformation d'une phrase source en une phrase cible) est appelé *décodage* dans la communauté de TA probabiliste. Ce terme vient de la théorie des codes et est inspiré par

[5] Nous avons simplifié le modèle de distorsion afin de rendre la représentation plus claire.

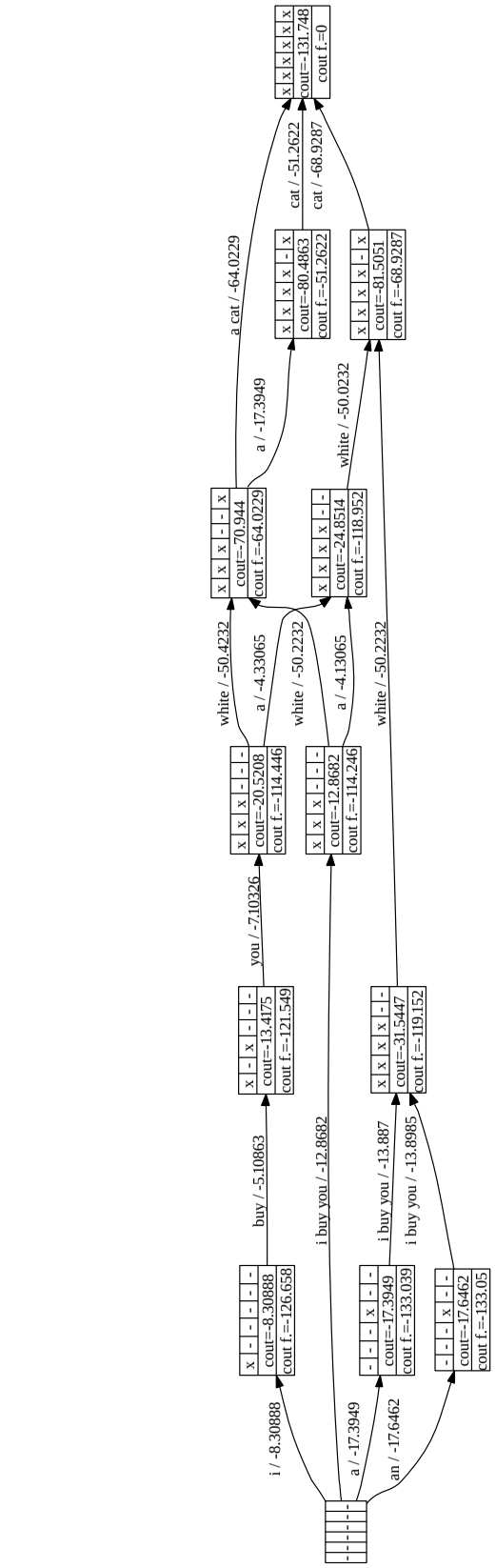


FIG. 1.2 – Exemple de graphe des hypothèses générant la traduction d'une phrase : *"je vous achète un chat blanc"*. Ce graphe a été généré par le système de TA probabiliste Moses qui sera présenté dans la section 1.2.4

l'idée de Warren Weaver, ancien cryptographe militaire, de considérer une phrase en russe comme une phrase anglaise chiffrée, et la tâche de traduction comme son décodage. Le *décodage* est la recherche d'une solution optimale de l'équation 1.3.

L'approche du décodage la plus courante en TA probabiliste ([Wang and Waibel, 1997]) est une généralisation de l'algorithme de décodage par piles (*stack decoding*) utilisé en reconnaissance vocale et introduit par [Jelinek, 1969].

Dans cet algorithme, l'espace de recherche est organisé en un graphe acyclique orienté (figure 1.2). Chaque nœud représente un état correspondant à une hypothèse de traduction (partielle ou complète). Un état est défini par un sous-ensemble de mots source couverts, les n derniers mots cible générés (pour le modèle de langage n-gramme), le coût de l'hypothèse selon le modèle utilisé ($P(e)P(f,d|e)$) et l'estimation du coût futur. Les nouveaux états du graphe sont générés en couvrant les mots source non couverts par l'état actuel et en rajoutant de nouveaux mots cibles générés par une hypothèse de traduction.

Le parcours de graphe avec les heuristiques du type A^* [Och et al., 2001] ou la conversion en problème de programmation linéaire en nombres entiers de [Germann et al., 2004] permettent de trouver une solution exacte de l'équation 1.3. Ce parcours, pourtant, devient très coûteux en temps pour des phrases longues. Des techniques d'*élagage* du graphe des hypothèses pour rendre la recherche plus efficace ont été proposées par [Tillmann and Ney, 2003] pour des modèles lexicaux.

1.2.2 Extension de la TA probabiliste lexicale

1.2.2.1 TA probabiliste à fragments

Les modèles lexicaux utilisent le mot comme unité de traduction. Bien que l'introduction du mot vide et du modèle de fertilité permette de traiter le problème des expressions à mots multiples, uniquement des traductions de type n–1[6] peuvent être traités par ces modèles.

En traduction réelle, les unités de traduction sont souvent des groupes des mots (contigus ou non). Prenons l'exemple du groupe anglais "*bank statement*", et sa traduction en français "*relevé de compte*". Bien qu'il ne soit pas impossible de générer cette traduction par un modèle IBM, ceci implique que les traductions *(bank, relevé)*, *(statement, compte)* soient choisies par le modèle de traduction lexicale, et que le mot "*de*" soit de fertilité 0. Il est quand même probable que beaucoup d'autres traductions des mots *bank* et *statement* plus fréquentes peuvent être proposée par le modèle de traduction lexicale. Quand ce groupe est traduit dans le contexte d'une phrase, la combinatoire des choix lexicaux et de leurs distorsions devient important, et les erreurs peuvent survenir facilement.

Le modèle de TA probabiliste à fragments permet de traiter ce problème, en choisissant un groupe de mots m–n comme unité de traduction. Les groupes de mots considérés peuvent être contigus ([Koehn et al., 2003], [Och and Ney, 2004]) ou non-contigus ([Simard et al., 2005], [Chiang, 2005]). La possibilité de générer une traduction en mettant en correspondance un fragment (groupe des mots) source avec un fragment cible remplace le modèle de fertilité et le mot vide introduits par des modèles lexicaux. Ainsi, le modèle de TA à fragments est réduit uniquement au tableau de traduction (bibliothèque des bi-fragments) et au modèle de distorsion ([Koehn et al., 2003]).

L'extraction d'une bibliothèque de bi-fragments est souvent faite à partir des alignements lexicaux qui sont à la base des modèles IBM en les recombinant avec des heuristiques ou autrement. Des exemples de bi-fragments extraits à partir d'une paire de phrases du corpus parallèle sont donnés à la figure 1.3.

Le passage des modèles lexicaux aux modèles à fragments augmente d'une façon importante la taille du graphe des hypothèses, car, en plus des traductions multiples pour chaque mot et des distorsions éventuelles, plusieurs couvertures pour une phrase source sont possibles. Ainsi, l'élagage du graphe des hypothèses est indispensable aux modèles de TA à fragments ([Och and Ney, 2004], [Koehn et al., 2003]).

[6] Grâce au modèle de fertilité, et à l'insertion du mot vide, n mots sources peuvent être traduits par 1 mot cible.

il	it	il est	it is
il est peu probable	it is unlikely	il est peu probable que	it is unlikely that
est	is	est peu probable	is unlikely
est peu probable que	is unlikely that	peu probable	unlikely
peu probable que	unlikely that	que	that
que la dernière tactique de moucharraf	that musharraf 's latest gambit	que la dernière tactique de moucharraf porte	that musharraf 's latest gambit will
la dernière tactique de	's latest gambit	la dernière tactique de moucharraf	musharraf 's latest gambit
la dernière tactique de moucharraf porte	musharraf 's latest gambit will	dernière	latest
dernière tactique	latest gambit	tactique	gambit
moucharraf	musharraf	porte	will
porte ses fruits , étant donné que	will succeed , as his	ses	his
ses fruits , étant donné que	succeed , as his	fruits	succeed
fruits ,	succeed ,	fruits , étant	succeed ,
fruits , étant donné	succeed ,	fruits , étant donné que	succeed , as
,	,	, étant	,
, étant donné	,	, étant donné que	, as
étant donné que	as	donné que	as
que	as	le	support
le soutien	support	le soutien public	popular support
le soutien public est	popular support is	le soutien public est au	popular support is at
le soutien public est au	popular support is at its	le soutien public est au plus bas	popular support is at its lowest ebb
soutien public	popular	public	popular
est	is	est au	is at
est au	is at its	est au plus bas	is at its lowest ebb
est au plus bas .	is at its lowest ebb .	au	at
au	at its	au plus bas	at its lowest ebb
au plus bas .	at its lowest ebb .	plus bas	lowest ebb
plus bas	its lowest ebb	plus bas .	lowest ebb .
plus bas .	its lowest ebb .	.	.

FIG. 1.3 – Bi-fragments extraits à partir du couple de phrases (normalisées) "*il est peu probable que la dernière tactique de moucharraf porte ses fruits , étant donné que le soutien public est au plus bas . – it is unlikely that musharraf 's latest gambit will succeed , as his popular support is at its lowest ebb .*"

1.2.2.2 TA probabiliste hiérarchique à fragments

Le modèle de TA probabiliste hiérarchique introduit par [Chiang, 2005] est une généralisation du modèle de grammaire synchrone hors-contexte (introduit par [Lewis and Stearns, 1966]). Ce modèle de grammaire est une extension des grammaires hors-contexte, permettant de générer une paire de chaînes (source et cible) plutôt qu'une chaîne unique. Voici un exemple d'une telle grammaire pour le français et l'anglais :

$$\begin{aligned} X &\rightarrow \langle \text{ chat , cat} \rangle \\ X &\rightarrow \langle \text{ Marie , Mary} \rangle \\ X &\rightarrow \langle X_1 \text{ de Marie , Mary's } X_1 \rangle \\ X &\rightarrow \langle X_1 \text{ de } X_2, X_2 \text{ 's } X_1 \rangle \end{aligned}$$

Chacune des deux chaînes générées par une règle de dérivation de cette grammaire peut contenir plusieurs symboles non terminaux, à condition que ce soient les mêmes non-terminaux entre la source et la cible.

La construction de cette grammaire est faite comme suit :
- tous les bi-fragments possibles sont extraits à partir du corpus parallèle ;
- les règles de grammaire sont dérivées en remplaçant des sous-fragments plus petits contenus à l'intérieur de bi-fragments plus grands ; par exemple, la paire des bi-fragments (*ne veut plus, do not want anymore*) et (*veut, want*) mène aux dérivations suivantes :

$$\begin{aligned} X &\rightarrow \langle \text{ veut , want } \rangle \\ X &\rightarrow \langle \text{ ne } X_1 \text{ plus , do not } X_1 \text{ anymore } \rangle \end{aligned}$$

Le problème de traduction dans ce cadre est équivalent au problème d'analyse d'une phrase source par une grammaire : une suite de dérivations qui génère une phrase source, génère sa traduction automatiquement. Les dérivations valides sont recherchées par un algorithme de type CYK, la meilleure est ensuite choisie avec un modèle log-linéaire (introduit ci-dessous).

1.2.3 Modèles log-linéaires de TA probabiliste

1.2.3.1 Modèle de canal bruité

Les premiers modèles de TA probabiliste étaient des modèles dits *génératifs*. Ainsi, dans le cas du modèle IBM4 (équation 1.6) les familles des paramètres sont définies par ses sous-modèles (fertilité $p(\phi_i|e_i)$, tableau de traduction $t(f_j|e_i)$, etc.).

La paramétrisation des modèles de type canal bruité est directement liée au processus de décodage : les paramètres optimisés sont des objets utilisés pour générer une nouvelle traduction. En même temps, pour pouvoir décrire le processus de décodage par de tels objets, des hypothèses d'indépendance forte doivent être faites. Ainsi, par exemple, dans les modèles à mots, la traduction d'un mot individuel de la phrase ne dépend, au mieux, que de ses n mots voisins (contrôlés par un modèle de langage et un modèle de distorsion).

1.2.3.2 Modèle log-linéaire

Le modèle log-linéaire propose de modéliser directement la probabilité $P(e, d|f)$ comme une combinaison linéaire de fonctions de traits $h_k(e, d, f)$ [Och and Ney, 2002].

$$P_\lambda(e, d|f) = \frac{\exp \sum_{k=1}^{K} \lambda_k h_k(e, d, f)}{\sum_{e', d'} \exp \sum_{k=1}^{K} \lambda_k h_k(e', d', f)} \tag{1.6}$$

Ce modèle contient K paramètres λ_k, qui déterminent la contribution de chaque trait $h_k(e, d, f)$ à la valeur finale de $P(e, d|f)$.

Au décodage, la meilleure traduction $(\hat{e}, \hat{d})$ est une solution de

$$\hat{e}, \hat{d} = \arg\max_{e,d}(\exp\sum_{k=1}^{K}\lambda_k h_k(e,d,f)) = \arg\max_{e,d}\sum_{k=1}^{K}\lambda_k h_k(e,d,f) \tag{1.7}$$

Notons que l'équation 1.3 correspond à un cas particulier de 1.6, où $h_1(e,f,d) = logP(e)$, $h_2(e,f,d) = logP(f,d|e)$ et $\lambda_1 = \lambda_2 = 1$.

Contrairement au modèle de canal bruité, il n'est pas nécessaire d'attribuer une probabilité unique à chaque élément des données (e,f), mais plusieurs probabilités peuvent être attribuées. Ainsi, les probabilités $P(e)$, $P(e|f)$, $P(f|e)$ peuvent être toutes intégrées dans le modèle log-linéaire en tant que traits. De plus, les traits du modèle log-linéaire peuvent être de natures différentes, et ne sont pas nécessairement des probabilités bien définies.

1.2.3.3 Entraînement à taux d'erreur minimal

La méthode d'optimisation des paramètres du modèle log-linéaire couramment utilisée dans la communauté de TA probabiliste est *la méthode de minimisation du taux d'erreur* (Minimum Error Rate Training, MERT) introduite par [Och, 2003]. Ce modèle cherche à minimiser l'erreur finale de traduction selon une mesure comme BLEU, NIST ou WER (plus sur les mesures d'évaluation dans la section 2.2.1).

Le but de cet entraînement est de minimiser l'erreur sur le corpus représentatif parallèle $(f_i, e_i)_{i=1..S}$. On introduit une fonction $E(t(f_i), e_i)$ mesurant l'erreur d'une traduction $t(f_i)$ par rapport à une traduction de référence e_i. Si $C^i = \{e^i_n\}_{\{n=1..N\}}$ est un ensemble de traductions possibles pour une phrase f_i, le problème d'optimisation des paramètres se résume à résoudre l'équation suivante[7] :

$$\hat{\lambda} = \arg\min_{\lambda}\sum_{i=1}^{S} E(\arg\max_{t\in C^i} p_\lambda(t|f_i), e_i) \tag{1.8}$$

Cette fonction implique le calcul de $\arg\max_t$ et n'est pas continue en λ. Deux méthodes proposées par Och dans [Och, 2003] permettent de calculer une solution approchée de l'équation 1.8.

La première solution consiste à trouver l'optimum d'une variante "lissée" de la fonction objective (par une méthode de descente du gradient) :

$$\hat{\lambda} = \arg\min_{\lambda}\sum_{i=1}^{S}\sum_{t\in C^i} E(t,e_i)\frac{p_\lambda(t|f_i)^\alpha}{\sum_{t_{n'}\in C^i} p_\lambda(t_{n'}|f_i)^\alpha} \tag{1.9}$$

L'autre solution est une variante de l'algorithme de Powell avec recherche linéaire (pour plus de détails se référer à [Och, 2003]). Och a montré empiriquement que les deux solutions mènent à des performances comparables sur un ensemble de test.

Ce type d'entraînement nécessite l'ensemble de toutes les traductions possibles C^i pour chaque phrase source f_i. Les systèmes de TA probabiliste peuvent produire une liste des N meilleures traductions pour une phrase. Pourtant, l'optimisation des paramètres λ uniquement sur cet ensemble des N meilleures traductions peut mener à une augmentation du taux d'erreur au décodage. Cela est dû au fait que, lors du décodage, le système peut avoir accès aux traductions qui ne se trouvent pas parmi les N meilleures traductions. Il est donc nécessaire d'avoir une liste plus complète des traductions pour chaque phrase, afin d'apprendre les paramètres λ minimisant le taux d'erreur quand ils sont intégrés dans le décodeur.

Afin de générer une telle liste de traductions, Och [2003] propose une procédure itérative d'optimisation des paramètres (algorithme 1). L'idée de cette procédure itérative est la suivante :

[7] Notons que dans le cas des scores NIST et BLEU, une fonction d'erreur globale sur le corpus parallèle $\mathbf{f}^S, \mathbf{e}^S$ (f^S et e^S sont l'ensemble des phrases source et cible respectivement), $E(t(\mathbf{f}^S), \mathbf{e}^S)$ ne se décompose pas en somme des erreurs individuelles pour chaque phrase. Il est toutefois possible d'adapter l'algorithme de recherche linéaire de Och au cas des scores BLEU et NIST. Une adaptation d'une version lissée d'une fonction objective dans le cas du score NIST a été proposée par Simard et al. [2005].

– d'abord la liste des N meilleures traductions est générée avec des paramètres initialisés (à des valeurs uniformes ou aléatoires) ;
– ensuite, à chaque itération, de nouveaux paramètres sont appris et la liste des traductions (l'ensemble d'entraînement) possibles est étendue avec de nouvelles traductions, jusqu'à ce que l'ensemble d'entraînement ne change plus (les nouveaux paramètres n'ajoutent pas de nouvelles traductions à la liste). La convergence de cette procédure est garantie, car à la fin C^i contiendra toutes les traductions possibles de f_i. D'après les résultats empiriques de Och [2003], en pratique la méthode converge au bout de 5-7 itérations.

Algorithme 1 Entraînement à taux d'erreur minimal (MERT)

ENTRÉES : Ensemble d'entraînement $(f_i, e_i)_{i=1..S}$
SORTIES : Paramètres λ
Initialiser λ^0 (aléatoires ou uniformes)
$T^0 = \text{N meilleures traductions}_{\lambda^0}(f_1, .., f_S)$
$l = 0$
tantque $T^l \neq T^{l-1}$ **faire**
$\lambda^{l+1} = \arg\min_\lambda \sum_{i=1}^{S} E(\arg\max_{t \in T^l} p_\lambda(t|f_i), e)$
// ajouter la liste des traductions obtenue avec des paramètres λ^{l+1} aux traductions obtenues précédemment
$T^{l+1} = T^l \cup \text{N meilleures traductions}_{\lambda^{l+1}}(f_1, .., f_S)$
$l = l + 1$
fin tantque
Retourner λ^l

1.2.4 Moses : système de TA probabiliste à fragments

Nous présentons le modèle de traduction à fragments sur l'exemple du système Moses. Le système de TA Moses[8] [Koehn et al., 2007] est considéré aujourd'hui comme un système de référence dans la communauté de la TA probabiliste, et se base sur un modèle log-linéaire combinant les traits suivants :

– *modèle de langage :* $h_{lm}(e) = log(p(e))$, où $p(e)$ est un modèle de langage n-gramme entraîné sur un grand corpus monolingue ;

– *deux modèles de traduction à fragments (direct et inverse) :*

$$h_{seg}(f, e, d) = \sum_{l=1}^{L} log(p(\mathbf{e}_l|\mathbf{f}_l)), h_{seg_{rev}}(f, e, d) = \sum_{l=1}^{L} log(p(\mathbf{f}_l|\mathbf{e}_l))$$

où $p(\mathbf{e}_l|\mathbf{f}_l)$ sont les probabilités de traduire le fragment (groupe de mots contigus) source $\mathbf{f}_l$ par le fragment cible $\mathbf{e}_l$ ($f = \mathbf{f}_1..\mathbf{f}_L$ - décomposition d'une phrase f en fragments, définie par d) estimées sur un grand corpus parallèle ;

– *deux modèles lexicaux de traduction (directe et inverse) :* $h_{lex}(f, e, a_{lex}) = \sum log(p(e_i|f_j))$ et $h_{lex_{rev}}(f, e, a_{lex}) = \sum log(p(f_j|e_i))$, où $p(e_i|f_j)$ correspond au tableau de traduction du modèle IBM (section 1.2.1), et où a_{lex} est la correspondance (alignement) entre les mots source et cible ;

– *pénalité des fragments* $h_{pf}(f, e, d) = \sum_{l=1}^{L} 1$: nombre des fragments utilisés pour générer la traduction : moins ce nombre est grand, plus les fragments étaient longs, et plus la traduction est fluide ;

– *pénalité des mots* h_{pm} : nombre des mots cible dans une traduction générée ; ce trait permet de pénaliser des traductions trop courtes, qui sont souvent préférées par le modèle de langage ;

[8] `http://www.statmt.org/moses`

– *distorsion* $h_d(f, e, d)$: ce trait pénalise ou favorise certains déplacements de fragments dans une traduction ; si les fragments source ne sont pas traduits dans l'ordre, la distance (en nombre des mots) entre un fragments traduit et un nouveau fragment mesure la distorsion de la traduction ;

– *distorsion lexicalisée* $h_{d_{lex}}$: déplacement d'un fragment cible peut être conditionné par ce même fragment cible, et/ou par le fragment source correspondant, les déplacements des fragments précédents et/ou suivants, et les types de déplacement possibles pour un fragment (monotone, échange, discontinu).

Certains traits correspondent aux sous-modèles IBM, qui sont entraînés sur un grand corpus parallèle. Les poids de ces sous-modèles (correspondant à la contribution de ces sous-modèles dans la valeur finale de $P(e, d|f)$) et les poids des autres fonctions de traits sont optimisés sur un ensemble d'entraînement disjoint de celui utilisé par les modèles génératifs[9] (appelé *corpus de développement* par la suite). L'optimisation des paramètres du modèle log-linéaire se fait par la variante de recherche linéaire proposée par Och [Och, 2003], minimisant l'erreur en terme du score BLEU (défini dans la section 2.2.1).

Le décodage se fait par un algorithme de recherche en faisceau (*beam search*) (Koehn et al. [2003]) avec la bibliothèque des bi-fragments extraits à partir d'un grand corpus parallèle.

Moses permet aussi l'entraînement et le décodage de modèles factoriels [Koehn and Hoang, 2007], représentant chaque mot (*token*) comme un vecteur de facteurs (forme de surface, lemme, partie du discours, etc.).

1.3 TA hybride dans la littérature

L'approche hybride est un sujet de recherche très actif ces dernières années et les types d'hybridation existants dans la littérature sont nombreux. Dans cette section, nous allons présenter les approches hybrides en TA les plus représentatives, en mettant plus l'accent sur celles qui rajoutent des connaissances linguistiques expertes aux systèmes de TA probabiliste.

1.3.1 Extension de la TA experte par des méthodes empiriques

Le succès des méthodes empiriques dans la traduction a incité de nombreuses équipes travaillant uniquement avec des méthodes expertes à intégrer des méthodes empiriques dans le processus de traduction (à différents niveaux). Ainsi, le système commercialisé de TA SYSTRAN [10] utilise des méthodes empiriques intégrées à des niveaux différents [Senellart et al., 2003; Surcin et al., 2007] :

– au niveau de l'analyse morphologique : chaque mot est enrichi par des traits morphologiques et syntaxiques, extraits à partir du dictionnaire, dont certains peuvent être ambigus ; un niveau de confiance (*probabilité*) est attribué à chaque trait et catégorie potentiels de ce mot ; l'analyse morphologique la plus probable correpond à une trajéctoire désambiguïsée dans le graphe produit.

– au niveau du transfert : utilisation de terminologie extraite à partir de corpus parallèle bilingue en même temps que de dictionnaires créés par des experts ; l'intégration des expressions à mots multiples extraites à partir du corpus parallèle a permis de gagner 1,5% de BLEU lors de la campagne d'évaluation WMT-07, les évaluation manuelles supplémentaires ont permis de prouver un gain de qualité considérable.

– au niveau de la génération : un nouveau modèle de génération a été construit, permettant d'ajouter la notion de *probabilité* ou *confiance* à l'exécution des fonctions ; cette notion initialement prévue pour être extraite à partir du corpus, peut être également être adaptée par un linguiste. D'après [Surcin et al., 2007], ce nouveau modèle n'est pas encore mis en production, étant en cours de test.

[9] Il s'agit du modèle de langage et du modèle de traduction définis dans le cadre du modèle de canal bruité.
[10] `http://www.systranet.fr/`

D'autres travaux ([Simard et al., 2007], [Dugast et al., 2007]) proposent d'apprendre un système de postédition des traductions produites par SYSTRAN, en tant que modèle de traduction probabiliste à fragments sur la base du corpus parallèle qui a les traductions produites par SYSTRAN du côté source, et les traductions de référence du côté cible. Ce type d'approche a montré des améliorations en termes de scores BLEU (1,5-3%) et de légères améliorations en termes de fluidité et d'adéquation pour certaines tâches lors de la campagne d'évaluation WMT-07.[11]

Dans les travaux de [Oepen et al., 2007], un reclassement des hypothèses a été intégré dans le système de TA experte à transfert sémantique LOGON (du norvégien vers l'anglais). Chaque composante du système (analyse, transfert, génération) générant les n meilleures hypothèses, la traduction optimale est choisie en reclassant la liste des traductions finales (combinaisons des hypothèses produites par chaque étape) avec des traits externes probabilistes tels que les probabilités lexicales, le modèle de langage etc. Un tel reclassement a permit de gagner 3% de BLEU par rapport au système de base qui prend toujours la première hypothèse à chaque étape.

Dans [Sánchez-Martínez, 2008] l'auteur remplace certaines composantes du système de TA experte Apertium (système à transfert superficiel ou semi-direct) par des composantes empiriques (basées sur le corpus) : désambiguïsation lexicale basée sur le corpus parallèle et extraction automatique de règles de transfert à partir du corpus parallèle. Cela lui a permis de gagner 4% en taux d'erreur des mots pour la traduction de l'espagnol vers le catalan, et 1% pour la traduction de l'espagnol vers le portugais.

1.3.2 Exemples de systèmes hybrides

Le système MATADOR (Habash [2003]) vise la traduction d'une langue "peu dotée" (pauvre en ressources et outils informatisés) vers une langue "bien dotée". L'idée de base est de produire l'analyse de dépendances d'une phrase source par des règles expertes et d'appliquer un transfert descendant hybride pour générer la traduction finale. La première étape du transfert est une traduction des unités lexicales source en sacs de mots cible. La composante de "génération"[12] génère ensuite une traduction finale à partir des dépendances en source et des sacs de mots cible. Le système MATADOR s'est montré plus robuste par rapport au modèle IBM4 (entraîné sur un corpus de UN, 50 000 phrases parallèles) pour la traduction de l'espagnol vers l'anglais d'un texte hors-domaine (La Bible).

Un autre exemple de système de TA hybride est le système Stat-XFER [Lavie, 2008] de CMU. La traduction est faite en 3 étapes :

1. un treillis des analyses morphologiques possibles pour chaque mot est produit ;

2. toutes les règles de transfert (lexical et structurel) basées sur une grammaire hors-contexte synchrone sont appliquées aux différents niveaux pour créer le treillis des sous-traductions possibles. Contrairement aux systèmes de TA à fragments hiérarchiques (section 1.2.2.2), ces règles structurelles peuvent être soit extraites à partir du corpus des analyses syntaxiques parallèles[13] (anglais-français), soit créées par des experts (hébreu-anglais). Les règles de transfert lexical sont apprises à partir du corpus parallèle bilingue ou extraites à partir du dictionnaire.

3. l'étape de génération consiste à générer une traduction finale à partir du treillis généré par l'étape de transfert ; l'étape de génération (appelée *décodage*) consiste à choisir une séquence linéaire (sans chevauchement) des sous-traductions avec un modèle log-linéaire combinant l'ensemble des traits attribués à l'hypothèse de traduction.

[Lavie, 2008] montre que l'introduction d'un grammaire de transfert même très simple (36 règles) permet d'obtenir des améliorations de 4% de BLEU pour la traduction de l'hébreu vers l'anglais.

[11] Malheureusement, en contexte d'exploitation, cette approche n'a donné aucune amélioration significative, contrairement à l'introduction des probabilités au niveau morphologique (J. Senellart via Ch. Boitet)

[12] Utilisant des ressources expertes pour la langue cible et des méthodes empiriques basées sur un grand corpus monolingue cible.

[13] Dans ce contexte il s'agit des corpus parallèle alignés au niveau des mots.

1.3.3 Intégration de connaissances linguistiques sans changement d'architecture linguistique en TA probabiliste

1.3.3.1 Préparation des données

Beaucoup de travaux essayent de prétraiter le corpus parallèle à l'aide d'analyseurs linguistiques experts pour pallier la rareté des données et améliorer la qualité du modèle de traduction empirique appris à partir de ces données. Ainsi, [Ma et al., 2008] apprennent les alignements (et le modèle probabiliste de TA à fragments) à partir d'une bibliothèque d'arbres de dépendances, en introduisant des traits syntaxiques pour étendre le modèle d'alignement de base. Ils montrent une diminution relative de 5,5% d'AER[14] pour l'alignement d'un corpus parallèle chinois-anglais, ce qui résulte en une amélioration de 1% de BLEU pour une tâche de traduction.

[Crego and Habash, 2008] utilisent un analyseur en constituants du côté source et cible, pour éliminer les alignements erronés non cohérents avec la correspondance entre les constituants. Ce type de prétraitement mène à des améliorations de 1,5-2,5 de BLEU pour une traduction de l'arabe vers l'anglais.

1.3.3.2 Reclassement des traductions

Le reclassement des traductions ([Och et al., 2003; Shen, 2004]) est un moyen simple et rapide d'intégration des connaissances linguistiques dans un système de TA probabiliste. L'idée est de prendre la liste des N meilleures traduction produites par le système de référence, et à l'aide de fonctions de traits basées sur des informations linguistiques source et/ou cible, d'établir un nouvel ordre (des traductions) sur cette liste. Les traits externes (linguistiquement motivés ou non) permettent de contrôler les aspects non accessibles au modèle de base.

Les travaux de [Och et al., 2003] et [Shen, 2004] ont trouvé des améliorations de 1,3 en BLEU (de 31,6 à 32,9), tandis que les estimations d'une borne supérieure sur une liste de 1000 meilleures traductions était de 35,7 de BLEU[15].

1.3.4 Changement de l'architecture linguistique d'un système de TA probabiliste

Dans cette section, nous allons décrire des systèmes se basant sur le modèle de TA probabiliste, qui intègrent des connaissances linguistiques (expertes) dans le processus de décodage, ce qui implique un changement de l'architecture linguistique. Nous allons classer ces approches par niveau des connaissances linguistiques intégrées (morphologiques, syntaxiques), et par le type de transfert (ascendant, descendant, horizontal, autre).

Le transfert descendant nécessite l'analyse de la phrase source et génère une traduction cible à partir de cette analyse sans passer par des structures intermédiaires cible. Le transfert ascendant génère une structure intermédiaire cible directement à partir d'une chaîne source. Le transfert horizontal nécessite les étapes d'analyse, transfert d'une structure source vers une structure cible, et la génération d'une traduction finale à partir d'une structure intermédiaire cible.

1.3.4.1 Niveau morphologique

1.3.4.1.1 Transfert descendant. Quand il s'agit de traduction vers une langue plus riche en morphologie flexionnelle (c'est souvent le cas de traduction de l'anglais vers une autre langue), la génération des formes fléchies est une tâche difficile, car la langue d'origine n'a pas d'informations suffisantes pour générer une forme correcte. Selon [Ueffing and Ney, 2003], si on ne tient pas compte de la forme fléchie du mot, mais uniquement de son lemme, le taux d'erreur est réduit de 10% pour une tâche de traduction de l'anglais vers l'espagnol.

[14] Alignment Error Rate.

[15] La borne supérieure était choisie en prenant la meilleure traduction en terme des scores BLEU pour chaque phrase parmi les 1000 meilleures traductions proposées par le système de TA probabiliste.

Les solutions proposées pour affronter ce problème sont généralement basées sur la transformation de la phrase source en ajoutant des informations manquantes[16] : [Ueffing and Ney, 2003] propose de transformer les verbes précédés par un nom propre et les formes interrogatives d'une phrase source anglaise pour la rendre plus proche de l'espagnol. Différents prétraitements basés sur l'analyse morphologique d'une phrase source anglaise ont été proposés pour la traduction vers le danois [Stymne, 2009] ou le suédois [Stymne, 2008].

Bien que la traduction d'une langue morphologiquement riche vers l'anglais soit une tâche plus simple que la traduction inverse, les nombreuses formes fléchies du côté source causent une grande rareté des données au niveau des formes, ce qui résulte en des modèles de traduction erronés, et en des mots (formes) inconnus au moment du décodage.

Les solutions à ce problème proposées dans la littérature sont généralement spécifiques à la langue source : [Sadat and Habash, 2006] proposent différents schémas de segmentation des clitiques pour l'arabe, qui mènent à plus de 1,7% d'amélioration du score BLEU pour des traductions de l'arabe vers l'anglais. [Goldwater and McClosky, 2005] appliquent différents types de lemmatisation du côté source et montrent des améliorations de traduction du tchèque vers l'anglais de 6% (de 27% à 33%) en BLEU. [Koehn, 2003; Niessen and Ney, 2000] montrent que la décomposition des mots composés améliore la traduction de l'allemand vers l'anglais.

1.3.4.1.2 Transfert horizontal. Une approche tenant compte des analyses morphologiques source et cible est proposée par [Toutanova et al., 2008], qui introduit un modèle log-linéaire pour la génération des formes fléchies indépendamment du modèle de traduction. Trois méthodes différentes sont proposées pour l'intégration de ce modèle avec le modèle de traduction, toutes montrant des améliorations par rapport au système de TA à fragments et au système de TA *à treelets* [Quirk et al., 2005] pour la traduction vers l'arabe et le russe. Chacun de ces modèles est représenté par un ensemble de traits dépendant d'une analyse de dépendances source, et des traits morphologiques cible (partie de discours, lemme, formes fléchies possibles, etc.). La génération d'une phrase cible est faite en choisissant une bonne forme fléchie pour chaque mot par biais du modèle log-linéaire combinant ces traits.

1.3.4.1.3 Autres architectures. Le modèle factoriel de traduction est le cadre général d'exploration des informations morphologiques proposé par [Koehn and Hoang, 2007]. Ce modèle consiste à représenter chaque mot par un vecteur de facteurs morphologiques et à intégrer cette représentation dans le cadre d'un système de TA probabiliste à fragments, permettant ainsi au système d'avoir accès aux informations morphologiques au moment du décodage. Les facteurs source et/ou cible peuvent être exploités dans ce cadre. Le type de transfert dépend alors de quel côté les facteurs sont utilisés. La génération d'une phrase cible à partir de sa représentation factorielle est faite uniquement par des techniques empiriques, bien que l'analyse de la phrase source puisse être faite par un analyseur morphologique expert.

1.3.4.2 Niveau syntaxique

1.3.4.2.1 Transfert descendant. La motivation du travail de [Avramidis and Koehn, 2008] est initialement de rajouter des informations manquantes pour faire la traduction d'une langue morphologiquement pauvre vers une langue riche. L'idée de cette approche, similaire à celle de [Ueffing and Ney, 2003], est d'explorer les analyses syntaxiques source afin de rajouter les informations manquantes aux noms et verbes d'une phrase source anglaise, permettant d'établir la notion de cas (pour les noms) et personne (pour les verbes) lors de la traduction vers le tchèque et le grec. Cette approche a permit de réduire le taux d'erreur de conjugaison des verbes de 19% à 5,4%, et de choix d'une forme nominale fléchie de 9% à 6%.

1.3.4.2.2 Transfert ascendant. Dans [Yamada and Knight, 2001], les auteurs proposent un modèle de traduction basé sur les sous-modèles du modèle d'IBM-4 (section 1.2.1), en les appliquant à l'arbre de constituants source et non pas à la chaîne des mots source. Dans [Yamada and Knight, 2002], ce modèle

[16] Cette idée a été déjà mise en œuvre depuis les années 1980 dans une quinzaine de systèmes de TA japonais-anglais commercialisés au Japon (Ch. Boitet) [Sakamoto et al., 1986].

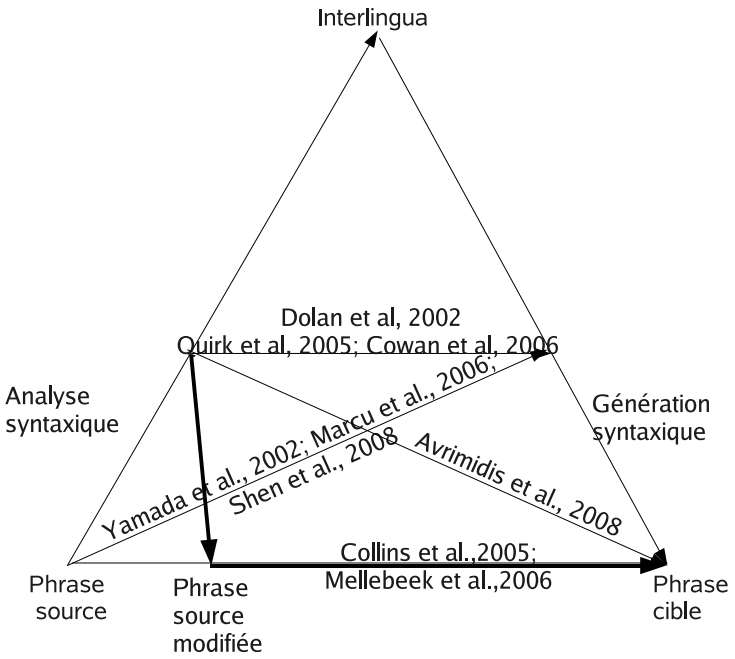


FIG. 1.4 – Les architectures linguistiques des différents méthodes d'hybridation proposées dans la littérature, intervenant au niveau syntaxique.

de traduction est appliqué dans le cadre du canal bruité[17] à la traduction du chinois vers l'anglais. Ainsi, le décodeur cherche à construire l'arbre des constituants cible (anglais) le plus probable à partir d'une chaîne source selon le modèle appris sur la base du corpus parallèle (chaîne source – arbre des constituants cibles).

Les travaux plus récents de [Marcu et al., 2006] vont dans la même direction, générant l'arbre cible en passant par des fragments (groupes des mots) source (et non plus par des mots uniques source comme dans [Yamada and Knight, 2002, 2001]). Ce modèle basé sur le modèle log-linéaire est plus proche des modèles à fragments de TA probabiliste, utilisant des objets de type *fragment-vers-arbre* au décodage. Ce modèle fut l'un des premiers modèles hybrides (changeant l'architecture linguistique) améliorant les scores BLEU par rapport aux systèmes de TA probabiliste à fragments, mais il a aussi montré des améliorations en termes d'évaluation humaine (chinois-anglais).

Dans [Shen et al., 2008], les auteurs explorent un modèle de type *chaîne-vers-dépendances*, qui génère une structure de dépendances cible (par opposition à la structure de constituants de [Yamada and Knight, 2002, 2001] et [Marcu et al., 2006]) à partir d'une phrase source. Ce modèle permet aussi d'introduire un modèle de langage à dépendances et d'améliorer des scores automatiques BLEU (de 1,48) et TER (de 2,53) (chinois-anglais) par rapport au modèle de TA probabiliste hiérarchique.

1.3.4.2.3 Transfert horizontal. Un exemple du transfert horizontal est le système de [Dolan et al., 2002] à Microsoft Research. Au moment de l'entraînement les analyseurs syntaxiques sont appliqués aux parties source et cible du corpus. Ces analyseurs transforment les phrases source et cible en représentations en formes logiques (LF), qui sont alignés ensuite. Les exemples de traductions (transformation de la

[17] Rappelons que le modèle de canal bruité transforme le modèle de la traduction comme ce qui suit : $P(e|f) = P(e)P(f|e)$ (f, e - phrases source, cible). Par conséquent, l'application du modèle introduit par [Yamada and Knight, 2001] dans ce cadre implique l'inversion des langues source et cible. Il s'agit, en fait, d'une analyse syntaxique cible dans ce contexte.

LF source en LF cible) sont extraits à partir de ce corpus. Au moment du décodage, la forme logique de la phrase source est obtenue par analyse, cette forme logique est ensuire transformée en une forme logique cible, à l'aide des exemples de traduction appris précédemment.

La traduction par des treelets de [Quirk et al., 2005] a poursuivi les travaux de [Dolan et al., 2002] mais n'utilise qu'analyseur syntaxique expert du côté source. L'apprentissage du modèle est fait à partir d'un corpus parallèle aligné au niveau des mots (avec GIZA++ [Och and Ney, 2003]), où l'arbre de dépendances source est aligné avec la chaîne des mots cible. Les dépendances source sont projetées ensuite sur la phrase cible et des objets du type *arbre-vers-arbre* (appelé *treelets*) sont extraits. Ces objets sont utilisés ensuite au moment de décodage, générant l'arbre de dépendances cible à partir de l'arbre de dépendances source sous le contrôle d'un modèle log-linéaire. La génération de la traduction finale à partir de l'arbre de dépendances se fait sous le contrôle du modèle d'ordonnancement (contrôlant l'ordre des mots dans la phrase cible).

[Cowan et al., 2006] est un exemple d'approche utilisant les analyses syntaxiques bilingues de type arbre à arbre. Dans ce papier, les auteurs proposent le modèle suivant : une phrase source (allemande) est analysée par l'analyseur syntaxique de Collins ([Collins, 2003] analyse empirique). L'étape de transfert consiste à appliquer une suite de transformations simples (dont l'ordre est prédéfini et peut être spécifique à une paire de langues) afin de générer une structure cible (anglaise) de forme spéciale (AEP - *aligned extended projections* [Frank, 2002]). Le modèle de transfert (de l'arbre d'analyse syntaxique source vers structure AEP cible) est un modèle linéaire combinant les fonctions de traits dépendant des décisions prises à chaque étape. Les paramètres de ce modèle sont appris sur la base d'un corpus parallèle avec l'algorithme du Perceptron Structuré. Pour des raisons de simplicité, ce modèle est utilisé uniquement pour la traduction des propositions (principales et subordonnées). Les évaluations du modèle (automatiques et manuelles) n'ont pas montré d'amélioration par rapport à Pharaoh ([Koehn, 2004], modèle probabiliste à fragments).

1.3.4.2.4 Autres architectures. Dans [Collins et al., 2005], la phrase source est analysée par un analyseur syntaxique et des règles de réarrangement sont appliqués à l'analyse source. Ensuite, la nouvelle phrase obtenue après réarrangement est traduite par un système de TA probabiliste à fragments. Ce modèle a montré des améliorations de 25,2% à 26,8% en BLEU et en jugements humains pour une traduction de l'allemand vers l'anglais (Europarl).

[Mellebeek et al., 2006] introduit des règles de décomposition d'une phrase source (en se basant sur l'analyse en constituants) en fragments plus simples. Les fragments sont traduits avec le système de TA probabiliste à fragments et la traduction finale est générée à partir de ces sous-traductions. Dans cette approche, les auteurs proposent une procédure de décomposition d'une phrase source basée seulement sur l'analyse source en constituants, et donc, dépendant seulement de la langue source. Cette approche a montré des améliorations en termes de scores automatiques (3,3% de BLEU) et d'évaluation manuelles pour la traduction de l'anglais vers l'espagnol. Nous nous inspirons de cette idée dans la partie III.

Les deux travaux cités ici ne rentrent pas dans le cadre d'architectures linguistiques classiques. Les deux approches utilisent l'analyse syntaxique source afin de transformer une phrase source (pour la rendre plus proche de la phrase cible, ou pour la simplifier). Il ne s'agit toutefois pas d'un transfert descendant, car le transfert est fait au niveau des formes de surface, et aucune structure linguistique intermédiaire n'intervient lors du décodage. Cette nouvelle architecture correspond au trajet en gras sur la figure 1.4.

Chapitre 2

Contexte et approches envisagées

2.1 Ressources disponibles au sein de XRCE

Notre direction de recherche a été inspirée par les ressources disponibles au sein de XRCE : l'analyseur syntaxique robuste XIP et le système de TA probabiliste Matrax. Dans cette section, nous allons présenter ces ressources avant de passer à la description des directions de recherche suivies dans la section suivante.

2.1.1 Ressources linguistiques expertes : XIP

L'analyseur syntaxique robuste XIP (*Xerox Incremental Parsing*) [Aït-Mokhtar et al., 2002] est un outil de traitement linguistique développé au sein de l'équipe Parsing & Semantics de XRCE. Il permet l'analyse syntaxique robuste de textes généraux. Pour un analyseur syntaxique, la robustesse correspond à la capacité de générer des analyses utiles, au moins partiellement correctes et utilisables, pour des textes réels. Cette robustesse sera surtout importante quand il s'agira d'analyses des sorties du système de TA probabiliste, qui ne sont pas nécessairement des phrases bien formées syntaxiquement (ce contexte d'utilisation de XIP est décrit dans la section 3.1).

XIP est un formalisme de règles adapté à l'analyse syntaxique superficielle et profonde. Il permet de traiter la désambiguïsation des parties de discours, la reconnaissances des entités nommées, le passage des constituants aux dépendances.

XIP n'est pas l'implémentation d'une théorie linguistique particulière. C'est un outil basé sur un LSPL (langage spécialisé pour la programmation linguistique)[1]. Il fonctionne de manière incrémentale en appliquant une séquence de traitements à une phrase d'entrée. L'application de cette séquence de traitements produit un arbre de *chunks* et une liste de dépendances. Les traitements appliqués sont :

Tokenization (segmentation en mots) et *analyse morphologique,* faites par des transducteurs d'états finis ;

Désambiguïsation des parties du discours (PoS) faite par des règles de désambiguïsation ou par des HMM (modèles de Markov cachés) ;

Chunking : une suite de règles (créées par des experts) organisée en couches (successives) sert à regrouper les séquences de catégories en *chunks* (structures syntaxiques minimales non ambiguës et non récursives).

Analyse des dépendances : une suite de règles est appliquée pour produire les relations entre les mots et/ou *chunks*. Les relations de dépendance peuvent être soit des dépendances fonctionnelles et syntaxiques superficielles, soit des dépendances profondes.

À la sortie de cette analyse, on dispose d'informations syntaxiques superficielles et profondes qui peuvent être exploitées par d'autres modules (pour l'analyse plus profonde). L'analyse de constituants proposé par XIP est toujours cohérente, mais elle peut ne pas être complète (il s'agit, en réalité, d'une

[1] Comme les outils construits par le GETA pour la TAO (traduction assistée par ordinateur) [Boitet, 1993].

liste d'arbres de *chunks*, qui sont des structures minimales non récursives). La structure de dépendances peut ne pas être cohérente, et ne pas être complète.

Nous avions à notre disposition les grammaires de XIP pour les langues suivantes : français, anglais, allemand, espagnol.

Les figures 2.1 et 2.2 donnent des exemples de sorties produites par XIP. Chaque mot et chaque *chunk* ont un certain nombre d'attributs syntaxiques (figure 2.3).

Arbre des *chunks* **:**

```
TOP{SC{NP{Ce matin} PUNCT{,} NP{j'} FV{ai}} NP{acheté} PP{des
NP{livres}} SENT{.}}
```

Liste des dépendances :

```
SUBJ(ai,j')
OBJ(ai,acheté)
VMOD(ai,matin)
NMOD_POSIT1(acheté,livres)
PREPOBJ(livres,des)
DETERM_DEM(matin,Ce)
```

FIG. 2.1 – Exemple d'analyse produite par XIP pour la phrase *Ce matin, j'ai acheté des livres.*

Arbre des *chunks* **:**

```
TOP{SC{NP{No one} FV{VERB{doubts NP{the NOUN{European NOUN{Central
Bank}} 's independence }}}} SENT.}
```

Liste des dépendances :

```
MOD_PRE(independence,European Central Bank)
DETD(European Central Bank,the)
SUBJ_PRE(doubts,No one)
OBJ_POST(doubts,independence)
MAIN(doubts)
PARTICLE(European Central Bank,'s)
HEAD(Bank, European Central Bank)
HEAD(independence,the European Central Bank 's independence)
ORGANISATION(European Central Bank)
```

FIG. 2.2 – Exemple d'analyse produite par XIP pour la phrase *No one doubts the European Central Bank's independence.*

```
NP{PRON{No one^no_one^+0+6+Pron+NomObl+3P+Sg+PRON :}
ORGANISATION(<NOUN :European^European^+18+26+NAdj+Sg+Country+NADJ,
Central^central^+27+34+NAdj+ModLoc+s_sc_pto+s_pto_adj+Sg+NADJ,
Bank^Bank^+35+39+Prop+orgHead+Sg+NOUN>)
```

FIG. 2.3 – Exemple de traits syntaxiques attribués lors de l'analyse.

2.1.2 Traduction automatique à XRCE : Matrax

Matrax (Simard et al. [2005]) est un système de TA probabiliste à fragments. La différence principale entre Matrax et les autres systèmes de TA probabiliste à fragments est sa capacité d'exploiter non seulement des bi-fragments connexes, mais aussi des bi-fragments à trous.

2.1.2.1 Extraction des bi-fragments à trous.

La construction empirique d'une bibliothèque de bi-fragments à trous est faite via une factorisation matricielle des correspondances lexicales entre phrases source et cible (Goutte et al. [2004]). Le but de cette factorisation est de trouver des objets pivot (*cepts*) via lesquels les mots source et cible sont alignés. Chaque *cept* définit les bi-fragments qui peuvent être extraits à partir d'une paire de phrases du corpus parallèle. Autrement dit, les cepts définissent des classes, et les mots source et cible doivent être affectés à une de ces classes. Ce type d'approche permet d'extraire des alignements de type m–n, sans contrainte de contiguïté (contrairement à l'approche de Och and Ney [2004] et Koehn et al. [2003]).

Les bi-fragments dans ce modèle sont définis par un *cept* ou par la combinaison de plusieurs *cepts*. Chaque bi-fragment se voit attribuer un vecteur de traits, dont les valeurs permettent de choisir ou non ce bi-fragment au moment du décodage. Ces traits correspondent aux probabilités conditionnelles $p(\mathbf{e_l}|\mathbf{f_l})$ et $p(\mathbf{f_l}|\mathbf{e_l})$ de traduire un fragment source $\mathbf{f_l}$ par un fragment cible $\mathbf{e_l}$, et inversement. Ces probabilités ont tendance à être surestimées pour des bi-fragments rares. Pour cette raison, les probabilités de correspondance lexicale sont également prises en compte, ce qui permet de contrôler la qualité d'alignement lexical entre les deux fragments. Ces traits sont calculés comme suit :

$$p_{lex}(\mathbf{f_l}|\mathbf{e_l}) = \frac{\sum_{\mathbf{f_i}\in\mathbf{f_l},\mathbf{e_j}\in\mathbf{e_l}} \mathbf{p}(\mathbf{f_i}|\mathbf{e_j})}{|\mathbf{e_l}|^{|\mathbf{f_l}|}}$$

La valeur de ce trait correspond à une moyenne de tous les alignement lexicaux possibles ($f_i - e_j$) à l'intérieur du bi-fragment ($\mathbf{f_l} - \mathbf{e_l}$). Les probabilités lexicales inverses $p_{lex}(\mathbf{e_l}|\mathbf{f_l})$ sont également prises en compte.

Une fois que tous les bi-fragments possibles sont extraits, différents filtrages sont appliqués, afin de réduire la taille de la bibliothèque des bi-fragments. Parmi les critères de filtrage figurent le nombre d'occurrences d'un bi-fragment dans le corpus d'entraînement, le nombre des trous source et cible (les bi-fragments contenant plus de 4 trous sont généralement abandonnés, car considérés comme peu utiles et peu fiables), et la taille des bi-fragments (les bi-fragments trop longs peuvent rarement être réutilisés).

La figure 2.4 donne une idée des bi-fragments à trous extraits à partir d'une paire de phrases.

2.1.2.2 Modèle de traduction.

Le processus de traduction est similaire à celui décrit dans la section 1.2.1.3. L'utilisation des bi-fragments à trous impose des contraintes supplémentaires sur la construction du graphe des hypothèses : lorsque tous les mots d'une phrase source sont couverts, la phrase cible correspondante ne peut contenir aucun trou.

Le modèle de traduction utilisé par Matrax est un modèle log-linéaire, similaire à celui de Moses présenté dans la section 1.2.4. Matrax utilise les mêmes traits que Moses, à l'exception des traits suivants :

- Matrax n'utilise pas de distorsion lexicalisée ;
- Matrax contient une fonction de trait supplémentaire pour contrôler le nombre de trous dans les bi-fragments : h_{gc}.

L'optimisation des paramètres du modèle de traduction est faite en maximisant le score NIST global lissé (Simard et al. [2005]) par une méthode de descente du gradient, en itérant l'optimisation, et cela en ajoutant de nouvelles traductions à chaque itération afin de générer un ensemble d'entraînement plus complet.

’s	de
,	,
, _ ’s	, _ _ de
, _ ’s authorities	, _ autorités de
, _ _ authorities	, _ autorités
.	.
authorities	autorités
authorities have not	autorités _ _ _ n’ ont pas
far	présent
have	ont
have _ _ to	ont _ _ à
have not	n’ ont pas
not	ne _ pas
not reacted _ the tharwa project .	n’ _ pas réagi _ le projet tharwa .
project	projet
reacted	réagi
reacted _ _ tharwa project	réagi _ _ projet tharwa
reacted _ the	réagi _ le
reacted to	réagi à
reacted to the tharwa project	réagi à le projet tharwa
so	jusqu’à
so far	jusqu’à présent
so far , syria ’s authorities have not reacted to the tharwa project .	jusqu’à présent , _ autorités de la syrie n’ ont pas réagi à le projet tharwa .
syria	la syrie
syria ’s authorities have not	autorités de la syrie n’ ont pas
syria _ authorities	autorités _ la syrie
syria _ authorities have	autorités _ la syrie _ ont
tharwa	tharwa
tharwa _ .	tharwa .
tharwa project	projet tharwa
the	le
the _ _ .	le _ _ .
the _ project	le projet
the tharwa project	le projet tharwa
the tharwa project .	le projet tharwa .
to	à

FIG. 2.4 – Exemples de bi-fragments à trous extraits par Matrax pour une paire de phrases du corpus parallèle : "*so far , syria ’s authorities have not reacted to the tharwa project . – jusqu’à présent , les autorités de la syrie n’ ont pas réagi à le projet tharwa* ."

2.2 Base expérimentale

2.2.1 Mesures d'évaluation automatique

Le but d'une mesure automatique est de mesurer la "qualité" de la traduction produite par un système de TA sans intervention de juges humains. La notion de qualité est à la base du problème d'évaluation de la traduction automatique. Cette notion est souvent liée à la tâche : simple compréhension ou traduction de haute qualité.

Beaucoup de mesures automatiques définissent la qualité de la traduction comme une mesure de similarité entre la traduction évaluée et la (les) traduction(s) de référence produite(s) par un traducteur professionnel. Cette approche est souvent critiquée (voir la sous-section 2.2.1.2). Malgré de nombreuses critiques, ces mesures sont souvent utilisées par la communauté de la TA (surtout de la TA probabiliste), et permettent de comparer facilement des instances du même système ou de systèmes différents entraînés sur les mêmes données.

Nous présentons d'abord quelques mesures d'évaluation automatique et nous discutons ensuite des avantages et des critiques des mesures automatiques.

2.2.1.1 Quelques mesures automatiques

La liste des mesures automatiques présentées ici est loin d'être exhaustive et a pour but de présenter les mesures d'évaluation automatique les plus utilisées, et surtout, les plus utilisées dans le développement des systèmes de TA probabiliste.

WER. *WER* (Word Error Rate, taux d'erreur de mots) est la première mesure automatique proposée pour l'évaluation des sorties de TA. L'idée de cette mesure empruntée au domaine de la reconnaissance vocale est de comparer la suite des mots reconnus avec la phrase de référence. Le taux d'erreur est dérivé d'une distance de Levenshtein et est défini par l'équation 2.1, où N est le nombre de mots dans la phrase de référence, et $D/I/S$ sont le nombre minimum d'opérations d'édition nécessaire pour transformer la suite des mots reconnus en la phrase de référence (D- suppression, I - insertion, S- substitution).

$$WER = \frac{S + I + D}{N} \tag{2.1}$$

Cette mesure n'est en fait pas adaptée au domaine de la TA, car, contrairement au domaine de la reconnaissance vocale, où une seule réponse correcte existe, il existe le plus souvent un très grand nombre de traductions correctes différentes, et très dispersées au sens de la distance de Levenshtein utilisée. Des variantes différentes de cette mesures ont été proposées, permettant de tenir compte de ce fait, telles que : taux d'erreur indépendant de la position (PER) [Tillmann et al., 1997], taux d'erreur à références multiples (*multi-reference WER*) [Nießen et al., 2000]. Mais aucune de ces variantes ne peut améliorer la qualité de la mesure de façon significative, car le nombre de références reste beaucoup trop petit.

BLEU. La mesure *BLEU* (BiLingual Evaluation Understudy)[Papineni et al., 2002] (une des premières mesures introduites après WER) est considérée aujourd'hui comme une mesure standard d'évaluation automatique de la TA dans la plupart des campagnes d'évaluation[2].

L'idée de *BLEU* est de comparer la traduction avec la référence non seulement mot à mot (comme les mesures du type WER), mais de tenir compte également des suites de N mots en commun (N-grammes) entre la sortie du système de TA et la traduction de référence. La comparaison de la traduction de référence avec une traduction évaluée est faite au niveau d'un ensemble de traductions, et non au niveau de

[2]Notons tout de même que les campagnes d'évaluations n'utilisent des mesures automatiques qu'à titre indicatif. Les évaluations manuelles servent de base principale aux jugements sur les traductions produites par les divers systèmes.

chaque phrase.[3]

La mesure *BLEU* est proportionnelle à la *précision modifiée* p_n définie par l'équation 2.2, où C est l'ensemble des phrases du corpus de traductions évalué, c une phrase individuelle de l'ensemble des traductions, $ngram$ une séquence de n mots, $Count(ngram)$ le nombre d'occurrences de $ngram$ dans la traduction c, $Cooc(ngram)$ le nombre maximal d'occurrences, dans lesquelles $ngram$ est partagé entre la traduction c et sa référence.

$$p_n = \frac{\sum_{c \in C} \sum_{ngram \in c} Cooc(ngram)}{\sum_{c' \in C} \sum_{ngram' \in c'} Count(ngram)} \tag{2.2}$$

Afin de pénaliser les traductions trop courtes, un facteur de pénalisation BP (brevity penalty) est introduit (L_c, L_r sont le nombre des mots dans l'ensemble des traductions évaluées et dans les traductions de référence) :

$$BP = \begin{cases} 1 & \text{, si } L_c > L_r \\ e^{(1-L_r/L_c)} & \text{, si } L_c \leq L_r \end{cases} \tag{2.3}$$

L'expression générale pour *BLEU* est définie par l'équation 2.4, où le paramètre w_n donne la possibilité de pondérer des n-grammes de tailles différentes. Dans la version standard de *BLEU*, $w_n = 1/N$, et $N = 4$.

$$BLEU = BP \cdot exp\big(\sum_{n=1}^{N} w_n \cdot \log p_n\big) \tag{2.4}$$

La mesure *BLEU* est conçue pour évaluer la traduction uniquement au niveau du corpus, les scores prédits par *BLEU* au niveau des traductions individuelles peuvent être très loin des jugements humains. De plus, un intérêt de *BLEU* est dans l'exploration de références multiples, ce qui a pour but de prendre en compte des styles et des structures grammaticales différents. Malheureusement, il existe très peu de corpus à références multiples, et l'évaluation avec *BLEU* est souvent faite en utilisant une référence unique, ce qui donne souvent une corrélation faible avec les jugements humains.[4]

NIST. La mesure *NIST* [Doddington, 2002] est basée sur une idée proche de celle de BLEU, avec certaines modifications. Ainsi, plutôt que d'attribuer les mêmes poids à tous les n-grammes dans le décompte des précisions, *NIST* pondère les n-grammes par leur "informativité"[5] *Info*$(w_1 \cdots w_n)$:.

$$\textit{Info}(w_1 \cdots w_n) = \log_2\big(\frac{Count(w_1 \cdots w_{n-1})}{Count(w_1 \cdots w_n)}\big) \tag{2.5}$$

La formulation finale de *NIST* est définie par la formule 2.6.

$$NIST = BP_{nist} \sum_{n=1}^{N} \frac{\sum_{c \in C} \sum_{ngram \in c} Info(ngram) \cdot Cooc(ngram)}{\sum_{c \in C} \sum_{ngram \in c} Count(ngram)} \tag{2.6}$$

$$BP_{nist} = exp\big(\beta \log_2 min(L_c/L_r, 1)\big) \tag{2.7}$$

A cause de la pondération des n-grammes par leur "informativité", la valeur de *NIST* n'a pas de borne supérieure (BLEU se place toujours dans l'intervalle $[0, 1]$). De plus, la valeur de *NIST* varie avec la taille du corpus évalué. Aujourd'hui, *NIST* a tendance à disparaître des campagnes d'évaluation.

[3] Cette incapacité à noter les traductions au niveau des segments est un défaut majeur de ce type de mesure.

[4] En réalité, même avec des références multiples, la corrélation de BLEU peut être faible, car l'ensemble des références n'est jamais assez dense dans l'ensemble des traductions correctes possibles. D'autre part, utiliser des références multiples peut dégrader intrinsèquement la validité de BLEU, comme l'ont montré Callison-Burch and Osborne [2006] car cela mène à juger bonne une mauvaise traduction qui contient 2 n-grammes incompatibles mais communs à 2 références différentes.

[5] L'idée de ce facteur est de privilégier les n-grammes "rares", et donc plus informatifs, aux n-grammes fréquents. Ainsi, le bi-gramme *of the* est beaucoup plus fréquent (moins informatif), que le bi-gramme *the house*.

wpF, wpBLEU. Les mesures *wpF* et *wpBLEU* [Popovic and Ney, 2009] sont basées sur les décomptes des suites de formes de surface et des parties de discours en commun entre la traduction proposée par le système et la traduction de référence.

wpF - c'est une F-mesure basée sur des n-grammes prenant en compte des formes de surface et des parties du discours.

On a classiquement :

$$\begin{aligned}
\text{F-mesure} &= \frac{2 \cdot \text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}} \\
\text{Précision} &= \frac{\sum_{c \in C}(\sum_{ng_w \in c} Cooc(ng_w) + \sum_{ng_p \in c} Cooc(ng_p))}{\sum_{c \in C}(\sum_{ng_w \in c} Count(ng_w) + \sum_{ng_p \in c} Count(ng_p))} \\
\text{Rappel} &= \frac{\sum_{r \in R}(\sum_{ng_w \in r} Cooc(ng_w) + \sum_{ng_p \in r} Cooc(ng_p))}{\sum_{r \in R}(\sum_{ng_w \in r} Count(ng_w) + \sum_{ng_p \in r} Count(ng_p))}
\end{aligned}$$

wpBLEU est une combinaison du score BLEU classique (sur les formes de surface) avec le score BLEU sur les partie du discours.

Les deux mesures ont montré une meilleure corrélation (par rapport aux mesures NIST et BLEU) au niveau des traductions individuelles avec les jugements humains au cours de la campagne d'évaluation WMT-2009 ([Callison-Burch et al., 2009]).

Autres mesures.
METEOR [Banerjee and Lavie, 2005] - moyenne harmonique entre la précision et le rappel des unigrammes, tenant compte de la synonymie, et des variantes morphologiques des mots.
TER, HTER [Snover et al., 2006] - *TER*[6] est défini comme le nombre minimal d'opérations d'édition nécessaires pour transformer la traduction évaluée en une traduction de référence. *HTER*[7] compte le nombre d'opérations d'édition nécessaires pour transformer une hypothèse de traduction en une bonne traduction (pas nécessairement une des traductions de référence).

2.2.1.2 Critique des mesures automatiques

Bien que BLEU reste toujours une mesure officielle dans la plupart des campagnes d'évaluation, elle est de plus en plus critiquée pour sa faible corrélation avec les jugements humains [Callison-Burch and Osborne, 2006; Blanchon and Boitet, 2007]. En effet, la corrélation de BLEU avec les jugements humains est assez forte pour des traductions mauvaises, mais cette même corrélation devient très faible quand la qualité des traductions augmente.

Ainsi, cette mesure (comme les autres mesures automatiques de même nature) n'est pas capable de tenir compte des différences subtiles des structures syntaxiques.

Une des plus grandes critiques de cette mesure est qu'elle ne permet pas de comparer deux systèmes de natures différentes. BLEU (surtout utilisant une référence unique) pénalise sévèrement les traductions proposées par un système de TA expert ce qui est lié aux choix lexicaux différents de ceux utilisés dans le corpus (et donc utilisés par les systèmes de TA probabiliste entraînés sur ce corpus), à la formulation différente, etc. Les traductions de TA probabiliste sont privilégiées car ces systèmes ont été entraînés sur un corpus de même origine que les traductions de référence utilisées par BLEU.
De plus, il arrive souvent que les traductions produites par les systèmes experts ont de meilleurs scores donnés par des juges humains, car ce sont souvent des phrases grammaticalement bien formées, et car les

[6] Translation Edit Rate.
[7] Human-targeted Translation Edit Rate.

choix lexicaux faits par le système expert sont tout à fait acceptables du point de vue de la langue cible [Callison-Burch et al., 2009].

2.2.1.3 Utilité des mesures automatiques

Malgré de nombreuses critiques, les mesures automatiques s'avèrent très utiles au cours du développement d'un système de TA pour comparer rapidement (sans intervention humaine) deux versions successives de ce système. Un système de TA probabiliste contient un grand nombre de paramètres, qui peuvent avoir plus ou moins d'influence sur la qualité de la traduction en fonction du corpus, du domaine, du couple de langues etc. Ainsi, l'optimisation de ces paramètres en faisant l'évaluation manuelle à chaque instant semble peu réaliste. L'utilisation des scores automatiques, en revanche, permet d'évaluer les états intermédiaires facilement, et ce type d'optimisation des paramètres donne en pratique de bons résultats.

2.2.2 Corpus

L'apprentissage d'un système de TA probabiliste demande un corpus parallèle de taille suffisante pour entraîner le modèle de traduction, et un corpus monolingue pour l'entraînement du modèle de langage.

2.2.2.1 News Commentary

Dans la majorité de nos expériences, nous avons utilisé le corpus de News Commentary fourni au cours de la campagne d'évaluation WMT-08[8]. Les caractéristiques de ce corpus sont présentées dans le tableau 2.1.

La partie parallèle (les paires de langues : en-fr, en-es, en-de, en-cz) du corpus a été créée dans le cadre du projet européen EuroMatrix[9] à partir d'éditoriaux en ligne[10]. Ce corpus parallèle contient 1 à 2 millions de mots par langue. Il a été augmenté ensuite par les deux ensembles de test (2000-3000 phrases) créés pendant les campagnes d'évaluation WMT-(07,08,09). Ces ensembles de test sont des traductions d'articles d'actualité (depuis août 2007) provenant de sites Web différents (tchèques, hongrois, français, anglais, allemands, italiens et espagnols). Les traductions ont été faites par des traducteurs professionnels.[11]

Dans nos travaux, nous cherchons à montrer que l'introduction de structures syntaxiques est importante surtout dans le cas où peu de textes parallèles sont disponibles. C'est une des raisons pour lesquelles nous avons utilisé le corpus de News Commentary dans la plupart des nos expériences.

La partition du corpus de News Commentary en des ensembles d'entraînement (modèle de traduction), de développement (apprentissage des paramètres) et de test est représentée dans le tableau 2.2. Le corpus *dev2006* correspond au corpus de développement utilisé pour l'optimisation des paramètres qui était proposé lors de la campagne d'évaluation WMT2007 (en domaine). Les corpus *test2007* et *test2006* sont des corpus de test (hors domaine) proposés lors des campagnes WMT-(06,07). Nous avons utilisé le corpus *test2006* pour évaluer nos modèles et *test2007* pour des entraînements supplémentaires.

2.2.2.2 Europarl

Pour certaines de nos expériences, nous avons également utilisé le corpus Europarl [Koehn, 2005], construit à partir des procès verbaux des réunions du parlement européen, traduits par des professionnels et disponibles dans 11 langues européennes. Le tableau 2.1 montre que ce corpus est d'une taille beaucoup plus importante que News Commentary. Nous l'avons utilisé, dans nos expériences, afin d'enrichir la couverture lexicale de News Commentary, pour réduire le nombre des mots non traduits dans le corpus

[8] `http://statmt.org/wmt08/shared-task.html`
[9] `http://www.euromatrix.net/`
[10] `http://www.project-syndicate.org/`
[11] Ces données sont considérées *hors domaine* car elles proviennent de sources différentes de celles du corpus lui-même.

TAB. 2.1 – Caractéristiques du corpus : corpus parallèle bilingue

Europarl

	en-fr		**en-de**		**en-es**	
phrases	1 428 799		1 418 115		1 411 589	
nb mots	40 067 498	44 692 992	37 431 872	39 516 645	41 042 070	40 067 498
nb mots uniques	107 733	129 166	104 269	320 180	108 116	154 971

News Commentary

	en-fr		**en-de**		**en-es**	
phrases	64 223		82 740		74 512	
nb mots	1 560 274	1 831 149	1 977 200	2 051 369	1 799 312	2 052 186
nb mots uniques	38 821	46 056	43 383	92 313	41 592	56 578

TAB. 2.2 – Les sous-corpus du corpus de News Commentary (anglais, français).

corpus	*train*		*dev2006*		*test2007*		*test2006*	
phrases	55 420		1057		2007		1064	
nb mots	1 177 636	1 358 640	22 646	26 391	43 777	50 848	22 521	26 907

de test, et pour faciliter la tâche d'évaluation manuelle.

2.2.3 Protocole d'évaluation

2.2.3.1 Modèle de référence (baseline)

La plupart de nos expériences sont basées sur le système de TA à fragments Moses (section 1.2.4), entraîné dans les conditions décrites ci-dessous. Cela nous permet d'obtenir des résultats qui peuvent être directement comparés aux résultats trouvés dans la littérature, car Moses est considéré aujourd'hui comme une base de référence forte par la communauté de la TA probabiliste. Cela nous permet également d'utiliser certains des accesoires disponibles dans Moses (et non disponibles dans Matrax), qui permettent de simplifier la procédure d'expérimentation[12]. Par la suite, sauf si le contraire est précisé, nous nous référons à ce modèle en parlant du *modèle de base* ou *baseline*.

Le corpus d'entraînement utilisé pour l'extraction d'une bibliothèque de bi-fragments est le corpus de News Commentary (tableau 2.2). Le corpus *train* est utilisé pour l'entraînement du modèle de traduction. La partie cible de ce corpus, jointe avec la partie cible du corpus Europarl, est utilisée pour l'entraînement du modèle de langage. Les paramètres du modèle sont optimisés sur le corpus *dev2006*, et les résultats que nous présentons sont calculés sur le corpus *test2006*.

Nous avons enrichi la bibliothèque des bi-fragments extraits du corpus News Commentary avec des bi-fragments extraits d'Europarl afin d'améliorer la couverture lexicale. Pour ne pas privilégier les traductions spécifiques au corpus Europarl et non adaptées à la traduction de News Commentary, nous n'ajoutons pas tous les bi-fragments extraits à partir de Europarl, mais uniquement ceux qui augmentent la couverture de la bibliothèque des bi-fragments de News Commentary : si un fragment source contient des mots non observés dans le corpus de News Commentary, il est ajouté à la bibliothèque des bi-fragments[13].

[12] Notons que toutes les expériences présentées dans la suite de ce mémoire peuvent directement s'appliquer à n'importe quel système de TA probabiliste à fragments. Ainsi, nous avons appliqué quelques modèles de reclassement parmi ceux présentés dans la section 5.1 aux traductions de Matrax [Nikoulina and Dymetman, 2008]. Les résultats que nous avons obtenus étaient similaires à ceux présentés dans cette thèse.

[13] À notre connaissance, ce type d'enrichissement de la bibliothèque de bi-fragments n'a pas encore été décrit dans la littérature.

2.2.3.2 Évaluation automatique

Afin d'évaluer l'impact des modèles que nous allons proposer, nous avons besoin de définir le protocole d'évaluation. Les modèles proposés dans cette thèse sont des variantes ou des extensions du modèle de base. Afin de comparer nos résultats avec ceux trouvés dans la littérature, nous utiliserons des mesures automatiques standard telles que NIST, BLEU. Nous utilisons aussi une variante de BLEU, wplBLEU, introduite plus loin (section 4.3.2) pour comparer les nouveaux modèles proposés par rapport au modèle de base.

Certaines différences entre les deux traductions peuvent être assez subtiles et ne seront pas visibles au niveau des scores automatiques. Ainsi, des changements mineurs, tels qu'un accord, par exemple, pourront être ignorés par BLEU ou NIST, si cela ne mène pas à une traduction de référence, car ces mesures ne prennent nullement en compte la formation linguistique des phrases. Nous évaluerons donc les traductions manuellement afin d'étudier l'impact réel des informations linguistiques.

2.2.3.3 Évaluation manuelle

Dans notre contexte, l'évaluation manuelle des traductions vise à étudier l'impact de l'introduction d'informations linguistiques sur la "qualité" de la traduction produite par un système de TA à fragments. Définissons la notion de qualité dans notre contexte.

Généralement, on introduit des connaissances syntaxiques dans le modèle à fragments pour améliorer la structure de la traduction afin de la rendre plus compréhensible. Nous utilisons pour cela des critères d'évaluation standard, tels que l'adéquation et la fluidité de la traduction. Ainsi, quand nous parlons d'amélioration de la qualité de la traduction dans la suite, nous parlons d'une meilleure formation (de point de vue de la syntaxe), qui rend la traduction plus adéquate et/ou plus fluide (au moins partiellement).

Le scénario que nous adoptons est le suivant. Le juge doit comparer deux traductions, dont une est la traduction produite par le modèle de base (*baseline*), et l'autre est une traduction produite par le système que nous cherchons à évaluer.

Les critères que le juge doit considérer en comparant ces deux traductions peuvent être différents : soit on compare la fluidité des deux traductions, soit leur adéquation. Ainsi, pour chaque paire de traductions, un juge humain doit dire :

- si la traduction t_1 est plus fluide ou moins fluide que la traduction t_2 (sans voir la phrase source) ;
- si la traduction t_1 est plus adéquate ou moins adéquate que la traduction t_2 (en voyant la phrase source).

Le juge peut juger les deux traductions comme étant de qualité équivalente. Le juge ne connaît pas la provenance de chaque traduction.

2.2.4 Évaluation du système de référence

Les scores automatiques du modèle entraîné sur News Commentary et du modèle enrichi avec Europarl sont présentés dans le tableau 2.3. Dans ce tableau, nous voyons que le lexique supplémentaire extrait de Europarl nous a permis de réduire le nombre des mots non traduits.

TAB. 2.3 – Évaluation automatique du système de référence. NC : performance du système entraîné uniquement sur le corpus de News Commentary ; NC + Europarl : système enrichi avec le lexique bilingue du corpus Europarl. NTW (non translated words) - nombre des mots non traduits.

mesures / données	fr-en			en-fr		
	NIST	BLEU	NTW	NIST	BLEU	NTW
NC training	7,1824	0,2754	745	6,9113	0,2716	510
NC + Europarl training	7,1614	0,2742	670	6,9918	0,2712	377

Regardons quelques exemples de traductions proposées par le modèle de base (annexe A, page 115) et analysons des erreurs typiques qu'on peut espérer corriger à l'aide de connaissances linguistiques.

Observations générales :
- les mots non traduits sont plus fréquents pour la traduction du français vers l'anglais. La raison principale en est la plus grande richesse de la morphologie flexionnelle du français : en l'absence d'analyse morphologique, le système considère chaque forme du même mot comme un nouveau *token*.
- les problèmes d'accord sont courants pour une traduction de l'anglais vers le français (annexe A, phrases 12, 25, 49). La phrase source anglaise ayant une morphologie plus pauvre ne contient pas suffisamment d'informations (à ce niveau d'interprétation) pour générer les bonnes formes fléchies.
- l'ordre des mots dans une traduction est aussi un problème fréquent : le modèle de langage choisit les n-grammes les plus fréquents, ce qui peut générer une traduction fluide, mais pas adéquate (en-fr : 11, 25, 54 ; fr-en : 25).
- de nombreux mots-outils (articles, prépositions) manquent ou sont mal placés dans les traductions vers le français : cela a un grand impact sur l'adéquation et sur la fluidité.
- le modèle a tendance à omettre certains mot-clés lors de la traduction vers l'anglais (phrases 5, 32, 42), ce qui change le sens de la traduction (mais probablement a un impact positif sur les scores BLEU).

La plupart des erreurs typiques entre l'anglais et le français se produiront aussi dans le cas de la traduction entre l'anglais et une autre langue ayant une morphologie flexionnelle riche. Le problème de l'ordre des mots persiste dans le cas de la traduction entre deux langues ayant un ordre des mots différents.

2.3 Approches envisagées

Bien que l'analyse des erreurs montre que l'analyse morphologique est un problème important pour une traduction de et/ou vers une langue riche en morphologie flexionnelle, cela ne sera pas notre préoccupation principale. De nombreuses solutions ont été proposées à ce problème (nous en avons présenté quelques-unes dans la section 1.3.4.1). Nous avons choisi de nous concentrer sur le problème d'intégration des connaissances syntaxiques dans le modèle de TA probabiliste à fragments. Ayant l'analyseur syntaxique XIP à notre disposition, notre motivation principale a été d'introduire des connaissances sur la structure source et cible dans le processus de traduction, afin d'améliorer l'adéquation de la traduction.

2.3.1 Approche initiale : traduction par des graphelets

Au début de nos travaux, nous avons envisagé de construire un nouveau modèle de traduction basé sur des objets que nous appelons *graphelets*. L'idée de cette approche est d'apprendre le modèle de traduction sur la base des structures de dépendances produites par XIP pour les parties source et cible d'un corpus parallèle. Nous décrivons ici quelques idées et les avancées réalisées avec cette approche, ainsi que les problèmes rencontrés.

2.3.1.1 Présentation de l'approche

L'idée de cette première approche est similaire à celle de la traduction par des *treelets* de Quirk et al. [2005], en utilisant des analyses de dépendances du côté source et cible.[14] La projection des dépendances source sur la partie cible peut être erronée, soit suite aux erreurs d'alignement, soit à cause de la divergence lexicale. Dans le cas d'un corpus bilingue de taille relativement faible, GIZA++ a tendance à faire plus d'erreurs d'alignement. Dans ce cas de figure, l'exploitation des structures linguistiques bilingues semble être plus utile car elle permet à la fois d'induire des alignements de meilleure qualité et d'apprendre des objets structurés qui servent à produire une nouvelle traduction.

Les structures de dépendance opèrent au niveau des mots et ont tendance à être plus proches entre les deux langues que leurs structures de constituants [Fox, 2002]. De plus, nous espérons que, grâce à

[14] Notre approche suit donc celle proposée par Dolan et al. [2002].

l'utilisation d'un analyseur du même type du côté source et cible, les structures produites seront uniformes et que les incohérences syntaxiques seront minimales (ce seront seulement celles causées par des divergences entre les deux langues et non pas celles résultant de différences des formalismes). Cela nous donne des raisons de croire que l'utilisation de tels objets structurés bilingues aura plus d'impact positif sur la qualité de la traduction que les essais précédents (comme ceux de Cowan et al. [2006]).

Formalisation. Nous allons d'abord définir formellement les structures que nous allons utiliser dans notre modèle. XIP produit une liste de dépendances étiquetées entre les mots et/ou les *chunks*. Un tel graphe de dépendances ne forme pas nécessairement une structure arborescente.

Pour simplifier notre modèle, nous allons considérer uniquement les dépendances binaires lexicales (relations entre des mots, et non pas entre des *chunks*) proposées par XIP ; ces dépendances forment un *graphe de dépendances orienté*, dont les nœuds correspondent aux mots d'une phrase.

Définition. Un *bi-graphe* B est un triplet (G_S, G_T, A_{st}) où G_S et G_T sont des graphes de dépendances , et A_{ST} une matrice d'alignement qui met en correspondance les mots des phrases source et cible.

La figure 2.5 est une illustration du bi-graphe (l'analyse XIP de la phrase source est donnée sur la figure 2.1). La figure 2.6 est la matrice d'alignement correspondante.

Définition. Un *graphelet* g est un sous-graphe du graphe de dépendances $G = (V, E)$.

Définition. Un *bi-graphelet* b du bi-graphe $B = (G_S, G_T, A_{st})$ est un triplet (g_s, g_t, a_{st}), où g_s (g_t) est un sous-graphe de G_S (G_T) et a_{st} est une matrice d'alignement entre les nœuds de g_s ou g_t.

Quelques exemples de bi-graphelets (ainsi que de bi-fragments correspondants) extraits à partir du bi-graphe de la figure 2.5 sont donnés dans la figure 2.7.

2.3.1.2 Apprentissage

Par analogie avec les modèles de TA probabiliste, nous extrayons notre bibliothèque de bi-graphelets en suivant les étapes principales décrites ci-dessous.

2.3.1.2.1 Alignement lexical structuré. La plupart des modèles de TA probabiliste sont basés sur les alignements lexicaux, quand il s'agit de TA à fragments, ou de TA utilisant des objets plus complexes (*arbre-vers-arbre*, *chaîne-vers-arbre*, etc.). Les alignements lexicaux produits uniquement par GIZA++ ne tiennent pas compte des structures linguistiques source et cible ; [Ma et al., 2008; Fossum et al., 2008] ont montré que les connaissances sur des structures syntaxiques bilingues permettent de réduire certains types d'erreur d'alignement.

Afin de tenir compte des structures linguistiques des phrases source et cible, nous construirons une matrice d'alignement $A = \{a_{i,j}\}$, $a_{i,j} \in \{0, 1\}$ minimisant *l'énergie d'alignement* entre les deux graphes, qui est définie par l'équation 2.8 (BinLio and Hancock [2001]) :

$$E(S, T) = -\frac{1}{2} Tr[S^t A T A^t] \tag{2.8}$$

S et T sont les matrices d'adjacence du graphe de dépendances source et du graphe de dépendances cible respectivement.[15] Autrement dit, nous cherchons un alignement entre les deux phrases qui rende leurs structures de dépendances plus proches.

[15] Pour une matrice X de taille $n \times n$, $Tr(X) = \sum_{i=1}^{n} x_{ii}$.

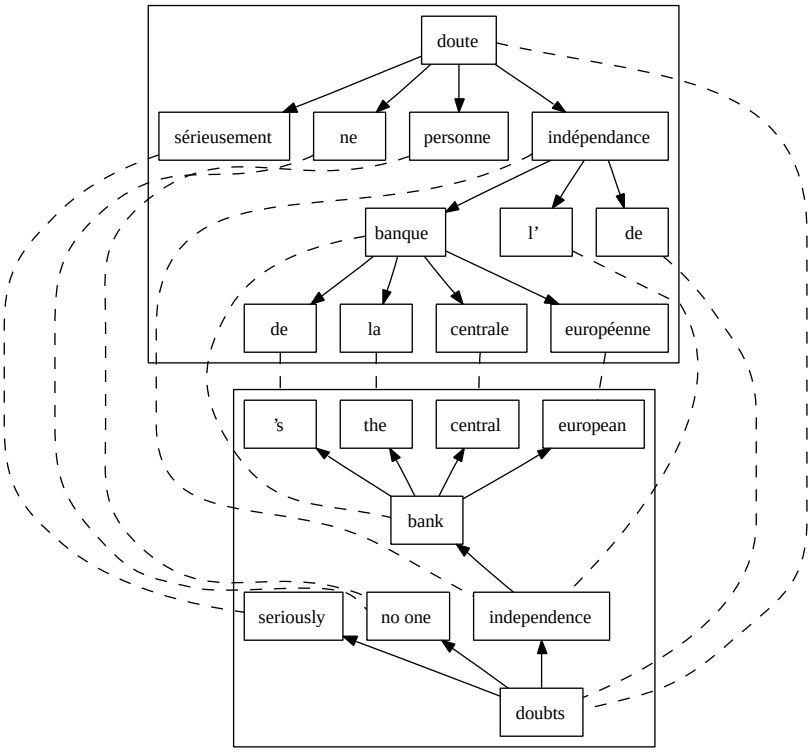


FIG. 2.5 – Exemple de bi-graphe pour une bi-phrase du corpus parallèle : "*personne ne doute sérieusement de l’ indépendance de la banque centrale européenne – no one seriously doubts the european central bank ’s independence*"

	No	one	seriously	doubts	the	european	central	bank	’s	independence
personne	1	1	0	0	0	0	0	0	0	0
ne	0	0	0	1	0	0	0	0	0	0
doute	0	0	0	1	0	0	0	0	0	0
sérieusement	0	0	1	0	0	0	0	0	0	0
de	0	0	0	0	1	0	0	0	0	0
l’	0	0	0	0	1	0	0	0	0	0
indépendance	0	0	0	0	0	0	0	0	0	1
de	0	0	0	0	0	0	0	0	1	0
la	0	0	0	0	0	0	0	0	1	0
banque	0	0	0	0	0	0	0	1	0	0
centrale	0	0	0	0	0	0	1	0	0	0
européenne	0	0	0	0	0	1	0	0	0	0

FIG. 2.6 – La matrice d’alignement A_{ST} pour une bi-phrase di corpus parallèle :"*personne ne doute sérieusement de l’ indépendance de la banque centrale européenne – no one seriously doubts the european central bank ’s independence*"

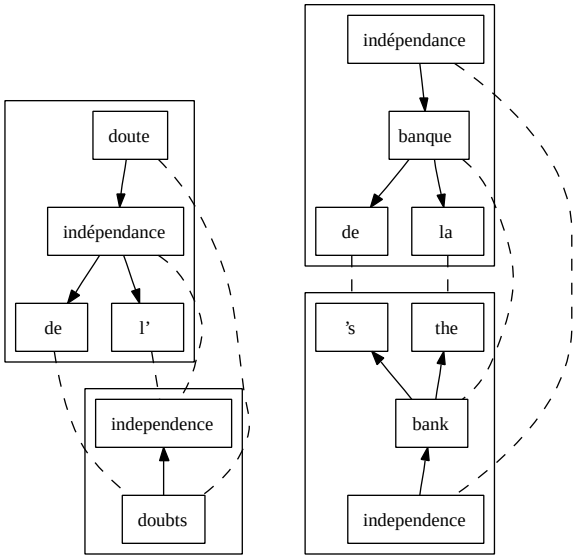


FIG. 2.7 – Exemples de bi-graphelets extraits à partir du bi-graphe de la Figure 2.5. Ces bi-graphelets correspondent aux bi-fragments : *doute _ de l'indépendance – doubts _ _ _ _ _ independence* (à droite), *indépendance de la banque – the _ _ bank's independence* (à gauche).

Nous utilisons les N meilleurs alignements produits par GIZA++ dans les deux directions[16] comme alignements de base. La combinaison des alignements source-cible et cible-source avec la méthode de symétrisation raffinée comme dans Och and Ney [2004] donne la liste des N^2 alignements possibles, parmi lesquels nous choisissons un meilleur alignement en suivant le critère de minimisation de l'énergie (équation 2.8).

2.3.1.2.2 Génération d'une bibliothèque de bi-graphelets. L'intérêt majeur des bi-graphelets lors de la traduction est que l'analyse syntaxique permet d'établir des dépendances à longue distance, auxquelles les systèmes de TA à fragments n'ont souvent pas accès. En même temps, l'imposition de contraintes strictes sur une structure syntaxique de bi-graphelets risque de générer un grand nombre d'objets rares, ayant une couverture bien plus faible que les bi-fragments (DeNeefe et al. [2007]).

Ainsi, nous nous proposons d'extraire une bibliothèque de bi-fragments (par analogie avec les modèles de TA à fragments) sans imposer de contraintes de cohérence syntaxique lors de l'extraction (cette cohérence a déjà été imposée implicitement lors de l'alignement au niveau lexical). Nous espérons ainsi obtenir une bibliothèque de bi-graphelets ayant une aussi bonne couverture que la bibliothèque de bi-fragments. Le nombre de bi-graphelets peut être tout de même plus important que le nombre de bi-fragments, car le même fragment peut correspondre à des graphelets différents, à cause de l'existence d'analyses syntaxiques différentes, en fonction du contexte dans lequel le fragment se trouve.

Cette construction de la bibliothèque de bi-graphelets implique qu'un bi-graphelet n'est pas nécessairement connexe : certains fragments peuvent correspondre à des graphelets non-connexes. Ainsi, un graphelet peut être vu plus généralement comme un fragment, avec des informations supplémentaires sur certaines relations syntaxiques entre les mots.

Afin de pouvoir exploiter les dépendances à longue distance, il est intéressant de construire une bibliothèque de bi-graphelets à partir d'une bibliothèque de bi-fragments qui ne sont pas nécessairement connexes. Nous pouvons utiliser le mécanisme de Matrax (section 2.1.2.1) ou un mécanisme similaire à

[16] Source vers cible : produit les alignements pour chaque mot source (un mot cible correspondant, ou un mot vide), et cible vers source : produit des alignements pour chaque mot cible.

celui de Chiang [2005] pour l'extraction de fragments à trous.

Les techniques de filtrage de la bibliothèque de bi-fragments doivent être adaptées au contexte des bi-graphelets. Ainsi, il est intéressant de garder les fragments à trous tels que les parties non-connectées du fragment sont reliés par une relation de dépendance dans un graphelet correspondant. Par exemple, un bi-graphelet *doute _ de l'indépendence, doubts _ _ _ _ _ independence* est plus intéressant qu'un bi-graphelet *doute _ _ _ _ _ la, doubts the,* même s'il contienne un nombre de trous plus important.

À chaque bi-graphelet, nous attribuons les scores suivants par analogie avec les scores des bi-fragments : $p(g_s|g_t)$, $p(g_t|g_s)$, $p_{lex}(g_s|g_t)$, $p_{lex}(g_t|g_s)$.

2.3.1.3 Décodage par bi-graphelets

Dans notre modèle, la traduction correspond à un chemin passant par le transfert horizontal du triangle de Vauquois (Figure 1.1). Les étapes principales de la traduction sont les suivantes :

1. l'analyse d'une phrase source (par XIP) donne un graphe de dépendances source G_s ;
2. le transfert transforme le graphe source G_s en un graphe cible G_t ;
3. la génération produit une traduction finale cible à partir du graphe de dépendances G_t et de contraintes sur l'ordre induites au niveau des chaînes.

L'étape 2 fera l'objet de cette sous-section. Dans la suite, le terme *décodage par bi-graphelets* se rapporte à la transformation d'un graphe de dépendances source en un graphe cible.

Cette architecture permet d'un côté de réduire l'espace de recherche du graphe cible optimal, car, contrairement aux modèles lexicaux, nous n'avons pas besoin de considérer le modèle de distorsion dans le cadre du décodage par des bi-graphelets : l'ordre des mots est décidé au moment de la génération à l'étape 3. D'autre part, le seul moyen de désambiguïsation dans ce type de modèle repose sur les relations syntaxiques cible.

En suivant les modèles log-linéaires de TA probabiliste, nous définissons le but du décodage comme la recherche d'une solution de l'équation :

$$\hat{G}_t = \arg\max_{G_t} P_\lambda(G_t|G_s) = \arg\max_{G_t,B} \lambda \cdot \Phi(G_t, G_s, B) \tag{2.9}$$

Dans ce cadre, $B = \{b_1, .., b_k\}$ est un ensemble de bi-graphelets compatibles avec le graphe source (la partie source de chaque b_i est un sous-graphe du graphe source G_s) tel que les parties cible des bi-graphelets forment un graphe cible G_t *cohérent* (cette notion est définie précisément ci-dessous).

Définition. Une *couverture* $p = \{g_1, .., g_k\}$ est un ensemble de graphelets source tel que chaque mot source soit couvert au moins une fois.

Définition. Une *partition* est une couverture telle qu'aucun graphelet n'en contient un autre.

Lors du décodage, nous allons chercher une partition du graphe source, permettant de couvrir le même mot source plusieurs fois.

Une couverture définit une transformation d'un graphe source en un *hypergraphe* dont les sommets correspondent aux graphelets de cette couverture[17], et les arcs correspondent aux parties communes entre deux graphelets. La figure 2.8 illustre une telle transformation.

De la même manière, l'ensemble des parties cible des bi-graphelets peut être transformé en un hypergraphe cible.

[17] Notons qu'ici la notion de graphelet est similaire au *scope* défini dans UNL [Uchida et al., 2005].

Définition. Un ensemble de graphelets cible forme un *graphe cible cohérent*, si l'hypergraphe cible formé est isomorphe[18] à l'hypergraphe source.

Cette définition implique que, pour chaque paire de graphelets source se chevauchant, l'intersection des graphelets cible correspondants doit correspondre à un sous-graphelet dans l'intersection des graphelets source (Figure 2.9).

D'un côté, ce type de décodage doit considérer plus d'hypothèses, ce qui augmente l'espace de recherche de la solution optimale. Considérer des objets se chevauchant donne la possibilité d'explorer mieux le contexte (des arcs différents peuvent contenir le même nœud mais appartenir à des graphelets différents) et une capacité de désambiguïsation plus forte. Cela permet aussi de considérer des objets de taille plus petite, et de réexploiter le contexte lors du décodage.

Le décodage peut être réalisé par un algorithme de type recherche en faisceau (*beam-search*), chaque nouvel état étant rajouté en conservant la cohérence du graphe cible.

Problèmes

Lors de l'avancement dans cette direction initiale, nous nous sommes aperçue de certains obstacles et risques liés à cette approche. Non seulement cette direction nécessitait le développement complet d'un décodeur et d'un générateur, ce qui représente un travail important, mais encore, nous n'avions pas de preuve forte d'améliorations significatives possibles grâce à l'introduction d'objets plus complexes. De plus, contrairement aux espoirs qu'on avait au début, certains travaux ont montré des dégradations suite à l'introduction d'objets complexes dans le processus de décodage (DeNeefe et al. [2007], Cowan et al. [2006]). Cette dégradation est souvent liée à la rareté des données, plus importante que dans le cas du modèle à fragments. De plus, la taille du modèle augmente dramatiquement, et des algorithmes de recherche beaucoup plus efficaces sont nécessaires pour pouvoir tirer avantage des connaissances supplémentaires introduites.

Les bi-graphelets, même formés à partir de bi-fragments, représentent tout de même des objets plus rares que les bi-fragments, ce qui risque de mener à une dégradation par rapport aux systèmes de TA probabiliste à fragments.

D'un côté, ne pas considérer l'ordre des mots lors du décodage permet de réduire significativement l'espace de recherche. De l'autre côté, cela peut introduire des problèmes pour résoudre certains ambiguïtés. Ces ambiguïtés sont généralement résolues par un modèle de langage dans des modèles probabilistes standard.

Après avoir analysé les risques et les perspectives de l'approche initiale, nous nous sommes décidée à prendre une autre direction de recherche, celle de l'extension du modèle de traduction à fragments, puisque l'approche de la TA à fragments représente une ligne de base forte dans le cadre de la TA probabiliste. À partir de maintenant, notre but sera donc d'enrichir ce modèle avec des connaissances sur la structure syntaxique, pour résoudre certaines des erreurs typiques dues au manque de généralisation et d'informations sur la structure de la phrase (section 2.2.4).

Le grand avantage de cette direction est la garantie d'avoir une traduction finale au moins aussi bonne que le modèle de base. Ainsi, nous avons moins de risque de dégrader la qualité de la traduction, et l'espoir d'apporter des améliorations là où c'est possible. Cette approche peut être vue comme une étape intermédiaire permettant d'étudier l'impact potentiel d'introduction des connaissances syntaxiques et d'étudier quel type de connaissances peut être plus intéressant à intégrer dans le cadre des modèles probabilistes. Nous pouvons ensuite procéder à l'élaboration de modèles plus complexes, en ayant une idée plus claire des avantages et désavantages qui peuvent être apportés par l'introduction des différents types d'informations syntaxique.

[18] Les hypergraphes ont le même nombre de sommets qui sont connectés de la même façon.

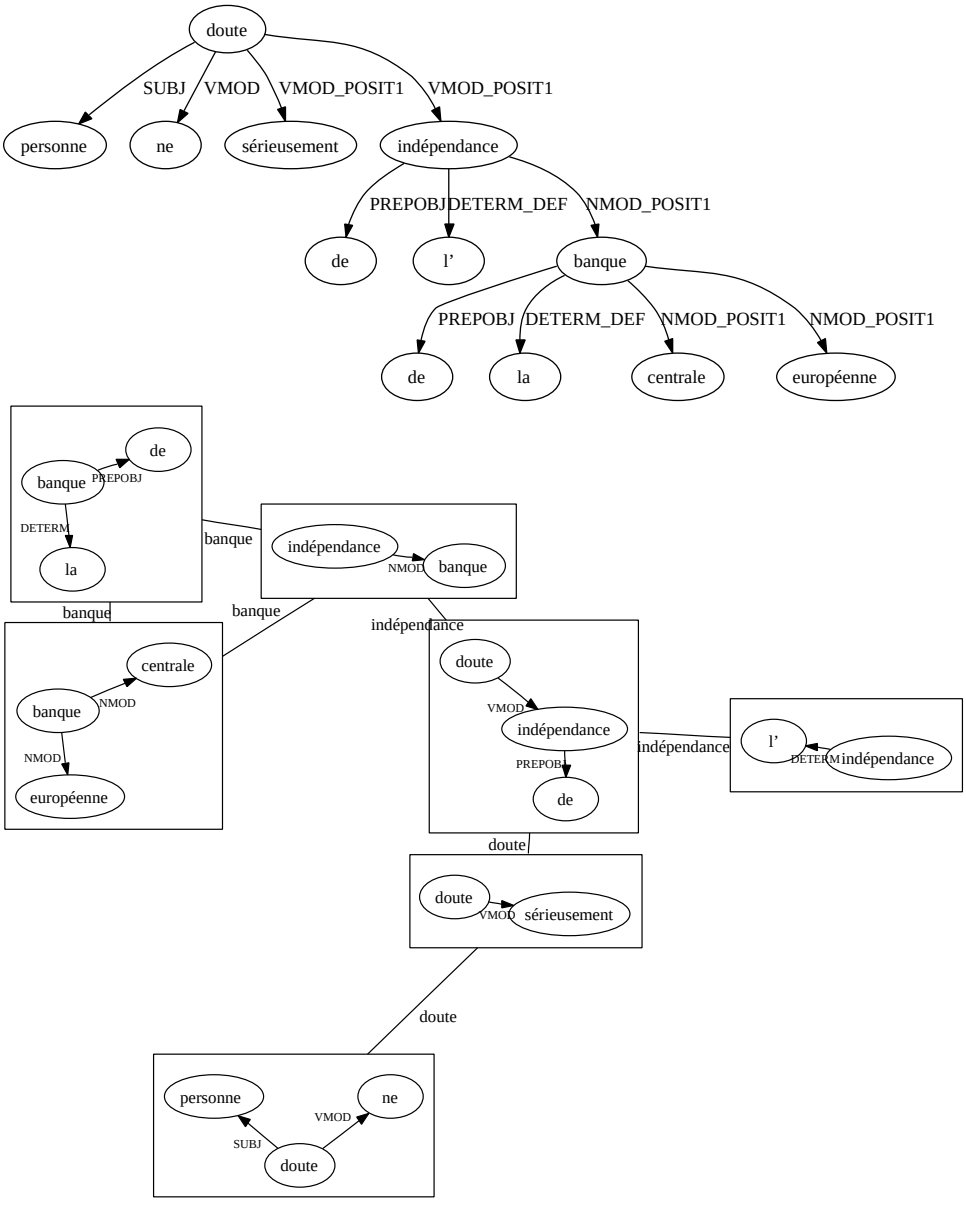


FIG. 2.8 – Exemple d’hypergraphe source

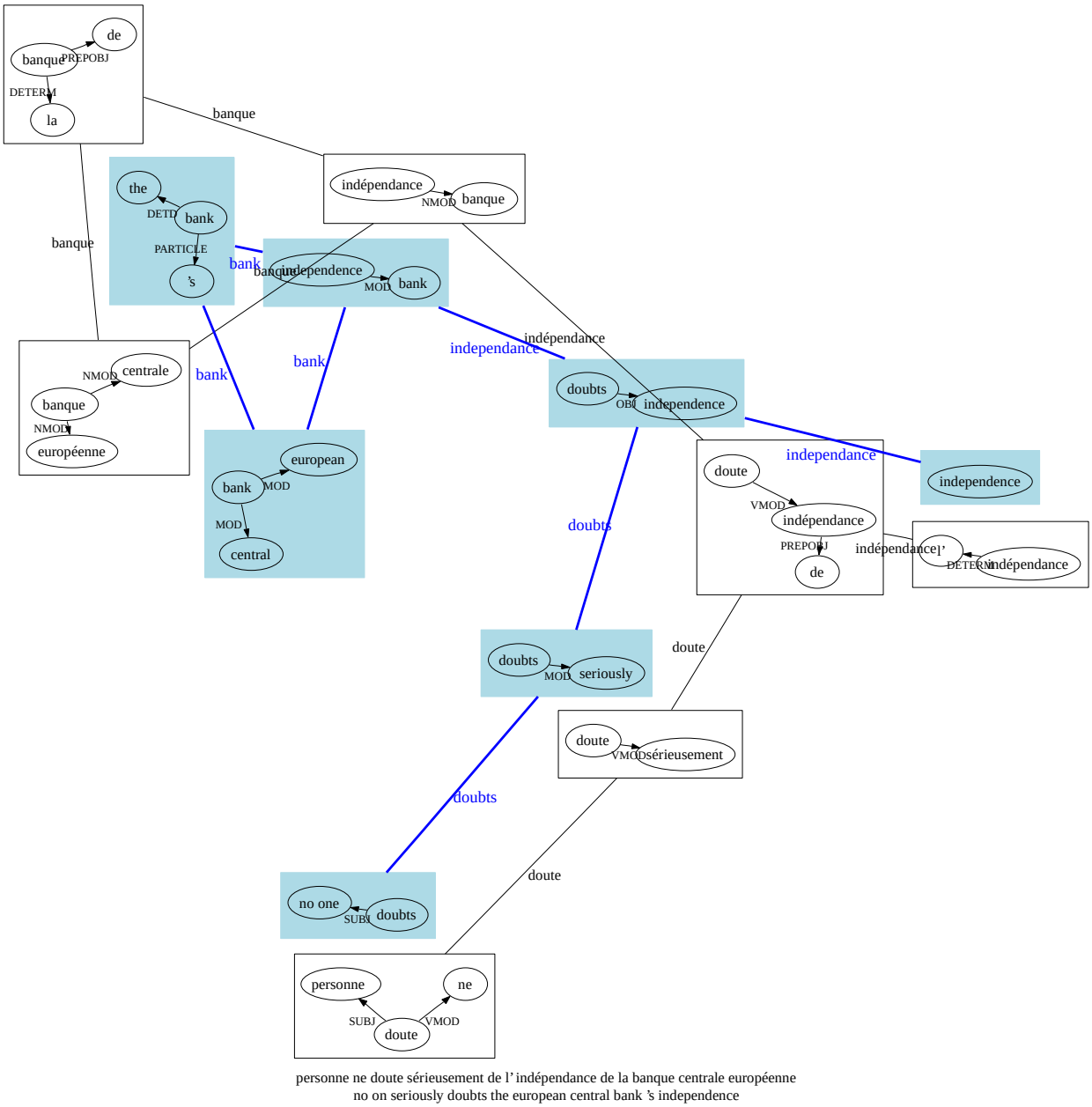


FIG. 2.9 – Exemple d'hypergraphe cible généré à partir d'un hypergraphe source

2.3.2 Directions de recherche prises : extension du modèle à fragments avec la structure syntaxique

2.3.2.1 Reclassement avec fonctions linguistiquement motivées

Motivation. La première approche que nous avons adoptée est l'intégration des fonctions de traits syntaxiques (bilingues et monolingues) dans le cadre du reclassement (section 1.3.3.2). L'approche par reclassement permet d'introduire facilement des traits linguistiquement motivés et d'estimer leur impact potentiel sur la qualité de la traduction. Il n'est pas nécessaire de créer une nouvelle base (nouveau décodeur, modèle de traduction ou génération, comme c'était le cas de la direction initiale), et nous avons un point de départ solide (modèle de la TA à fragments).

Les fonctions de traits que nous définissons sont basées sur les analyses de dépendances bilingues produites par XIP. Les parseurs XIP sont assez robustes, c'est-à-dire qu'ils peuvent produire des analyses (éventuellement fragmentaires) même pour des phrases mal formées et non grammaticales (ce qui est souvent le cas des traductions produites par un système de TA probabiliste). Ils sont donc bien adaptés à cette tâche.

Description de l'approche. L'idée du reclassement est d'introduire des fonctions de traits pour la liste des N meilleures traductions produites par un système de référence (modèle de TA probabiliste à fragments), et de choisir la traduction ayant le meilleur score défini par un modèle de reclassement discriminatif exploitant des informations linguistiques.

Nous introduisons des fonctions de *couplage* bilingue ayant pour but de mesurer la distance entre deux structures de dépendances source et cible produites par XIP. Les traits de *couplage* que nous avons introduits sont inspirés par la notion d'énergie d'alignement (équation 2.8) que nous avons utilisée pour choisir le meilleur alignement entre deux graphes. Le but des traits de *couplage* est de choisir un meilleur alignement, qui rende la structure cible plus proche de la structure source (ce qui est également le but du décodage par des graphelets).

Nous avons introduit plusieurs variantes de couplage basées sur l'énergie de l'alignement, en tenant compte des probabilités lexicales et des étiquettes des arcs correspondants.

Nous avons aussi introduit des traits monolingues, afin de tenir compte de la fluidité de la traduction, dont certains peuvent être adaptés au cas de la traduction par graphelets.

Limitations. L'approche par reclassement est limitée à la liste générée par le système de base. D'un côté, cela permet d'espérer que la traduction finale sera au moins aussi bonne que la traduction de base. De l'autre côté, le modèle de reclassement ne peut par contre pas générer de nouvelles hypothèses dans le cas où la liste des N meilleures traductions ne contient pas de bonne traduction. De plus, plus les traductions sont longues, moins leurs listes des N meilleures traductions sont variées.

Des essais précédents d'intégrer des informations linguistiquement motivées dans le cadre du reclassement (Och et al. [2003]) n'ont pas produit de résultats convaincants en termes de scores automatiques. Cependant, à notre connaissance, aucune analyse systématique manuelle des résultats n'a été faite afin de mieux comprendre l'impact réel des informations linguistiques introduites.

Objectifs. Nous allons introduire des ensembles de traits différents et étudier leurs impacts sur les scores automatiques, et, plus important, sur la qualité de la traduction, jugée par un évaluateur humain. Notre but est de déterminer :

- quel est le potentiel réel d'amélioration qu'on peut obtenir à partir d'une liste de traductions,
- si les faibles améliorations des scores automatiques correspondent à des améliorations réelles des traductions,
- comment les améliorations réelles sont corrélées avec les scores automatiques.

Nous allons en déduire l'impact de différents ensembles de traits syntaxiques sur l'adéquation et la fluidité de la traduction (en appliquant le protocole d'évaluation défini au 2.2.3.3).

2.3.2.2 Traduction hybride compositionnelle

Cette approche prend une direction orthogonale à l'approche par reclassement. Nous proposons un modèle qui a pour but de simplifier le processus de la traduction à l'aide d'informations syntaxiques source. Notre but était principalement de diriger le processus de la traduction vers les hypothèses les plus intéressantes du point de vue de la structure syntaxique source.

Plusieurs travaux ont déjà tenté d'extraire des sous-fragments syntaxiques à partir d'une phrase source en espérant simplifier la tâche de traduction :

- Koehn [2003] a essayé de prétraduire les syntagmes nominaux et de les intégrer dans la bibliothèque des bi-fragments ; il a cependant montré que cela est redondant avec la notion de bi-fragment dans les systèmes de TA à fragments, et n'apporte pas d'améliorations supplémentaires.

- Hewavitharana et al. [2007] sont allés plus loin dans cette direction en proposant de remplacer les syntagmes nominaux par un symbole non-terminal, et d'entraîner un modèle de traduction contenant des nœuds non terminaux. Ce modèle de traduction en combinaison avec le modèle à fragments est utilisé ensuite pour le décodage de la phrase source. Ce modèle a montré des améliorations significatives pour la traduction de l'arabe vers l'anglais. L'introduction de symboles non-terminaux permet de généraliser les bi-fragments extraits du corpus parallèle. Cette capacité de généralisation n'est toutefois pas tout à fait présente au moment du décodage, car seuls les syntagmes nominaux vus lors de l'apprentissages peuvent être généralisés : aucune analyse syntaxique n'intervient à ce stade.

- Mellebeek et al. [2006] propose de décomposer une phrase source en "squelette" et en des fragments correspondant à l'expansion des mots simples du squelette, et traduisibles séparément. Les traductions du squelette sont faites indépendamment (par un ou plusieurs systèmes de TA) et des fragments et sont ensuite combinées afin de construire une traduction finale. Une telle décomposition est définie à partir de l'analyse syntaxique source, et les erreurs d'analyse peuvent influencer la qualité de la traduction finale.

- Cherry [2008] vérifie, lors du décodage, que la traduction est cohérente avec la structure de dépendance source, en faisant l'hypothèse que, lors de la traduction d'un sous-arbre de dépendances, seules des distorsions locales sont possibles. Une variante plus générale de cette idée propose d'intégrer un trait indiquant si la contrainte de cohésion au niveau des sous-arbres de dépendances est respectée. Cela permet de tenir compte des erreurs de l'analyse syntaxique et rend le modèle plus générique.

Nous proposons un modèle introduisant des contraintes de distorsion à partir de l'analyse syntaxique d'une phrase source. Ce modèle tient compte non seulement de la structure de constituants et de l'analyse de dépendances, mais aussi d'autres traits externes à l'analyse syntaxique de la phrase source (nombre d'occurrences des fragments choisis dans la partie source du corpus). L'introduction de ces traits de nature différente rend le modèle plus résistant aux erreurs d'analyse.

Nous proposons ensuite une extension de ce modèle qui suit l'approche de Mellebeek et al. [2006] et simplifions la phrase source initiale afin de traduire de façon autonome certains de ses fragments. La simplification d'une phrase source signifie le remplacement de certains fragments d'une phrase initiale par des sous-fragments plus simples. Par exemple, la traduction de la phrase "*J'ai lu ce fameux livre dont tu m'as tant parlé* " peut se résumer à la traduction des fragments suivants :

- *j'ai lu ce livre*
- *fameux livre dont tu m'as tant parlé*

La traduction de chaque sous-segment est une tâche plus simple pour un système de TA probabiliste car le

graphe d'hypothèses construit pour chacune des sous-phrases est plus petit que le graphe des hypothèses de la phrase initiale. De plus, moins d'hypothèses inintéressantes sont considérées lors du décodage (par exemple les distorsions inutiles), ce qui nous laisse espérer trouver une meilleure traduction en fin de compte.

Dans le cas de chacun de ces modèles, le processus de traduction est décomposé en une suite d'étapes successives autonomes. Cela introduit le risque d'accumuler les erreurs d'une étape à l'autre. Nous proposons une procédure d'optimisation des paramètres des étapes intermédiaires, ayant la qualité de la traduction finale comme objectif. Ce type d'apprentissage nous permet de mesurer la qualité de la décomposition en fonction de la qualité de la traduction finale, et non en introduisant des mesures artificielles pour des étapes intermédiaires.

Il est ensuite possible d'introduire une étape de reclassement dans le modèle décomposant une phrase source. Grâce au modèle de décomposition, nous espérons générer des hypothèses plus intéressantes et, par conséquent, l'approche par reclassement aura plus de potentiel : on peut espérer que l'ajout de connaissances sur la structure syntaxique cible permettra de choisir une meilleure traduction parmi les hypothèses initiales.

Deuxième partie

Reclassement avec des fonctions de traits linguistiquement motivées

Introduction

Le but des expériences sur le reclassement avec des traits linguistiquement motivés est d'étudier l'impact potentiel de l'introduction de connaissances linguistiques bilingues dans un modèle de TA probabiliste.

L'idée du reclassement consiste à produire une liste d'hypothèses de traduction par le système de base, et à choisir la meilleure traduction dans cette liste en utilisant des informations externes, non accessibles au décodeur. L'avantage de ce type d'approche est qu'elle n'exige pas de modification du système de traduction, elle nous permet d'utiliser des fonctions de traits complexes qui seraient coûteuses à intégrer dans le décodeur (par exemple : l'analyse syntaxique de la phrase cible), et d'évaluer l'utilité et l'impact possibles de ces fonctions de traits non seulement sur les scores automatiques, mais aussi en évaluant les traductions manuellement.

Le modèle de reclassement est entraîné par un Perceptron Structuré, qui permet d'optimiser les paramètres du modèle assez rapidement, même pour des modèles assez grands.

Nous introduisons des traits basés sur l'analyse syntaxique produite par XIP pour une phrase source et une phrase cible afin d'améliorer l'adéquation de la traduction. L'adéquation d'une traduction est estimée comme la similarité entre les structures de dépendances source et cible produites par XIP. Nous introduisons des mesures de similarité différentes, et étudions leur impact sur la qualité de la traduction en les intégrant en tant que traits dans le modèle de reclassement discriminatif.

Nous introduisons aussi quelques traits monolingues basés sur l'analyse de dépendance et l'analyse morphologique de la cible pour contrôler la fluidité de la traduction.

Nous évaluerons l'impact des fonctions de traits à l'aide de scores automatiques du type BLEU/NIST, mais aussi en faisant des évaluations manuelles.

Chapitre 3

Fonctions de traits pour le reclassement

Nous sommes intéressée par deux critères principaux de la qualité de traduction : la fluidité et l'adéquation. Nous introduisons donc des fonctions de traits linguistiquement motivées permettant de contrôler chacun de ces deux aspects. Ces traits se séparent naturellement en deux groupes :

- les traits bilingues (prenant en compte les structures linguistiques source et cible) qui visent à contrôler l'adéquation de la traduction ;
- les traits monolingues cible, qui visent à contrôler la fluidité de la traduction.

3.1 Fonctions de traits bilingues

Pour contrôler l'adéquation de la traduction, nous introduisons une mesure de similarité entre les deux structures de dépendances proposées par XIP. Grâce à sa robustesse, XIP sera toujours capable de produire une analyse d'une phrase cible, même si elle n'est pas bien formée, ce qui est souvent le cas des sorties des systèmes de TA probabiliste. Les structures de XIP que nous allons utiliser sont les *graphes de dépendances*, définis dans la section 2.3.1.
Nous allons faire l'hypothèse que, si deux mots source sont reliés par une relation de dépendance, il est probable que les deux mots cible correspondants sont aussi reliés par une relation de dépendance.
Pour rendre cette idée plus claire, prenons l'exemple de la Figure 3.1. La première traduction de l'exemple est une traduction proposée par le système, et la deuxième traduction a été trouvée plus loin dans la liste des N meilleures traductions. Bien qu'aucune des deux traductions ne soit parfaite, la deuxième traduction semble meilleure, car le mot "substantiel" est mieux placé, ce qui rend la traduction plus adéquate.

Nous supposons connues les correspondances mot à mot entre la phrase source et sa traduction pour chacune de ces deux traductions. Si l'on regarde les structures de dépendances produites par XIP pour chacune des deux traductions, le meilleur emplacement du mot "substantiel" est équivalent à une préservation de la relation de dépendance source *(assistance -> substantial)* dans la traduction (relation *(aide -> substantielle)*). La mesure de similarité entre les deux structures produites par XIP repose sur le nombre des relations de dépendances préservées entre la structure source et la structure cible. Cette mesure de similarité sera appelée *couplage* par la suite.

3.1.1 Couplage générique

Nous avons mis en place plusieurs variantes de cette idée de base. Nous commençons par décrire différentes fonctions de couplage "génériques" dérivées à partir de l'idée de base, en supposant que les alignements de mots ont déjà été déterminés, puis nous décrivons l'option de prendre en considération la matrice d'alignement lexical pondéré et les étiquettes spécifiques de dépendance.

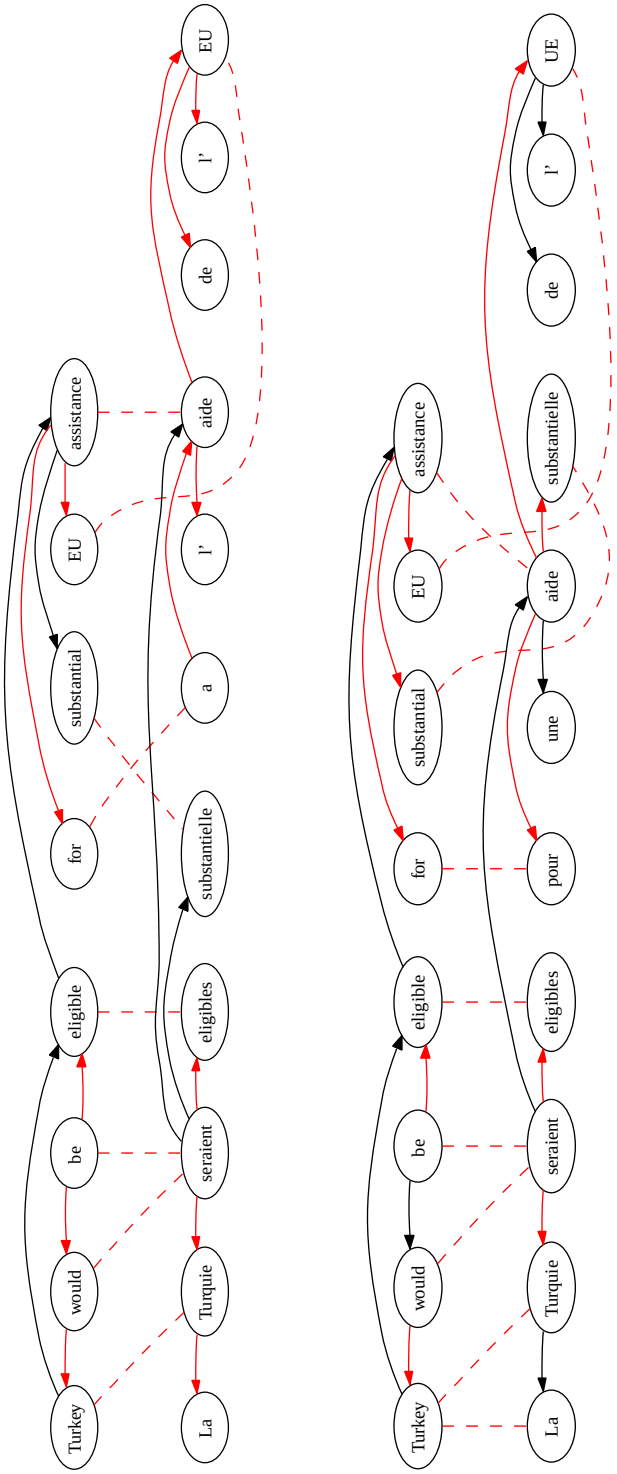


FIG. 3.1 – Exemple de deux phrases parmi les N meilleures traductions produites par le système de TA probabiliste. La première phrase est proposée par le système comme solution optimale (en haut), la deuxième se trouve plus bas dans la liste (en bas).

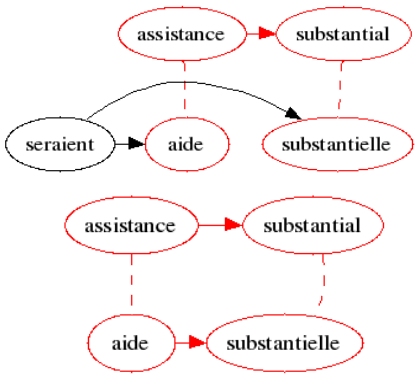


FIG. 3.2 – Exemple de rectangle.

Nous allons d'abord présenter quelques fonctions que l'on définit comme *traits de couplage générique*. Les premières mesures de couplage sont basées sur des alignements de mots simples, non pondérés.

3.1.1.1 Décompte

Notre approche générale pour le calcul du couplage entre la structure de dépendance de la phrase source et celle d'une traduction candidate est la suivante :

- nous commençons par aligner les mots entre la phrase source et la traduction candidate,
- nous produisons les graphes de dépendances XIP de la phrase source et de la phrase cible,
- nous comptons le nombre de configurations ("rectangles") qui sont du type suivant : $((s_1, s_{12}, s_2), (t_1, t_{12}, t_2))$, où s_{12} est un arc (relation de dépendance) entre s_1 et s_2, t_{12} est un arc (relation de dépendance) entre t_1 et t_2, s_1 est aligné avec t_1 et s_2 est aligné avec t_2.

Ainsi, dans notre exemple de la Figure 3.2, s_1 correspondra à *assistance*, s_2 à *substantial*, t_1 à *aide*, et t_2 à *substantielle*. La différence dans les décomptes des rectangles entre les deux phrases de l'exemple de la Figure 3.1 sera le rectangle correspondant à (*substantial assistance - aide substantielle*), qui est présent dans la deuxième traduction, mais n'est pas présent dans la première traduction.

La fonction de décompte que nous avons présentée peut être vue, plus généralement, comme l'énergie d'appariement des graphes (équation 2.8). Supposons que D_S et D_T représentent les matrices d'adjacence des graphes de dépendances source et cible ($d_{ij} = 1$ s'il existe une relation de dépendance entre les mots i et j, sinon $d_{ij} = 0$), et que A est une matrice d'alignement entre les mots de la phrase source et de la phrase cible ($a_{ik} = 1$ si le mot s_i de la phrase source correspond au mot t_k de la phrase cible, sinon $a_{ik} = 0$). L'énergie d'appariement entre les graphes de dépendances source et cible E_{ST} est définie comme suit [BinLio and Hancock, 2001] :

$$E_{ST} = -Tr(D_S^t A D_T A^t)/2 \tag{3.1}$$

La décomposition nous donne (N_s est le nombre de mots source, N_t le nombre de mots cible) :

$$E_{ST} = -\frac{1}{2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\sum_{k=1}^{N_t}\sum_{l=1}^{N_t} d_{ij}^s d_{kl}^t a_{ik} a_{jl} = -\text{Décompte} \tag{3.2}$$

Cette décomposition est équivalente (au signe près) à la notion du décompte que nous avons introduite, car cela revient à compter les relations de dépendances source et cible d_{ij}^s et d_{kl}^t, telles que le mot source

s_i soit aligné avec le mot cible t_k ($a_{ik} = 1$) et que le mot source s_j soit aligné avec mot cible t_l ($a_{jl} = 1$).

3.1.1.2 Variantes des fonctions de décompte

La valeur de *décompte de couplage* est fortement corrélée avec la taille des structures de dépendances source et cible. Nous présentons quelques variantes normalisant cette fonction, afin de neutraliser l'effet de la taille des phrases :

– *La précision de couplage* compte quelle proportion des arcs source est présente dans le graphe cible :

$$\text{Précision} = \frac{\text{Décompte}}{\text{Nombre des arcs source}} \tag{3.3}$$

– *Le rappel de couplage* compte quelle proportion des arcs cible a des arcs source en correspondance :

$$\text{Rappel} = \frac{\text{Décompte}}{\text{Nombre des arcs cible}} \tag{3.4}$$

– *F-mesure de couplage*. Dans le cas de structures de dépendances parfaitement isomorphes (une situation qui se produit rarement naturellement en raison des divergences linguistiques entre les deux langues), la précision et le rappel seraient égaux à 1. Afin de mesurer la divergence par rapport à ce cas idéal, nous proposons une fonction que nous appelons *Couplage-F-mesure*, qui est définie comme la moyenne harmonique des deux fonctions précédentes :

$$\text{F-mesure} = \frac{2 * \text{Précision} * \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{3.5}$$

3.1.1.3 Alignements lexicaux

Les fonctions de couplage que nous avons introduites supposent connus les alignements au niveau des mots. Certains systèmes de TA à fragments, comme Moses, peuvent produire ces alignements avec les listes des N meilleures traductions. Dans ce cas-là, nous pouvons exploiter directement les alignements fournis par le système.

D'autres systèmes à fragments, comme Matrax, produisent les alignements au niveau des bi-fragments qui ont été utilisés afin de générer la traduction, mais ne donnent pas d'informations sur des alignements plus fins. Pour obtenir des alignements lexicaux dans ce cas de figure, nous proposons deux solutions.

– La première consiste à générer les alignements lexicaux à l'aide de GIZA++ [Och and Ney, 2003]. Une fois que les alignements sont générés, nous pouvons calculer nos fonctions de couplage générique comme défini précédemment. L'avantage de cette solution est qu'elle peut s'appliquer aux cas plus généraux que les systèmes de TA probabiliste à fragments. Par exemple, dans les cas où on n'a aucun accès aux informations internes du système qui a produit les traductions, et donc aucune information ni sur les alignements lexicaux, ni sur les bi-fragments utilisés, ni même sur la nature du système qui a produit ces traductions. Cela peut être généralisé au cas où les traductions proviendraient de systèmes différents (combinaison de systèmes de TA).

– Une autre solution est moins générale, mais convient bien au cas d'un système de TA probabiliste à fragments. Plutôt que de considérer les relations de dépendance entre mots, nous allons les transformer en dépendances entre fragments, pour lesquelles les alignements sont connus.
Les transformations des arcs sont illustrées sur la Figure 3.3. Seules les relations de dépendance reliant les parties source (cible) de deux bi-fragments différents seront prises en compte pour la structure source (cible). Aucune dépendance à l'intérieur d'un bi-fragment n'est prise en compte dans ce cas de figure (Exemple : une relation de dépendance reliant les mots *would* et *be* n'est pas

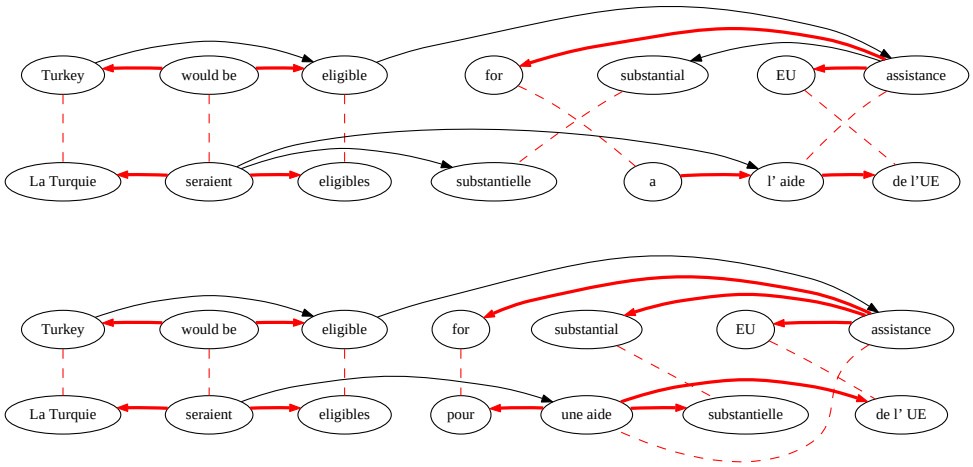


FIG. 3.3 – Exemple de couplage au niveau des fragments. En gras : les relations de dépendance partagées entre la structure source et la structure cible. La première phrase est proposée par le système comme solution optimale (en haut), la deuxième se trouve plus bas dans la liste (en bas).

prise en compte dans le décompte des rectangles). La figure 3.3 montre que le passage des relations au niveau lexical aux relations au niveau des fragments rend les structures de dépendances plus proches (par rapport à celles de la Figure 3.1).

3.1.2 Extensions des fonctions de couplage générique

3.1.2.1 Couplage lexical

Les fonctions de couplage générique se basent sur des alignements binaires de mots, et donc ne prennent pas en considération la probabilité d'aligner un mot source avec un mot cible. Ainsi, si l'alignement lexical est incorrect (l'alignement produit par GIZA++), cela peut mener à des couplages erronés. Comme solution possible à ce problème, nous introduisons une fonction *Couplage-Lex* qui prend en compte les probabilités lexicales de la manière suivante : chaque rectangle trouvé entre la structure source et cible est pondéré par le produit des probabilités conditionnelles $p(t_1|s_1)$ et $p(t_2|s_2)$[1] .

$$\text{Décompte-Lex} = \sum_{\text{rectangles}(s_1,s_2,t_1,t_2)} p(t_1|s_1) \cdot p(t_2|s_2) \tag{3.6}$$

En termes d'énergie d'appariement des graphes, cela revient à remplacer la matrice d'alignement A, $a_{ij} \in \{0,1\}$ par la matrice A_{lex}, où $a_{ij} = p(t_i|s_j)$[2].

$$\text{Précision-Lex} = \frac{\text{Décompte-Lex}}{\text{Nombre des arcs source}}$$
$$\text{Rappel-Lex} = \frac{\text{Décompte-Lex}}{\text{Nombre des arcs cible}}$$

De la même façon, nous introduisons une fonction de trait *Couplage-Rev-Lex* qui prend en compte les

[1] $p(t_1|s_1)$ ($p(t_2|s_2)$) est la probabilité que le mot s_1 (resp. t_1) soit traduit par le mot t_1 (resp. s_1) estimée sur un grand corpus parallèle.

[2] Notons que la précision et le rappel lexicaux atteignent rarement la valeur 1.

probabilités conditionnelles inverses.

$$\text{Décompte-Rev-Lex} = \sum_{\text{rectangles}(s_1,s_2,t_1,t_2)} p(s_1|t_1) \cdot p(s_2|t_2)$$

$$\text{Précision-Rev-Lex} = \frac{\text{Décompte-Rev-Lex}}{\text{Nombre des arcs source}}$$

$$\text{Rappel-Rev-Lex} = \frac{\text{Décompte-Rev-Lex}}{\text{Nombre des arcs cible}}$$

Alignements au niveau des fragments. Les traits de couplage lexical sont facilement adaptables au couplage au niveau des fragments : il suffit de prendre les probabilités $p(S|T)$, $p(T|S)$, où S et T sont des fragments (groupes des mots) de la phrase source ou cible, $S = s_1 \cdots s_k$, $T = t_1 \cdots t_l$. Dans le cas d'un système de TA probabiliste à fragments, ces probabilités se trouvent dans la bibliothèque de bi-fragments.

3.1.2.2 Couplage étiqueté

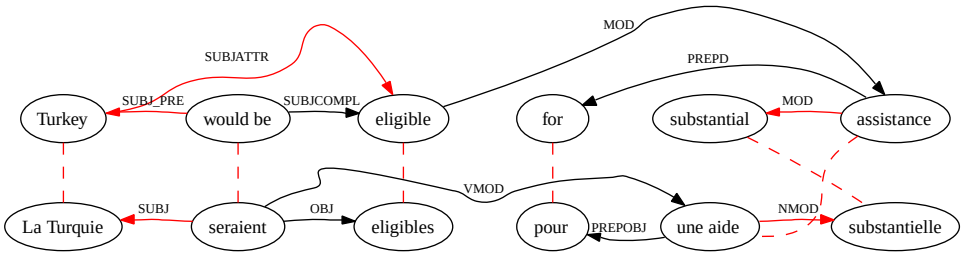


FIG. 3.4 – Exemple de couplage avec des relations de dépendance étiquetées.

Les fonctions de traits précédemment définies ne prennent pas en considération les étiquettes liées aux arcs dans les structures de dépendance. Pourtant, bien que les rectangles de la forme $((s_1, \text{subj}, s_2), (t_1, \text{subj}, t_2))$ soient assez systématiques entre des langues telles qu'anglais et français, d'autres rectangles peuvent l'être beaucoup moins, d'une part à cause de divergences linguistiques réelles entre les deux langues, d'autre part à cause de la différence des grammaires utilisées par XIP (Figure 3.4 : dans le cas de cet exemple, la dépendance du type "SUBJATTR" est définie en anglais, mais pas en français).

Nous souhaitons apprendre les étiquettes pour lesquelles l'appariement est systématique (tels que *subj*), et les étiquettes qui n'ont systématiquement pas d'arc correspondant cible (source).

Nous définissons une collection de traits *Couplage étiqueté*, chacun correspondant à une paire spécifique d'étiquettes de source et cible (lab_{src}, lab_{tgt}). La valeur d'un trait défini par (lab_{src}, lab_{tgt}) est le nombre d'occurrences de cette paire spécifique parmi les rectangles du couplage.

Pour clarifier cette idée, nous présentons la collection des traits correspondant à l'exemple de la Figure 3.4 :

$\phi_{SUBJ\_PRE-SUBJ} = 1$; $\phi_{SUBJCOMPL-OBJ} = 1$; $\phi_{PREPD-PREPOBJ} = 1$; $\phi_{MOD-NMOD} = 1$

Nous employons uniquement les paires d'étiquettes qui ont été observées alignées dans le corpus de développement (c'est-à-dire, celles contenues dans les rectangles observés). Lors de l'apprentissage, nous apprenons les poids $\lambda_{lab_s\_lab_t}$ pour chaque paire (lab_s, lab_t). Lors du reclassement, la valeur correspondant à $\sum_{lab_s\_lab_t} \lambda_{lab_s\_lab_t} \cdot \phi_{lab_s\_lab_t}$ sera équivalente à la fonction de décompte, où chaque rectangle

est pondéré par le poids correspondant à la paire des étiquettes des arcs de ce rectangle.

3.2 Fonctions monolingues

Le but des fonctions de couplage est de mesurer la similarité entre les structures source et cible. La fluidité de la traduction n'est pas du tout prise en compte par les traits de couplage. De plus, les traductions qui ne sont pas fluides peuvent avoir une structure parfaitement isomorphe avec une phrase source grâce à la robustesse de XIP.

Exemple :

Phrase source :	*The debate is closed.*
Traduction proposée :	*Le débat est clos.*
	Le débat est close.
	La discussion est clos.
	La discussion est close.
	Les débat est clos.
	Les débats sont clos.

Les graphes de dépendances proposés par XIP pour chacune de ces traductions seront exactement les mêmes et sont parfaitement isomorphes entre la phrase source et chacune des traductions candidates. Les fonctions de couplage décrites précédemment ne sont pas en mesure de distinguer entre ces traductions. Dans cette section, nous introduisons des fonctions de traits supplémentaires se basant sur les informations linguistiques de la phrase cible pour distinguer entre les traductions bien formées et les traductions mal formées.

Nous proposons ici deux types de fonctions de traits monolingues permettant de contrôler la fluidité de la traduction.

3.2.1 Fonction de cohésion linguistique

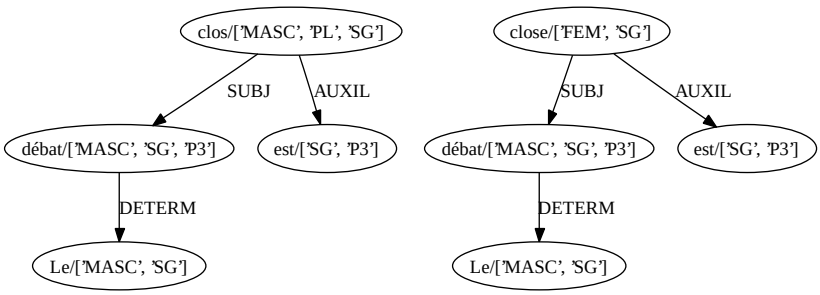


FIG. 3.5 – Structures de dépendances produites par XIP, avec les traits morphologiques sur les mots. À gauche : Le débat est clos. À droite : Le débat est close.

3.2.1.1 Motivation

Considérons de nouveau les deux premières traductions de notre exemple. Cette fois-ci, nous utilisons non seulement les dépendances entre les mots, produites par XIP, mais aussi les traits morphologiques de ces mots (figure 3.5).

Sur la figure 3.5[3], nous voyons que la différence entre les deux analyses est dans les traits linguistiques portant sur les mots *clos* et *close*. Nous voyons également que, dans la première analyse, les mots "*clos*" et "*débat*" qui sont connectés par la relation de dépendance *SUBJ* partagent le trait *MASC*, ce qui n'est

[3]Le mot *clos* contient à la fois les traits morphologiques du pluriel et du singulier : l'analyse morphologique est faite au niveau du mot et des traits à valeurs multiples peuvent être attribués.

pas le cas de la deuxième analyse. L'intuition serait alors de garder les traits morphologiques les plus cohérents entre le dépendant et le gouverneur.

Notons tout de même que cela ne sera pas nécessairement vrai pour toutes les relations de dépendance. Ainsi, si nous prenons par exemple la relation de type *NMOD*, où le nom est modifié par un nom (*ex : les livres de cuisine*), le gouverneur et le dépendant ne partagent pas forcément les mêmes traits morphologiques. Nous souhaitons apprendre les dépendances pour lesquelles la cohésion de chaque type doit être respectée.[4]

3.2.1.2 Définition

Nous définissons les traits de *cohésion linguistique* afin de tenir compte des traits morphologiques cohérents ou non-cohérents entre le gouverneur et le dépendant d'une dépendance.

Nous définissons une collection de traits, chacun concernant une étiquette spécifique lab et un trait morphologique t (qu'on spécifie manuellement pour la langue cible[5]).

La valeur d'un trait défini par (lab, t) est le nombre d'occurrences où le trait morphologique t est partagé entre un gouverneur et un dépendant reliés par une dépendance du type lab.

Les collections des traits correspondant aux traductions de l'exemple de la Figure 3.5 seront définies ainsi :

Exemple :
Le débat est clos.
$\phi_{subj\_genre} = 1$; $\phi_{subj\_nb} = 1$; $\phi_{auxil\_nb} = 1$; $\phi_{determ\_genre} = 1$; $\phi_{determ\_nb} = 1$.

Le débat est close.
$\phi_{subj\_genre} = 0$; $\phi_{subj\_nb} = 1$; $\phi_{auxil\_nb} = 1$; $\phi_{determ\_genre} = 1$; $\phi_{determ\_nb} = 1$.

La différence entre les deux traductions de cet exemple est la valeur du trait ϕ_{subj_genre} qui indique si le trait morphologique du type *genre* est partagé entre le gouverneur et le dépendant d'une relation du type $subj$. Dans la première traduction, le gouverneur *clos* de l'arc $SUBJ$ contient le trait du type $genre$: *MASC*, ainsi que son dépendant *débat*. Cela n'est pas le cas pour la deuxième traduction et, donc, peut servir d'indicateur de non-fluidité de la traduction.

De façon similaire, nous pouvons introduire des traits de non-cohésion, qui correspondent aux arcs pour lesquels la cohésion d'un certain type entre le gouverneur et le dépendant n'est pas respectée. Ces traits sont complémentaires des traits de cohésion : dans le cas de deux phrases où les traits de cohésion sont identiques, les traits de non-cohésion permettent de distinguer entre ces deux phrases.

3.2.2 Modèle de langage factoriel discriminatif

3.2.2.1 Motivation

L'idée du modèle de langage factoriel (MLF) est d'exploiter non seulement les n-grammes des formes de surface des mots, mais aussi les n-grammes de facteurs différents sur chaque mot (tels que lemme, surface, partie du discours), ce qui lui donne plus de capacité de généralisation ([Koehn and Hoang, 2007; Mahé and Cancedda, 2008]).

Ce trait monolingue permet de contrôler que la suite des n mots est possible dans la langue cible, contrairement à la fonction de cohésion qui contrôle plus la cohérence entre les formes morphologiques des mots. Ainsi, la fonction de cohésion ne sera pas capable de distinguer entre les constructions *"se souvenir à"* (fausse) et *"se souvenir de"* (juste). Un modèle de langage n-gramme serait capable de le faire, à condition que le trigramme correct (*"se souvenir de"*) ait été vu dans le corpus d'entraînement.

[4] Ainsi, ce modèle peut être facilement adapté à d'autres analyseurs de dépendances, et à d'autres langues.

[5] Ainsi, si on traduit vers l'anglais, nous pouvons considérer les traits *nombre* (pluriel, singulier) et *personne* (p1, p2, p3) ; dans le cas du français, on peut aussi considérer le trait *genre* (masculin, féminin).

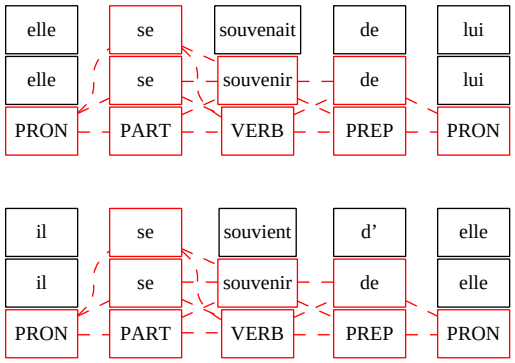


FIG. 3.6 – Exemple des représentations factorielles des deux phrases. En haut : elle se souvenait de lui ; en bas : il se souvient d'elle.

Il est toutefois peu probable que toutes les formes de verbe aient été observées dans le corpus d'entraînement. Ainsi, si la forme *"se souvenait"* n'est pas présente dans le corpus d'entraînement, le modèle de langage n-gramme classique ne sera pas capable de distinguer entre les constructions *"se souvenait à"* et *"se souvenait de"*. Un modèle de langage factoriel, en revanche, aura accès non seulement aux formes de surface, mais aussi aux lemmes. Ainsi, il suffit de trouver un trigramme des lemmes *"se souvenir de"* dans le corpus d'entraînement (où le verbe "se souvenir " peut avoir n'importe quelle forme de surface), et un modèle de langage factoriel sera capable de discriminer la construction *"se souvenait de"* de la construction *"se souvenait à"*.

3.2.2.2 Description d'un modèle de langage discriminatif basé sur des noyaux de séquences factoriels

Les traits du modèle de langage discriminatif factoriel. Le modèle de langage discriminatif factoriel proposé dans [Cancedda et al., 2008] utilise des n-grammes basés sur les formes de surface, les lemmes et les parties de discours comme fonctions de traits. D'après [Mahé and Cancedda, 2008], l'énumeration de tels n-grammes communs entre les deux phrases correspond aux "noyaux de séquences factoriels" d'ordre n[6].

Nous définissons $V = V_s \cup V_l \cup V_p$, où V_s est un ensemble de toutes les formes de surface possibles, V_l l'ensemble de tous les lemmes, et V_p l'ensemble de toutes les parties du discours. Les traits d'ordre n sont alors définis comme les élements (vus comme des n-grammes) de V^n. Les fonctions de traits d'ordre n sont définies par $\Phi_{MLF_n} : V^n \rightarrow \mathbf{N}$ et font correspondre à chaque élément de V^n son nombre d'occurrences dans la phrase.

Prenons l'exemple de la figure 3.6. Dans cet exemple, nous cherchons à comparer les deux phrases : t_1 = "il se souvient d'elle" et t_2 = "elle se souvenait de lui". Les n-grammes communs d'ordre 3 entre t_1 et t_2 sont : *"PRON se souvenir"*, *"PRON PART souvenir"*, *"PRON se VERB"*, *"se souvenir de"*, *"PART souvenir de"*, *"se VERB de"*, *"PART VERB de"*, etc. Chaque 3-gramme définit un trait du modèle. Par exemple, quelques valeurs de ces fonctions de traits pour chacune des deux phrases t_1 et t_2 sont :

t_1 : $\Phi_{\text{PRON se souvenir}} = 1$, $\Phi_{\text{elle se VERB}} = 1$, $\Phi_{\text{PRON se souvient}} = 0$, ...
t_2 : $\Phi_{\text{PRON se souvenir}} = 1$, $\Phi_{\text{elle se VERB}} = 0$, $\Phi_{\text{PRON se souvient}} = 1$, ...

Ces traits servent à entraîner un classificateur binaire. Le trait (générique) utilisé par le modèle de reclassement est une marge de ce classificateur : pour chaque phrase cible x, $\Phi_{MLF}(x) \in \mathbb{R}$.

[6] Pour plus de détails, se référer à [Mahé and Cancedda, 2008]

Entraînement du modèle de langage discriminatif

Entraînement avec SVM. L'entraînement de MLF proposé dans [Vuorinen et al., 2008] consiste à entraîner un modèle de classification binaire SVM [Vapnik, 1998] sur l'ensemble des exemples positifs (phrases bien formées) et des exemples négatifs (phrases mal formées). Pour que les exemples négatifs soient proches des traductions réelles sur lesquelles nous allons appliquer le modèle de langage discriminatif, la procédure suivante est proposée pour la génération de l'ensemble d'entraînement :

1. générer les listes des N meilleures traductions à partir de la partie source d'un corpus parallèle de taille suffisante ;
2. la partie cible de ce corpus forme un ensemble d'exemples positifs de l'ensemble d'entraînement ;
3. les traductions de la liste des N meilleures traductions ayant les scores NIST les plus bas constituent les exemples négatifs.

La valeur d'un trait intégré dans le modèle de reclassement est définie par la "marge souple"[7] du SVM prédit par le modèle de classification entraîné sur un tel corpus.

Nous proposons dans ce qui suit une variante de cette approche.

Estimation de confiance pour le choix des exemples négatifs. Dans l'approche décrite, on suppose que, pour chaque phrase source, il existe toujours une traduction mal formée. La taille d'une liste de traductions étant fixée à 20, il n'est pas garanti qu'une mauvaise traduction soit présente dans la liste[8]. Il est aussi possible que, pour une phrase source, toutes les 20 traductions soient très mauvaises, alors que pour une autre phrase, toutes les 20 traductions soient de bonnes traductions. Dans le premier cas, toutes les traductions peuvent être prises comme des exemples négatifs, alors que dans le deuxième cas, on préférerait ne pas choisir d'exemple négatif du tout. De plus, la corrélation du score NIST au niveau des phrases avec les jugements humains est très faible.

Plutôt que de choisir une traduction dans la liste des 20 meilleures, nous proposons de calculer une estimation disant s'il existe une bonne traduction, ou s'il existe une mauvaise traduction dans la liste des 20 meilleures traductions proposées par le système. Pour cela, nous allons utiliser la notion de score de confiance d'une traduction définie par [Specia et al., 2009].

Les scores de confiance sont calculés à partir d'un certain nombre de traits dépendant de la phrase source et de sa traduction, et ont montré une meilleure corrélation avec les jugements humains que les scores NIST ou BLEU.

Pour chaque phrase du corpus d'entraînement, un score de confiance est attribué à chaque traduction d'une liste de 20 meilleures traductions. S'il n'existe pas de traduction ayant un score de confiance au-dessous d'un certain seuil ("mauvaises" traductions) dans la liste, aucun exemple négatif n'est extrait, autrement, la traduction ayant un score de confiance minimum est prise comme exemple négatif[9].

[7] C est un paramètre qui permet de contrôler le compromis entre le nombre d'erreurs de classement, et la distance entre l'hyperplan séparateur et l'exemple le plus proche.

[8] Parmi 20 traductions de la liste, il est possible d'avoir certaines traductions ayant les mêmes formes de surface.

[9] Si plusieurs traductions ont le même score de confiance, nous prenons celle qui est le plus bas dans la liste proposée par un système de TA (et qui a par conséquent le score le plus faible attribué par le système).

Chapitre 4

Apprentissage des paramètres avec un Perceptron Structuré

Dans ce chapitre, nous introduisons l'algorithme du Perceptron Structuré que nous avons utilisé pour l'apprentissage des paramètres, et nous décrivons ensuite comment nous l'avons appliqué à notre problème.

4.1 Perceptron Structuré

L'algorithme dit du Perceptron Structuré (ou PS) a été introduit par Collins [Collins, 2002]. C'est un algorithme simple, pouvant exploiter un grand nombre de traits, et très rapide à l'entraînement. Le Perceptron Structuré a été appliqué avec succès à de nombreuses tâches du TALN telles que l'analyse syntaxique [Collins and Roark, 2004], la reconnaissance des entités nommées [Collins, 2001], la modélisation de langages [Roark et al., 2004], la TA probabiliste à fragments [Liang et al., 2006], etc.

4.1.1 Algorithme

Nous suivons l'approche proposée par Collins qui a introduit l'algorithme du Perceptron Structuré dans le cadre de l'analyse syntaxique. La tâche est d'apprendre une fonction qui mette en correspondance des entrées $x \in X$ avec des sorties $y \in Y$. Nous utiliserons les notations suivantes :

- GEN est une fonction qui donne un ensemble de traductions possibles $GEN(x)$ pour une phrase source x ;
- (x_i, y_i^*) est un exemple de l'ensemble d'entraînement, où x_i est une phrase source, et $y_i^* \in GEN(x_i)$ est une traduction de référence ;
- $\Phi : X \times Y \rightarrow \mathbb{R}^K$ associe un vecteur de valeurs de fonctions de traits $\Phi(x, y)$ à chaque $(x, y) \in X \times Y$;
- $\lambda \in \mathbb{R}^K$ est le vecteur des paramètres.

Les composantes GEN, Φ et λ définissent une fonction F mettant en correspondance une phrase source x avec une phrase cible $F(x)$ via l'équation suivante

$$F(x) = argmax_{y \in GEN(x)} \lambda^T \Phi(x, y) \tag{4.1}$$

La tâche de reclassement est alors de trouver une traduction $\hat{y}_i \in GEN(x_i)$ qui maximise le produit scalaire $\lambda^T \Phi(x_i, y_i)$.

L'apprentissage des paramètres λ se fait sur la base d'un corpus parallèle et a pour but de trouver un vecteur de paramètres λ qui décrit au mieux les données d'apprentissage. Autrement dit, dans le cas idéal, $\hat{y}_i = \arg\max_{y \in GEN(x)} \lambda^T \Phi(x_i, y_i)$ doit correspondre à la traduction de référence y_i^*.

Le but de l'apprentissage est de trouver un hyperplan séparant les traductions de référence y_i des autres traductions pour chaque phrase. Il n'est pas garanti qu'une telle séparabilité soit possible en pra-

tique. [Collins, 2002] donne une borne supérieure sur le nombre des erreurs faites par PS après une première itération en cas de non-séparabilité.

Algorithme 2 *Perceptron structuré.*

ENTRÉES : Exemples d'entraînement $\{(x_i, y_i^*)\}_{i\in[1..N]}$
SORTIES : Paramètres $\lambda \in \mathbb{R}^K$
$\lambda_k = 0$, $k \in [1...K]$
pour t de 1 à T **faire**
  pour i de 1 à N **faire**
    $\hat{y}_i = argmax_{y\in GEN(x_i)}\lambda^T\Phi(x_i, y)$
    si $\hat{y}_i \neq y_i^*$ **alors**
      $\lambda = \lambda + \Phi(x_i, y_i^*) - \Phi(x_i, \hat{y}_i)$
    finsi
  fin pour
fin pour
Retourner λ

Le nombre d'itérations T est défini par validation croisée sur des exemples disjoints de l'ensemble d'entraînement.

4.1.2 Version moyenne du Perceptron Structuré

Pour une meilleure performance, nous gardons tous les paramètres intermédiaires λ_i^t obtenus à l'itération t pour une phrase i. Les paramètres finals sont calculés comme la moyenne sur toutes les phrases et toutes les itérations : $\bar{\lambda}_{AVG} = \sum_{i,t} \frac{\bar{\lambda_i^t}}{NT}$. D'après [Freund and Schapire, 1999; Collins, 2002], cette version moyenne du Perceptron Structuré permet de limiter l'effet de surapprentissage et donne une meilleure performance sur le corpus de test.

4.2 Choix de l'objectif dans le cadre de la TA

L'entraînement du type Perceptron demande la présence d'une traduction y_i^* annotée comme la référence parmi les exemples de traductions $GEN(x_i)$ pour chaque phrase source x_i. Cela est nécessaire pour que nous puissions définir les valeurs des fonctions de traits pour y_i^* : $\Phi(x_i, y_i^*)$.

Dans le cas général,la traduction de référence produite par un humain peut ne pas être dans le domaine de Φ. Par exemple, si Φ contient des traits spécifiques au décodeur ou dépend des bi-fragments qui ont été utilisés pour générer une traduction, et si la traduction de référence n'est pas accessible au décodeur, Φ ne peut pas s'appliquer à la traduction humaine.

4.2.1 Notion de pseudo-référence

Dans les cas où la référence n'est pas dans le domaine de Φ, nous définissons un critère pour choisir une meilleure traduction dans la liste des traductions produites par le système de TA. Nous appelons cette meilleure traduction de la liste *une pseudo-référence*, pour ne pas la confondre avec une vraie référence, qui est une traduction produite par un humain. Toute *pseudo-référence* est donc dans le domaine de Φ, et peut être utilisée lors de l'entraînement du perceptron structuré.

Idéalement, on souhaiterait choisir la meilleure traduction au moyen de jugements humains. En pratique, étant donné la taille du corpus d'entraînement (1000 phrases avec une liste de 1000 meilleures traductions pour chaque phrase), il nous semble peu réaliste d'annoter les données d'entraînement manuellement. Nous suivons donc le critère utilisé dans les travaux précédents (Och et al. [2003]; Shen [2004]), et utilisons des mesures automatiques d'évaluation pour choisir la pseudo-référence. Nous essayons différents critères (décrits plus loin), afin de comparer l'impact de chacun des critères sur les

scores automatiques et sur la "qualité" (adéquation, fluidité) de la traduction en termes de jugements humains.

Selon le critère de choix de la pseudo-référence, nous pouvons définir une fonction $Score$ applicable à toute traduction y. $Score(y)$ mesure la similarité entre la traduction y et la référence y^* selon la mesure d'évaluation automatique choisie. La traduction $y = \arg\max_{y \in GEN(x)} Score(y)$ est choisie comme pseudo-référence pour la phrase x.

Afin d'estimer les gains possibles, nous introduisons la notion de score *oracle*, qui est l'estimation du score maximal qui peut être obtenu sur la liste des N meilleures traductions : mesure automatique calculée pour un corpus formé des pseudo-références.

4.2.2 Rayon ε de la pseudo-référence

L'algorithme du PS a pour but de trouver un vecteur des paramètres $\overrightarrow{\lambda} = (\lambda_1, ..., \lambda_k)$ qui sépare une pseudo-référence de toutes les autres traductions pour chaque phrase du corpus d'entraînement. En pratique, plusieurs candidats peuvent être optimaux selon une mesure automatique[1].

De plus, la corrélation des scores automatiques au niveau des phrases avec les jugements humains est assez faible ([Callison-Burch et al., 2009; Specia et al., 2009]). Ainsi, les petites variations des scores automatiques ne sont pas suffisamment fiables pour distinguer une bonne traduction d'une autre moins bonne. Afin de réduire les erreurs dues aux variations trompeuses des scores automatiques, nous considérons que les candidats proches en termes de scores automatiques sont de qualité équivalente.

Plutôt que de chercher à séparer une seule pseudo-référence de toutes les autres traductions de la liste, nous suivons l'idée proposée dans [Vuorinen et al., 2008] et considérons les traductions qui sont à une distance inférieure ou egale à ε de la pseudo-référence comme des exemples positifs, et les traductions qui se trouvent plus loin comme des exemples négatifs. Nous cherchons à apprendre le vecteur des paramètres λ afin de séparer les exemples positifs des exemples négatifs.

Les ensembles d'exemples positifs $G(x)$ et d'exemples négatifs $B(x)$ sont définis comme suit ([Vuorinen et al., 2008]) :

$$PR(x) = \arg \max_{y \in GEN(x)} Score(y)$$

$$G(x) = \{y \in GEN(x) | Score(PR(x)) - Score(y) \leq \varepsilon\}$$

$$B(x) = GEN(x) - G(x)$$

La version modifiée de l'algorithme du perceptron structuré tenant compte du rayon ε est l'algorithme 3 ci-dessous. La valeur de ε et le nombre d'itérations T sont définis par validation croisée sur un ensemble disjoint de l'ensemble d'entraînement. Comme avant, nous utiliserons une moyenne des paramètres λ_{AVG} pour réduire le risque de surapprentissage.

4.3 Critère de choix de la pseudo-référence

Dans nos expériences, nous considérons différents critères pour le choix de la pseudo-référence. Tout d'abord, elle peut être une vraie traduction de référence, dans le cas où les fonctions de traits peuvent être définies pour une vraie référence.
Dans le cas contraire, nous avons besoin de définir un critère pour le choix de la pseudo-référence. Le critère le plus important pour le choix d'une mesure automatique à utiliser pour le choix de la pseudo-référence est sa corrélation avec les jugements humains au niveau des phrases.

[1] Il peut y avoir des traductions représentées par des vecteurs différents dans l'espace des traits, mais ayant la même forme de surface. Cela est dû à des combinaisons différentes de bi-fragments. Par exemple, *the black cat* peut être traduit avec des bi-fragments *(the - le, black - noir, cat - chat)*, ou par des bi-fragments *(the - le, black cat -chat noir)*

Algorithme 3 *ε-sensitive perceptron structuré.*

ENTRÉES : Exemples d'entraînement $\{x_i, G(x_i), B(x_i)\}_{i\in[1..N]}$
SORTIES : Paramètres $\lambda \in \mathbb{R}^K$

```
λ_k = 0, k ∈ [1, K]
pour t = 1..T faire
  pour i = 1..N faire
    ŷ_i = argmax_{y∈GEN(x_i)} λ^T Φ(x_i, y)
    si ŷ_i ∉ G(x_i) alors
      λ = λ + Φ(x_i, PR(x_i)) − Φ(x_i, ŷ_i)
    finsi
  fin pour
fin pour
Retourner λ
```

4.3.1 NIST, BLEU individuels

Afin de pouvoir comparer nos résultats avec des résultats précédents (Och et al. [2003]; Shen [2004]), nous utilisons les mesures automatiques classiques (NIST, BLEU) pour le choix d'une pseudo-référence. Notons que ce type d'entraînement est différent de l'entraînement standard de type MERT [Och, 2003], car l'objectif du perceptron structuré est de placer des phrases ayant des scores individuels plus élevés plus haut dans la liste, mais nous ne tenons pas compte d'un score automatique global lors de l'entraînement (optimisé par MERT), qui peut être différent des scores individuels.

Malheureusement, les mesures automatiques telles que BLEU ou NIST ne montrent pas une très bonne corrélation au niveau des phrases ([Callison-Burch et al., 2009; Specia et al., 2009]). De plus, ces mesures ne sont pas conçues au départ pour mesurer la qualité des traductions individuelles, mais tirent avantage de l'information globale sur le corpus.

4.3.2 wlpBLEU

En s'inspirant des résultats de [Callison-Burch et al., 2009] où les scores introduits par [Popovic and Ney, 2009] ont montré une assez bonne corrélation au niveau des traductions individuelles, nous avons introduit une nouvelle variante du score BLEU, que nous appelons *wlpBLEU*, tenant compte de la forme de surface, du lemme et de la parties de discours de chaque mot. Cette mesure est basée sur la même idée que la mesure BLEU, sauf qu'elle considère non seulement les n-grammes au niveau de surface, mais aussi au niveau des autres facteurs, tels que les parties du discours et les lemmes.

Nous adaptons la précision modifiée définie pour BLEU (équation 2.2) au cas des n-grammes au niveau des formes de surface (ng_w), des lemmes (ng_l) et des parties du discours (ng_p)[2].

$$p_n = \frac{\sum_{c\in C}(\sum_{ng_w\in c} Cooc(ng_w) + \sum_{ng_l\in c} Cooc(ng_l) + \sum_{ng_p\in c} Cooc(ng_p))}{\sum_{c\in C}(\sum_{ng_w\in c} Count(ng_w) + \sum_{ng_l\in c} Count(ng_l) + \sum_{ng_p\in c} Count(ng_p))} \tag{4.2}$$

La mesure wlpBLEU est alors définie par l'équation 2.4 qui intègre cette nouvelle précision. Notons qu'il est également possible d'associer des poids différents aux différents facteurs. Ainsi, par exemple, les n-grammes au niveau de surface peuvent être plus importants (et plus rares) que les n-grammes au niveau des partie du discours. Nous n'avons pas approfondi cet aspect, en attribuant les mêmes poids à tous les facteurs.

Afin de vérifier la corrélation de cette nouvelle mesure d'évaluation avec des jugements humains, nous avons pris un corpus de 4000 phrases traduites par 3 systèmes de traduction (Matrax, Sinuhe, Portage) de l'anglais vers l'espagnol et annotées par des traducteurs professionnels, où à chaque traduction

[2]Notons qu'il s'agit ici des n-grammes d'objets de même type (i.e. des formes de surface, des lemmes ou des parties du discours), contrairement au cas des traits du modèle de langage factoriel décrit dans 3.2.2.2.

TAB. 4.1 – Corrélation du score BLEU et wlpBLEU avec les annotations manuelles des traductions (anglais-espagnol).

	sinuhe	matrax	portage
Corrélation BLEU	0,3499	0,3369	0,3649
Corrélation wlpBLEU	0,3750	0,3686	0,3770

est attribuée une note de 1 à 4. Nous avons calculé la corrélation du score BLEU et du score wlpBLEU avec les scores attribués par ces traducteurs.

Les corrélations pour chaque système sont présentées dans le tableau 4.1. Nos expériences montrent que la corrélation du score wlpBLEU avec les annotations manuelles reste assez faible, bien qu'elle soit systématiquement meilleure que la corrélation des scores BLEU individuels.

Chapitre 5

Résultats du reclassement

5.1 Expériences

5.1.1 Protocole d'évaluation

Pour toutes les expériences, nous avons généré pour chaque phrase source une liste des 1000 meilleures traductions avec le système de base. Le modèle de reclassement est entraîné sur cette liste dans des cadres différents :

– en choisissant une pseudo-référence avec l'une des mesures d'évaluation automatique (BLEU, NIST, wlpBLEU) : ce type d'entraînement est possible pour tous les ensembles de traits que nous souhaitons utiliser, les paramètres du perceptron structuré (ε et le nombre d'itérations T) sont choisis par une validation croisée en 5 passes (*5-fold*).

– en ajoutant une vraie traduction de référence à la liste des traductions et en l'utilisant lors de l'entraînement du modèle de reclassement : cela est possible uniquement pour des ensembles de traits qui peuvent être calculés pour la traduction de référence (traits de couplage et de cohésion) ; le perceptron structuré ne contient alors qu'un paramètre, le nombre d'itérations, qui est choisi, comme dans le cas précédent, par une validation croisée en 5 passes.

5.1.1.1 Systèmes de base

Moses. La plupart des expériences conduites pour le reclassement ont été faites en utilisant le système de base décrit dans la section 2.2.3.1 basé sur le système de TA probabiliste à fragments *Moses*. *Moses* contient un paramètre permettant de générer une liste de traductions distinctes : ainsi, si une même traduction peut être obtenue par des compositions différentes des fragments, seule la première ayant le score le plus élevé est considérée. Cette option permet de générer des listes de traductions plus variées.[1]
Moses fournit des informations sur les alignements lexicaux et au niveau des fragments, et les valeurs des traits du modèle de base (section 1.2.4) pour chaque candidat de traduction.

Pour entraîner le modèle de reclassement, nous avons utilisé le corpus *test2007* (tableau 2.2) : ce corpus avec une liste des 1000 meilleures traductions constitue le corpus d'entraînement du modèle de reclassement. Le modèle de reclassement défini par un ensemble de traits est entraîné sur les 1000 premières phrases de ce corpus (*dev1*) ; la deuxième moitié de ce corpus (*dev2*) est gardée à part pour l'utiliser pour des entraînements supplémentaires.

[1] Une liste de 1000 meilleures traductions générées ainsi contient environ 800 traductions distinctes par phrase en moyenne. La génération d'une liste de 1000 meilleures traductions dans le mode standard ne génère qu'environ 250 traductions distinctes par phrase.

Sinuhe. Les expériences sur l'intégration du modèle de langage discriminatif factoriel (MLF) ont été conduites dans le cadre du projet européen SMART[2] en utilisant le système de TA probabiliste à fragments Sinuhe [Kääriäinen, 2009]. La différence principale de Sinuhe par rapport aux autre systèmes de TA probabiliste à fragments est l'utilisation de fragments qui se chevauchent lors du décodage.

Les expériences avec les traits de MLF ont été faites en utilisant des listes de traductions générées par le système de base (Sinuhe entraîné sur Europarl anglais-espagnol). Le système de base a été entraîné sur un corpus parallèle dont la partie cible a été analysée par XIP afin d'obtenir les formes des lemmes et des parties de discours pour chaque *token* cible. Ainsi, lors de la traduction, la phrase source simple est transformée en une phrase cible enrichie (chaque token cible est représenté par un vecteur de facteurs : forme de surface, lemme, partie du discours).
Nous avions deux ensembles de traductions à notre disposition :

- *dev2* : 2000 phrases parallèles, avec une liste des 1000 meilleures traductions pour chaque phrase ; ce corpus a été utilisé pour l'entraînement du modèle de reclassement ;
- *dev3* : 100 000 phrases parallèles avec une liste des 20 meilleures traductions pour l'entraînement de MLF.

Pour pouvoir comparer l'impact du reclassement avec les traits MLF et les autres traits (monolingues et bilingues), nous avons divisé le corpus *dev2* en deux parties, en utilisant les 1000 premières phrases pour l'entraînement du modèle de reclassement, et les 1000 dernières phrases pour l'évaluation du modèle.

Les listes de traductions contiennent des traits du modèle de base :

- les traits de base de Sinuhe sont similaires à ceux de Moses : modèle de traduction à fragments, modèle de langue, modèle de traduction lexical inversé, nombre des mots cible, distorsion.

Les listes des traductions produites par Sinuhe ne contiennent aucune information sur les alignements, donc nous avons utilisé l'outil GIZA++ pour entraîner le modèle d'alignement sur la partie parallèle du corpus Europarl (utilisée pour l'entraînement du modèle de traduction) et pour établir des alignements lexicaux.

5.1.1.2 Évaluation des résultats

Nous faisons d'abord une évaluation détaillée de différents modèles de reclassement à l'aide des mesures d'évaluation automatique. Nous tenons compte des scores NIST et BLEU globaux et individuels.

Pour les évaluations manuelles, nous avons extrait deux échantillons de 150 phrases du corpus de test : l'un a été utilisé pour l'évaluation de l'adéquation et l'autre pour l'évaluation de la fluidité des traductions.

Nous avons formé l'union des traductions proposées par différents modèles de reclassement (un modèle de reclassement étant défini par un ensemble de fonctions de traits et un type de pseudo-référence), la traduction de base proposée par le système, et les pseudo-références maximisant les scores NIST, BLEU ou wlpBLEU. Nous avons fourni la liste des traductions ainsi obtenue aux juges, en leur demandant de classer les traductions par ordre de préférence (adéquation ou fluidité). Les juges pouvaient juger des traductions différentes comme étant de même qualité, en leur attribuant le même rang de préférence.

Bien que la comparaison directe de chaque traduction de la liste avec la traduction de base puisse être établie à partir des rangs attribués par les juges, ce type d'évaluation est un peu différent des comparaisons directes, où le juge n'a que 2 traductions à comparer. Les comparaisons deux à deux établies à partir des rangs des traductions peuvent alors être biaisées (par l'ordre dans lequels les traductions ont été évaluées). Cependant, la comparaison simultanée de toutes les traductions nous permet non seulement de comparer les traductions proposées par différents modèles de reclassement avec une traduction produite par un système de base (*baseline*), mais aussi comparer les différents modèles de reclassement entre eux.[3]

[2] `http://www.smart-project.eu`

[3] Nous avons opté pour ce type d'évaluation, plutôt que pour l'attribution de notes à chaque traduction indépendamment, pour les raisons suivantes. Beaucoup de traductions d'une liste des n-meilleures traductions sont très proches entre elles. Les changements

Dans le cas d'évaluation de l'adéquation, la phrase source et la traduction de référence sont montrées aux juges, tandis que, pour l'évaluation de la fluidité, seules les traductions provenant de modèles différents sont visibles.

5.1.2 Reclassement avec des traits bilingues

5.1.2.1 Traits de couplage et traits du modèle de base

La première série d'expériences a été faite avec des traits de couplage. Nous produisons des alignements lexicaux à l'aide du modèle GIZA++ entraîné sur le corpus d'entraînement. Ce type d'alignement peut être établi pour une traduction de référence également, et par conséquent les traits de couplage peuvent être calculés pour les traductions de référence.
Les traits de couplage utilisés pour cette série d'expériences sont les suivants :

Couplage générique : les traits correspondent au décompte simple (non pondéré) des rectangles, et à des variantes normalisant ce décompte (précision, rappel, F-mesure) ;

Couplage lexical : les traits correspondent au décompte des rectangles (s_i, s_j, t_k, t_l) pondéré par les probabilités lexicales $p(s_i|t_k)$ et $p(t_k|s_i)$. Ces probabilités viennent de la table de traductions apprise par GIZA++ sur le corpus parallèle lors de l'entraînement du modèle de traduction.

Couplage étiqueté : la collection des traits est formée de traits tenant compte des paires des étiquettes de dépendance ;

En plus des traits de couplage, nous introduisons les traits du système de base dans le modèle de reclassement (*moses, sinuhe*).

5.1.2.2 Intégration des traits de couplage étiqueté avec les autres traits

Nous avons implementé 2 possibilités d'intégrer les traits de *couplage étiqueté* :

– la première possibilité consiste à introduire les traits de *couplage étiqueté* directement dans le module de reclassement (comme c'était fait dans la série d'expériences précédentes) et à apprendre simultanément les poids correspondant (sur un corpus *dev1*) : c'est une solution directe ne demandant pas d'entraînement supplémentaire ; en revanche, le vecteur des paires d'étiquettes peut être d'une taille importante (~1000 pour un corpus de 1000 phrases avec une liste de 1000 traductions pour chaque phrase), et la fusion directe avec des traits divers peut mener à un surapprentissage.

– la deuxième possibilité consiste à introduire un trait similaire au trait de décompte de couplage générique, tenant compte de la force de couplage entre des dépendances ayant des étiquettes spécifiques. La force de couplage sera définie par les poids des paires d'étiquettes qui correspondent aux traits de *couplage étiqueté*. Ainsi, l'entraînement se fait en plusieurs étapes :

1. entraînement du perceptron structuré uniquement avec des traits de *couplage étiqueté* (sur un corpus *dev2*). Cet entraînement donne les poids $\lambda^1_{lab_s\_lab_t}$ pour des paires des étiquettes $lab_s\_lab_t$ trouvés dans le corpus *dev2*.
2. calcul d'un nouveau trait correspondant à un *couplage étiqueté* avec le modèle appris ; ce nouveau trait $\phi_{lab_{gen}}$ (nous allons l'appeler *couplage étiqueté générique*) correspond aux décomptes pondérés des rectangles avec les poids appris à l'étape précédente :

$$\phi_{lab_{gen}} = \sum_{lab_s\_lab_t \in dev2} \lambda^1_{lab_s\_lab_t} \cdot f_{lab_s\_lab_t} \tag{5.1}$$

Notons ici que seuls les poids des paires $lab_s\_lab_t$ se trouvant dans le corpus *dev2* seront considérés lors des calculs du trait étiqueté générique.

apportés par le reclassement étant limités à cette liste, ne sont souvent pas très importants. Nous espérons que la comparaison directe des traductions force le juge à se concentrer sur la différence entre les différentes traductions de la liste, et à évaluer cette différence (comme positive, négative ou nulle), plutôt que d'évaluer la traduction elle-même. Ce type d'évaluation nous permet d'évaluer le modèle de reclassement, sans évaluation implicite du système de base.

3. entraînement du modèle de reclassement par perceptron structuré avec le trait de *couplage étiqueté générique* $\phi_{lab_{gen}}$ ajouté aux autres traits sur un corpus *dev1*.

Le trait de *couplage étiqueté générique* peut être entraîné dans les deux modes suivants :
– seuls les traits étiquetés sont utilisés, et leurs poids sont appris lors de l'entraînement avec une traduction de référence,
– les traits étiquetés entraînés avec les traits du modèle de base. Nous ajoutons les traits de base dans ce cas de figure car les résultats des premières expériences (tableaux B.1, B.2[4]) ont montré que la présence des traits de base a une influence importante sur les valeurs des scores automatiques ; nous espérons ainsi que les poids des étiquettes appris dans ce cadre seront de meilleure qualité. Seuls les poids des traits de couplage étiquetés sont utilisés pour calculer la valeur du trait générique à l'étape suivante.

5.1.2.3 L'influence de l'alignement

Les fonctions de couplage dépendent directement des alignements lexicaux. Nous avons expérimenté les trois types d'alignement introduits au 3.1.1.3 :
– alignements lexicaux fournis par le système (Moses) ;
– alignement au niveau des fragments fournis par le système ;
– alignement lexicaux produits par GIZA++.

Cette série d'expériences a été faite uniquement pour des traductions produites par Moses et avait pour but de voir l'impact de la qualité d'alignement sur les résultats du reclassement.

5.1.3 Reclassement avec des traits monolingues

5.1.3.1 Traits de cohésion

Les traits de cohésion/non-cohésion linguistique ont été introduits comme dans la section 3.2.1. Nous avons choisi de prendre en compte la cohésion pour les traits morphologiques suivants en fonction de la langue cible :
– anglais : nombre (singulier, pluriel), personne (p1, p2, p3) ;
– français : nombre (singulier, pluriel), personne (p1, p2, p3), genre (masculin, féminin) ;
– espagnol : nombre (singulier, pluriel), personne (p1, p2, p3), genre (masculin, féminin).

L'ensemble des traits de cohésion/non-cohésion sont intégrés soit directement, soit en passant par un trait générique similaire au trait de couplage générique étiqueté.

5.1.3.2 Modèle de langage factoriel discriminatif (MLF)

Nous avons expérimenté différents entraînements possibles du trait de MLF :
– MLF (NIST) : entraînement par SVM, prenant les phrases avec le score NIST le plus bas de la liste comme exemples négatifs.
– MLF (CE) : entraînement par SVM, prenant une traduction ayant le score de confiance le plus bas de la liste comme exemple négatif uniquement si son score de confiance est inférieur à 2,7 (sur une échelle de 0 à 4).

Le paramètre de "marge souple" C de SVM est défini par validation croisée.

5.1.4 Reclassement avec les traits de couplage et les traits monolingues

La combinaison des traits de couplage avec des traits monolingues doit être intéressante car ces traits visent à contrôler des aspects différents de la traduction. Nous choisissons les traits de couplage les plus représentatifs (en nous basant sur les résultats d'évaluation automatique des expériences précédentes) et à les combiner avec des sous-ensembles différents de traits monolingues. Un modèle de reclassement pour chaque type d'alignement et chaque sous-ensemble des traits est alors entraîné.

[4]Les tableaux se trouvent dans l'annexe B, page 117.

5.2 Résultats

5.2.1 Évaluation automatique

5.2.1.1 Reclassement avec traits bilingues

Les résultats de cette première série d'expériences sont présentés dans les tableaux et sur les figures suivants :

- reclassement avec différents traits de couplage et traits de base (basés sur des alignements GIZA++) : figure 5.1, tableaux B.2, B.1, B.3 ;
- intégration des traits de couplage (direct ou générique) étiquetés avec d'autres traits : figure 5.2, tableaux B.4, B.5 ;
- comparaison d'alignements de différents types (fournis par le système au niveau des mots ou au niveau des fragments, basés sur GIZA++) : figures 5.3 ; tableaux B.6, B.7.

Tout d'abord, notons l'importance des traits du modèle de base dans le modèle de reclassement (figure 5.1). Les traits de couplage ne tiennent pas compte de la fluidité de la phrase cible (sans compter le biais éventuel sur la structure produite par XIP), tandis que le modèle de langage n-gramme (un des traits de base) permet un certain contrôle de la fluidité. Les traits du modèle de base contiennent le trait de modèle de langage, et donc fournissent cette information au modèle. De plus, les traits de base contiennent d'autres informations relatives à la génération de la traduction, qui complètent les traits de couplage introduits et contrôlent mieux la qualité de la traduction.

Notons également que les scores individuels[5] sont systématiquement améliorés sur le corpus de test (tableaux B.1, B.2 , B.3), ce qui confirme que l'entraînement a été efficace, améliorant la moyenne des scores individuels. Cela n'est pas toujours le cas pour les scores globaux NIST et BLEU. Bien que NIST s'améliore, BLEU n'est amélioré que dans des cas très rares (pour la traduction vers l'anglais). Cela s'explique par le fait que l'optimisation du modèle de base a été faite en optimisant le score BLEU global (dans le cas de Moses et Sinuhe) : il est donc plus facile d'améliorer un score NIST qui n'a pas été optimisé préalablement.

Le trait de couplage étiqueté semble avoir l'impact le plus important sur les scores automatiques par rapport aux autres traits de couplage (figure 5.1, tableaux B.1, B.2 B.3).

L'intégration du trait de couplage générique étiqueté n'améliore pas les scores quand il est intégré uniquement avec les traits de base (figure 5.2). Son intégration à un ensemble de traits plus important (traits de couplage et traits de base) donne systématiquement de meilleurs scores (individuels et globaux) par rapport à l'intégration directe du trait de couplage étiqueté (c'est surtout vrai dans le cas de la traduction vers l'anglais).

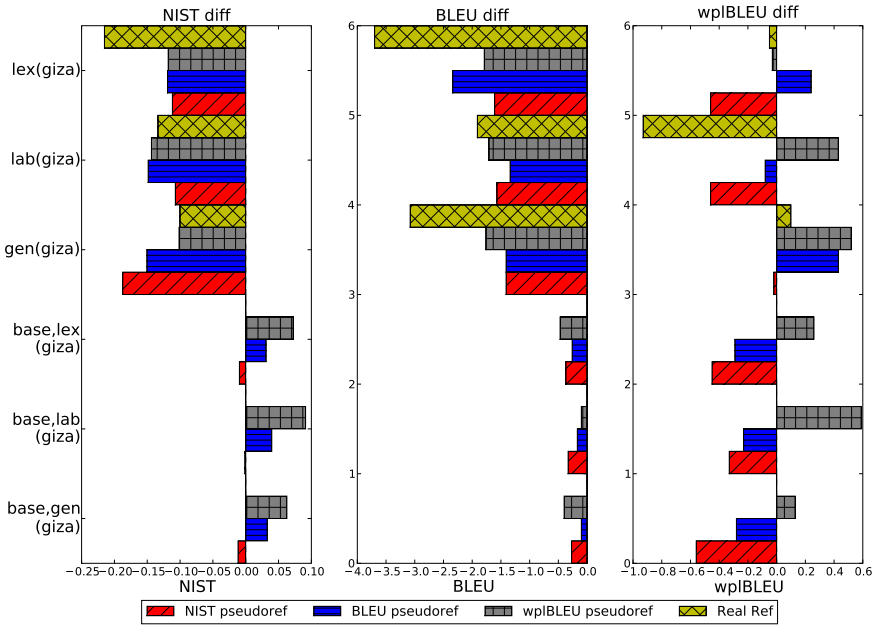
Le trait générique entraîné en utilisant une pseudo-référence avec des traits de base semble être plus utile (sur la base des évaluations automatiques) que le trait générique entraîné sur la base d'une vraie référence.

Nous avons fait une observation surprenante : dans le cas de la traduction vers l'anglais, l'entraînement avec la pseudoréférence wlpBLEU donne systématiquement de meilleurs scores NIST que le même modèle entraîné avec la pseudoréférence NIST.

En ce qui concerne le type d'alignement : la différence entre les alignements lexicaux produits par GIZA++ et ceux fournis par le système est peu visible. Les alignements au niveau des mots fournis par le système semblent donner des scores un peu meilleurs, mais la différence est faible et elle ne permet pas de conclure.

[5] La somme des scores (BLEU, NIST) au niveau des phrases individuelles pour toutes les phrases du corpus de test.

Français - anglais



Anglais - français

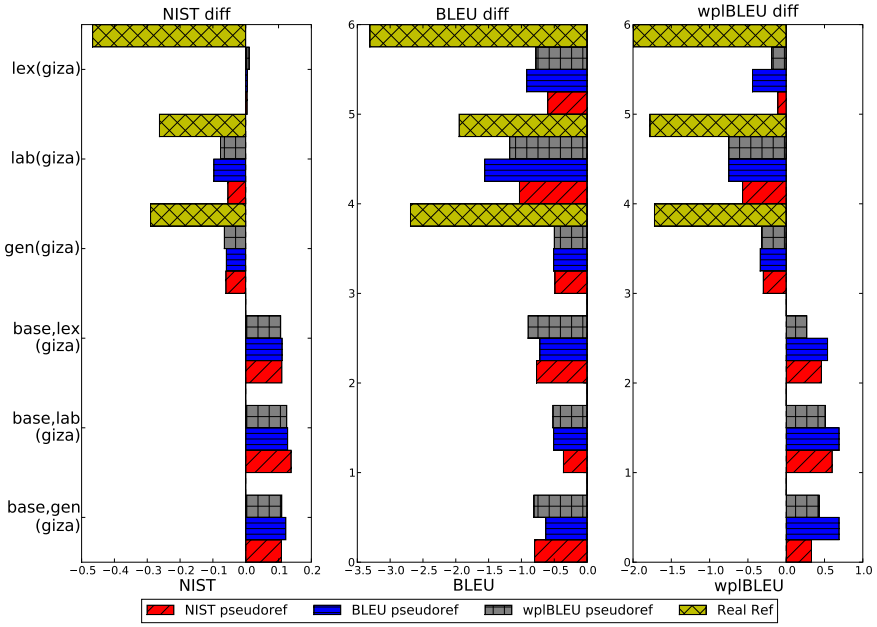
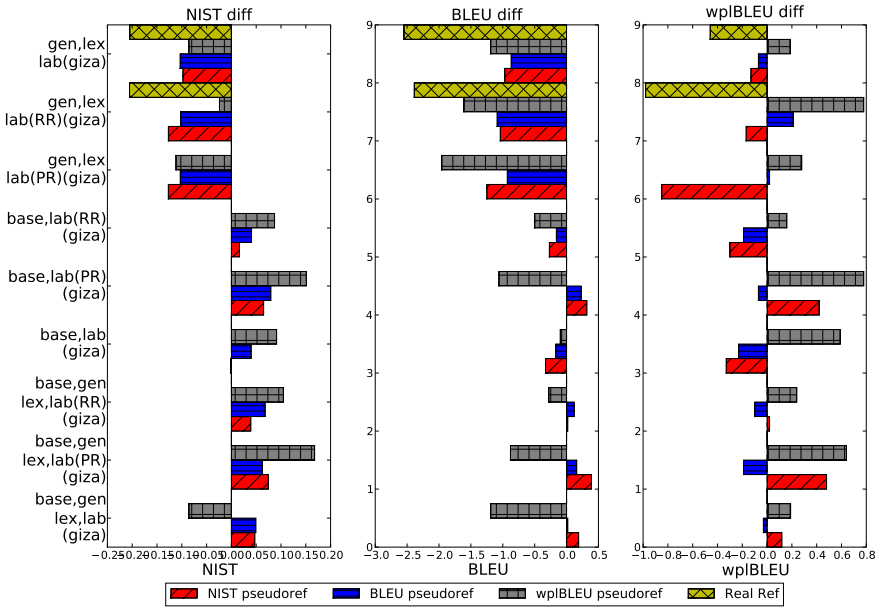


FIG. 5.1 – Résultats d'évaluation automatique du reclassement avec des traits de couplage et des traits de base. À gauche : l'ensemble des traits utilisé pour le reclassement. Base : les traits du modèle de base (Moses), gen : traits de couplage générique, lex : couplage générique lexical, lab : couplage étiqueté.

Français - anglais



Anglais - français

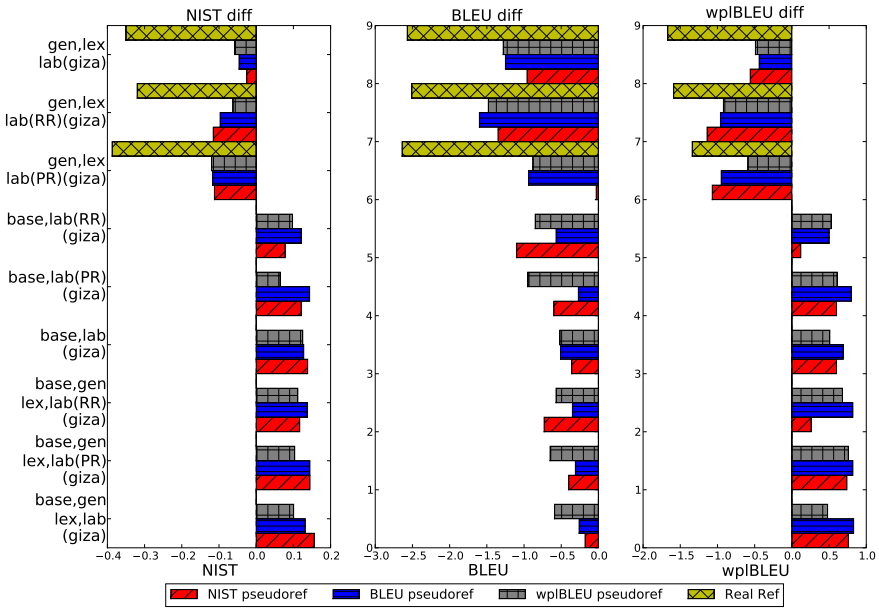
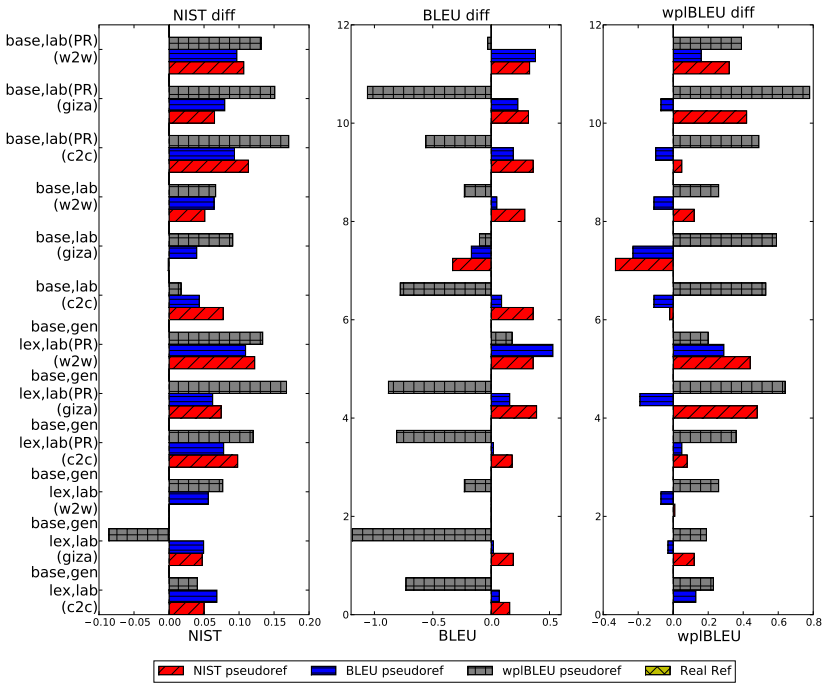


FIG. 5.2 – Résultats d'évaluation automatique de l'intégration du trait de couplage étiqueté. lab : intégration directe des traits de couplage étiqueté, lab(PR) : intégration du trait de couplage étiqueté générique, entraîné avec une pseudo-référence, lab(PR) : intégration du trait de couplage étiqueté générique, entraîné avec une vraie référence.

Français - anglais



Anglais - français

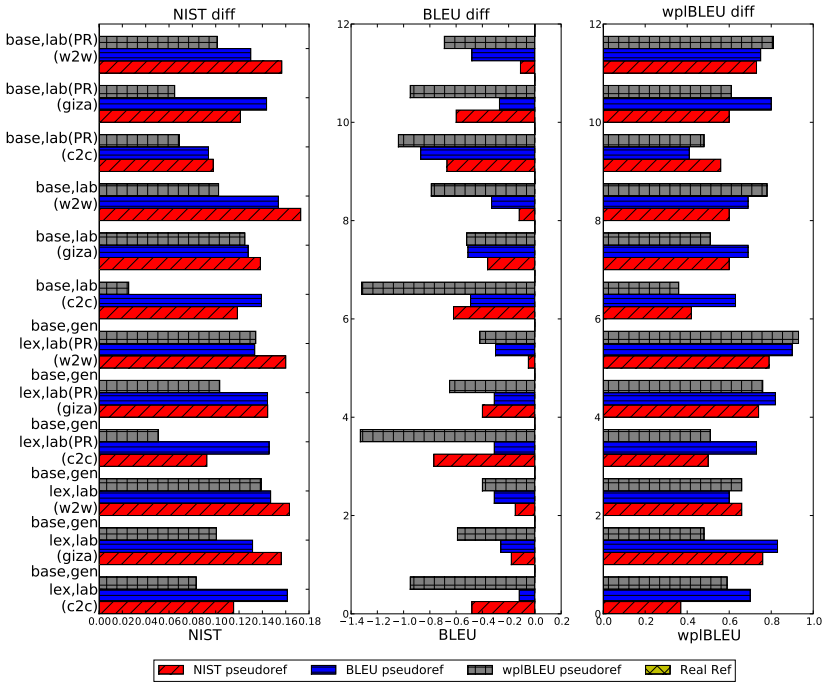


FIG. 5.3 – Influence de l'alignement sur les résultats d'évaluation automatique du reclassement avec les traits du couplage. w2w : alignement au niveau des mots fourni par le système de base, c2c : alignement au niveau des fragments fournis par le système de base, giza : alignement au niveau des mots produit par GIZA++.

5.2.1.2 Reclassement avec des traits monolingues

Les résultats du reclassement avec des modèles monolingues sont présentés sur la figure 5.4, et dans les tableaux :

- B.8, B.9 pour les traits de cohésion et non-cohésion ;
- B.10 pour le reclassement avec le modèle de langage factoriel.

L'intégration du trait de cohésion générique semble donner d'assez bons résultats (en termes des scores automatiques) dans le cas de la traduction du français vers l'anglais. Par contre, ce n'est pas le cas pour la traduction vers le français. Probablement, la raison qui explique cela est le fait que la taille du vecteur des traits de cohésion est plus importante dans le cas du français, la tâche d'apprentissage est donc plus complexe.

Notons également que l'intégration du trait de cohésion générique entraîné sans utilisation des traits de base mais en utilisant une vraie traduction de référence apporte plus d'amélioration par rapport au trait de cohésion générique entraîné en utilisant une pseudoréférence. Nous avons observé un effet contraire lors de l'intégration des traits de couplage étiqueté. Essayons de comprendre la raison de cette différence de comportement.

Les traductions générées par un système de TA sont généralement plus proches structurellement d'une phrase source, tandis que les traductions produites par un traducteur humain sont souvent moins littérales. Il est intuitivement clair qu'il doit être plus facile pour GIZA++ d'aligner une phrase source avec sa traduction générée par un système automatique, que d'aligner la même phrase source avec sa traduction de référence. De plus, quand il s'agit des traductions générées par un système de TA probabiliste, le vocabulaire utilisé par le système est celui du corpus parallèle aligné qui a été utilisé pour entraîner le modèle d'alignement de GIZA++.

Ainsi, des erreurs d'alignement sont possibles lors de l'alignement des phrases source avec des vraies références, ce qui mène aux traits de couplage bruités. Par conséquent, la valeur du couplage générique étiqueté est moins fiable dans le cas d'entraînement du reclassement avec une vraie référence. Dans le cas d'utilisation de la pseudo-référence, la présence des traits de base donne un contrôle supplémentaire et permet de compenser les erreurs d'alignement.

Le trait de cohésion, en revanche, ne dépend pas des alignements. La traduction de référence doit respecter la cohérence de structure, tandis que cela n'est pas nécessairement vrai pour une traduction de pseudo-référence. Ainsi, les trait génériques de cohésion et de non-cohésion entraînés avec une vraie référence sont plus fiables.

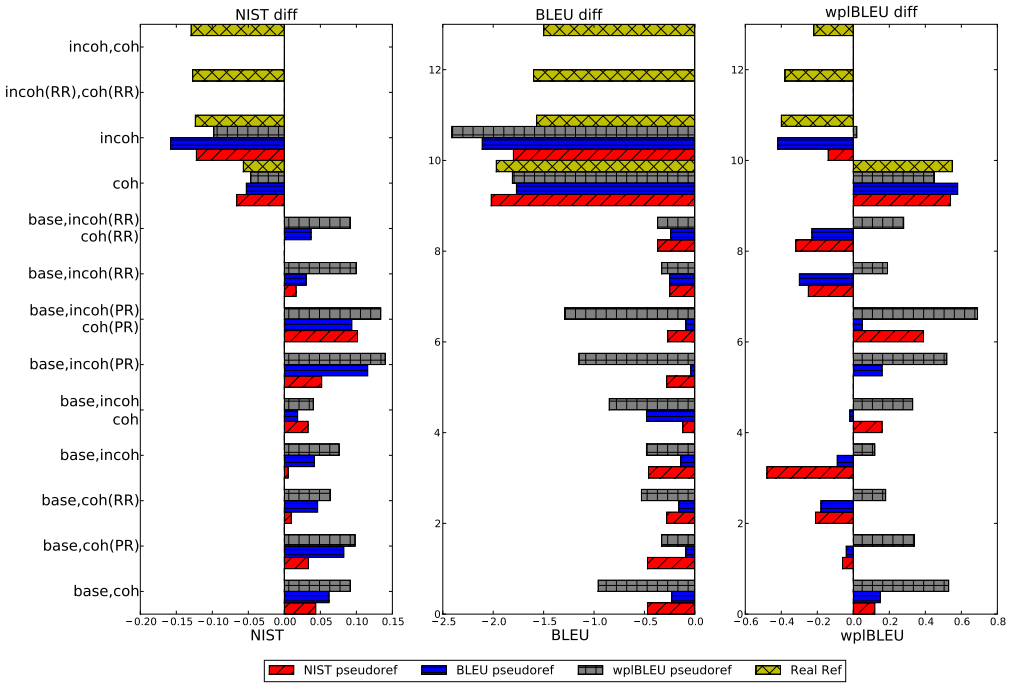
D'après les évaluations automatiques (tableau B.10), le trait de modèle de langage factoriel donne plus d'améliorations (en termes de scores NIST et wlpBLEU) que les traits de cohésion. L'utilisation du trait MLF appris en utilisant l'estimateur de confiance donne systématiquement de meilleurs scores automatiques, bien que la différence soit très faible.

5.2.1.3 Reclassement avec les traits de couplage et les traits monolingues

Les résultats sur l'intégration des traits de couplage avec des traits monolingues sont présentés dans les tableaux B.11, B.12, B.13, B.14, B.15, B.16.

Comme espéré, ce type de modèle a le plus de capacité d'améliorer les résultats. Comme précédemment, le trait générique de cohérence entraîné avec une vraie référence est plus intéressant que le même trait entraîné avec une pseudo-référence. Nous avons utilisé uniquement le trait de couplage générique étiqueté entraîné avec une pseudo-référence, suite aux résultats des expériences précédentes.

Français - anglais



Anglais - français

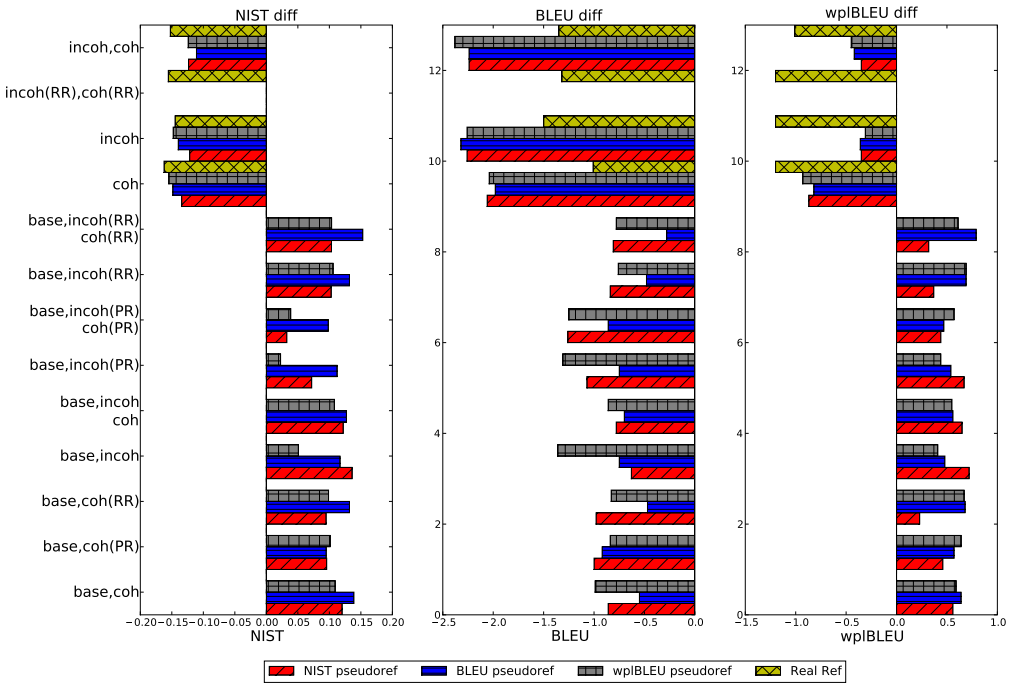


FIG. 5.4 – Résultats d'évaluation automatique du reclassement avec des traits de cohésion et traits de base.

TAB. 5.1 – Nombre de jugements obtenus par tâche

	en-fr		fr-en	
	adéquation	fluidité	adéquation	fluidité
phrases	112	101	81	51
jugements	1920	1758	1622	689
jugements multiples	405	488	410	0

5.2.2 Évaluations manuelles

En nous basant sur des résultats des évaluations automatiques, nous avons choisi les modèles de reclassement suivants pour des évaluations manuelles :

- les modèles de reclassement avec une *pseudo-référence* et utilisant les traits de base et :
  - traits de couplage uniquement (couplage générique, couplage lexical, couplage étiqueté générique (entraîné avec une *pseudo-référence*)),
  - les traits de cohésion ou non-cohésion uniquement (trait générique (entraîné avec une vraie référence ou une pseudo-référence) de cohésion, de non-cohésion),
  - les traits de couplage et les traits de cohésion/non-cohésion ;
- les modèles de reclassement avec une *vraie référence* utilisant :
  - les traits de couplage uniquement (couplage générique, couplage lexical, couplage étiqueté générique (entraîné avec une *vraie référence*)),
  - les traits de cohésion ou non-cohésion uniquement (trait générique (entraîné avec une *vraie référence*) de cohésion, de non-cohésion),
  - les traits de couplage et les traits de cohésion/non-cohésion.

Au total, nous avons une liste de traductions contenant 39 modèles de reclassement. En plus des modèles de reclassement, la liste des traductions proposées pour les évaluations manuelles contient une traduction du modèle de référence, et des pseudo-références (BLEU, NIST et wlpBLEU). Étant donné que des modèles différents peuvent correspondre aux mêmes traductions de la liste, nous avons obtenu environ 15 traductions par phrase pour une tâche de traduction du français vers l'anglais, et environ 13 traductions par phrase pour la traduction de l'anglais vers le français.

5.2.2.1 Accord entre les juges

En tout, 22 juges ont participé aux évaluations manuelles. Chaque juge avait reçu un échantillon de 10-15 phrases avec 10-15 traductions chacune (certains juges ont participé à plusieurs tâches). Les jugements de fluidité ont été faits principalement par des locuteurs maternels de la langue cible. Afin de pouvoir analyser l'accord entre les juges, certains échantillons ont été évalués plusieurs fois par des juges différents. Le tableau 5.1 résume le nombre de jugements que nous avons recueillis pour chaque tâche.

Le coefficient κ [Cohen, 1960] est utilisé pour mesurer l'accord entre deux juges et est défini comme suit :

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{5.2}$$

où $P(A)$ est la proportion où les 2 juges sont d'accord entre eux, et $P(E)$ est la proportion d'accords que les deux juges auraient par hasard.

Pour chaque paire de traductions (t_1, t_2), les jugements suivants sont possibles :

- t_1 est jugée meilleure que t_2,
- t_1 est jugée moins bonne que t_2,
- t_1 est jugée équivalent à t_2.

Si nous supposons que ces trois cas sont équiprobables, l'accord par hasard entre les deux juges est défini par $P(E) = 0.333$. Landis and Koch [1977] ont proposé un classement de l'accord en fonction de la valeur de κ, présenté dans le tableau 5.2.

Ainsi, les résultats d'accord du tableau 5.3 peuvent être interprétés comme un mauvais accord entre les juges.

TAB. 5.2 – Degré d'accord et valeur de κ

Accord	κ
Excellent	0,81
Bon	0,80 - 0,61
Modéré	0,60 - 0,41
Médiocre	0,40 - 0,21
Mauvais	0,20 - 0,0
Très mauvais	< 0,0

TAB. 5.3 – L'accord total entre les juges (3 types de jugements sont pris en compte : t_1 est jugée meilleure que t_2, t_1 est jugée moins bonne que t_2, t_1 est jugée équivalente à t_2).

	en-fr		fr-en	
	adéquation	fluidité	adéquation	fluidité
$P(A)$	0,37	0,42	0,37	-
κ	0,05	0,13	0,06	-

Ce faible accord est lié tout d'abord à la façon dont le problème a été posé. Effectivement, la tâche de classement des traductions était très difficile pour les juges. Bien qu'on demandé aux juges ont été demandés de se concentrer sur les différences entre les traductions, ces différences sont souvent très subtiles et il est difficile d'évaluer leur impact sur la qualité de la traduction finale. De plus, quand la traduction est assez mauvaise, les jugements sur les différences deviennent encore plus difficiles.

En effet, en analysant les désaccord entre les juges, nous avons constaté que la plupart de désaccords (67% pour l'adéquation, et 77% pour la fluidité) sont de type suivant : un juge considère une traduction t_1 équivalente à une traduction t_2, tandis que l'autre juge considère que t_1 est meilleure que t_2. En revanche, les situations où un juge considère t_1 meilleure que t_2, tandis que l'autre juge considére le contraire, sont rares.

Quelques exemples de jugements (du français) se trouvent dans l'annexe C.

5.2.2.2 Résultats de l'évaluation manuelle

Les résultats des évaluations manuelles sont présentés dans les figures 5.5 (anglais-français) et 5.6 (français-anglais). La partie barrée du milieu correspond aux traductions jugées équivalentes à la traduction de base, la partie à gauche correspond à des traductions jugées moins bonnes, et la partie à droite correspond à des traductions jugées meilleures que les traductions du système de base. Les traits marqués par "++" ou "- -" indiquent des modèles donnant des résultats significativement meilleurs (moins bons)[6] par rapport à la traduction du système de base.

Nous voyons que la proportion des traductions améliorées par les modèles de reclassement utilisant la vraie référence est plus importante que celle des modèles utilisant la pseudo-référence.

Les résultats de l'évaluation manuelle montrent que, contrairement à ce que les résultats de l'évaluation automatique nous ont montré, le reclassement avec une vraie référence mène systématiquement à des améliorations (significatives) de la fluidité et de l'adéquation de la traduction par rapport au modèle de base.

Le reclassement avec une pseudo-référence des traductions vers le français permet d'améliorer significativement la fluidité des traductions, mais dégrade dans certains cas l'adéquation (principalement pour des modèles de reclassement avec des traits monolingues).

Concernant le reclassement avec les modèles dépendant de la pseudo-référence des traductions vers l'anglais, la fluidité de la traduction est améliorée avec des modèles de reclassement utilisant uniquement

[6] Selon le test des signes, p-valeur = 0.01 (annexe D).

des traits monolingues. L'adéquation est améliorée par pratiquement tous les modèles, surtout dans le cas d'utilisation d'une pseudo-référence wlpBLEU.

5.3 Conclusions sur le reclassement

Nous avons tiré quelques conclusions de ces premières expériences sur le reclassement.

Tout d'abord, la notion de couplage s'est montrée utile dans le cadre du reclassement, plutôt en termes d'évaluations manuelles qu'en terme d'évaluations automatiques. La variante du couplage la plus utile (en terme d'évaluation automatique) est le couplage étiqueté.

Le reclassement avec une vraie référence s'est montré significativement meilleur que le modèle de base en termes de jugements humains, mais a mené à des dégradations importantes pour les mesures automatiques. Notons que les traductions de référence pour le corpus de test sont souvent peu littérales (nous avons une seule traduction de référence par phrase), ce qui peut expliquer en partie la faible corrélation avec les jugements humains.

Dans le cas de la traduction vers l'anglais, l'utilisation d'une pseudo-référence wlpBLEU semble donner de meilleurs résultats non seulement en termes d'évaluations manuelles, mais aussi en termes de scores automatiques. Ainsi, nous avons obtenu des scores NIST plus élevés en utilisant la pseudo-référence wlpBLEU qu'en utilisant la pseudo-référence NIST.

Cet effet est probablement lié à la meilleure corrélation de wlpBLEU avec la bonne formation syntaxique de la traduction (par rapport à la mesure NIST). Et de même, wlpBLEU a une meilleure corrélation avec les traits de couplage. La mesure NIST, n'ayant pas une bonne corrélation avec la bonne formation syntaxique, mène à des données d'apprentissage plus bruitées pour le perceptron, et rend l'apprentissage des paramètres plus difficile. L'utilisation d'une mesure automatique ayant une meilleure corrélation avec des jugements humains (par exemple, l'estimation de confiance de Specia et al. [2009]) permettrait probablement d'obtenir de meilleurs résultats du reclassement.

Nous avons ensuite étudié la capacité des listes de N-meilleures traductions en nous basant sur les résultats de l'évaluation manuelle. Nous voulions étudier quelle est la proportion des traductions pour lesquelles il existe une meilleure (en termes des jugements humains) traduction dans la liste. La liste des traductions dans ce cas est limitée aux traductions données par les modèles de reclassement choisis pour l'évaluation manuelle, et des pseudo-références choisies par chacune des mesures automatiques.

TAB. 5.4 – Évaluation du potentiel de la liste des N meilleures traductions.

	fr-en	en-fr
déjà bonne traduction	10%	10%
pas d'amélioration possible	6%	6%
Nombre d'améliorations obtenues par le meilleur modèle de reclassement		
fluidité	65%	47%
adéquation	55%	42%

Nous avons choisi les phrases pour lesquelles il n'existe pas de meilleure traduction que la traduction de base dans cette liste limitée. En regardant une liste entière de traductions, nous avons étudié quelle en est la raison :

- soit la traduction de base est déjà une bonne traduction et ne peut pas être améliorée ;
- soit il n'existe pas de traduction qui pourrait améliorer la traduction de base parmi toutes les 1000 traductions proposées par le système de base ;

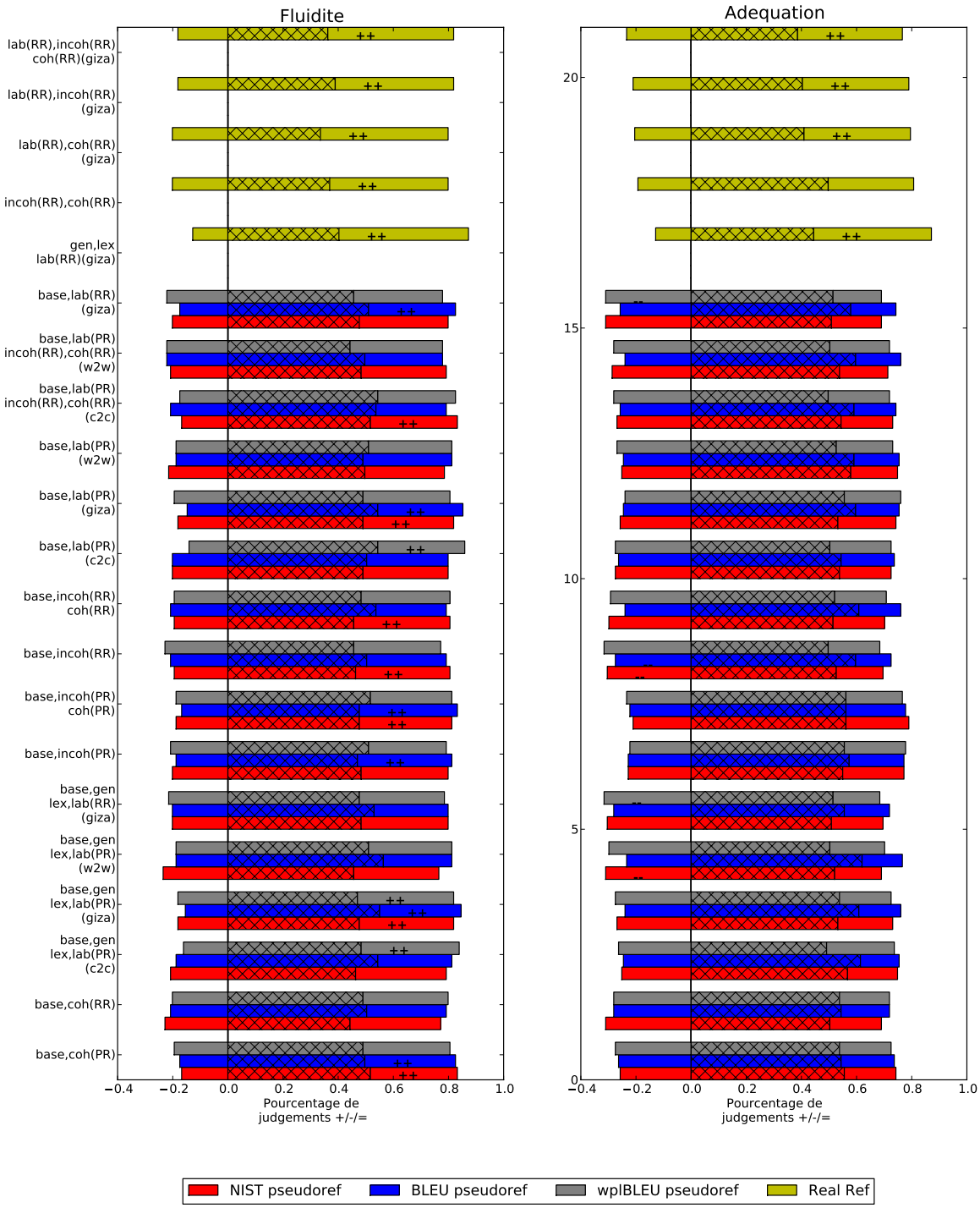


FIG. 5.5 – Résultats d'évaluation humaine subjective (anglais-français).

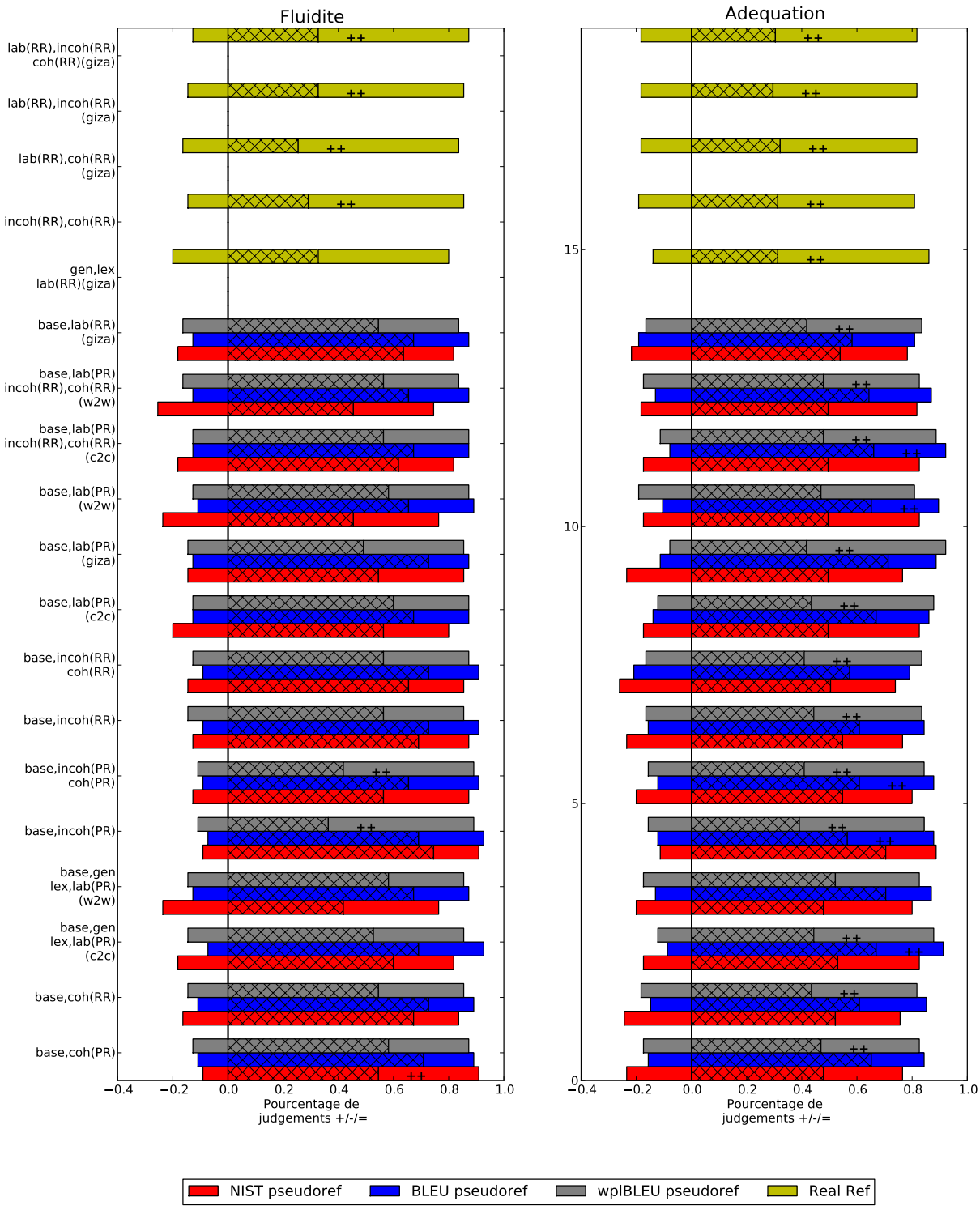


FIG. 5.6 – Résultats d'évaluation humaine subjective (français-anglais).

– soit il existe une meilleure traduction parmi les 1000 meilleures traductions, mais elle n'a été trouvée par aucun des modèles de reclassement choisis.

Les résultats de cette étude sont présentés dans le tableau 5.4. D'après ces résultats, seulement 16% des traductions de la liste ne peuvent pas être améliorées (aucune de traductions de la liste n'est peut être considérée meilleure que la traduction de base). Parmi les 84% des traductions qui peuvent être améliorées, le modèle de reclassement a amélioré (a choisi une traduction de la liste qui est meilleure que la traduction de base) 65% en terme de fluidité et 55 % en terme d'adéquation pour la traduction du français vers l'anglais. Ces résultats, ainsi que les résultats d'évaluations humaines (figures 5.6, 5.5) montrent qu'il est plus facile d'améliorer la fluidité de la traduction que son adéquation. De plus, la notion d'adéquation est souvent liée à la notion de fluidité, car une traduction peu fluide sera rarement jugée adéquate.

Les 6% des traductions pour lesquelles aucune amélioration possible n'existe représentent typiquement une limitation de l'approche par reclassement. De plus, beaucoup de traductions, bien qu'elles soient considérées comme améliorées, ne sont pas vraiment de bonnes traductions (quelques exemples de traductions sont donnés dans l'annexe C). Nous allons nous intéresser à ce problème dans la troisième partie.

Troisième partie

Traduction avec des contraintes syntaxiques source

Introduction

Dans cette partie, nous abordons une autre façon d'enrichir le modèle à fragments avec des connaissances syntaxiques. Les deux modèles proposés dans cette partie ont pour but de simplifier le processus de traduction (décodage), en utilisant des connaissances sur la structure syntaxique source. L'idée de ces deux modèles est de détecter les fragments d'une phrase source qui peuvent soit être traduits indépendamment du contexte de la phrase entière, soit permettre de limiter les permutations qui contredisent la structure syntaxique source.

La motivation derrière cette approche est aussi de donner des indices de la structure syntaxique au modèle à fragments, et ainsi de limiter l'espace de recherche de la solution optimale parcouru par le modèle de base. Nous espérons éviter l'exploration d'hypothèses qui violent la structure syntaxique, et probablement mènent à des mauvaises traductions. Ces hypothèses, donc, sont peu intéressantes. Cela permet également de traiter le problème de limitation de la liste des n-meilleures traductions produites par un système de TA à fragments, qui ne contient souvent pas d'hypothèse intéressante. Cela est dû au fait qu'une bonne hypothèse a été probablement perdue dans un grand nombre d'hypothèses non intéressantes considérées par le modèle de base.

Chapitre 6

Traduction avec des contraintes syntaxiques de distorsion

6.1 Modèle de contraintes syntaxiques de distorsion

6.1.1 Exemple pour présenter et motiver l'approche

Nous nous sommes inspirée du travail de [Cherry, 2008], qui a montré que la préservation de la cohésion[1] au niveau de la structure de dépendances entre la phrase source et la phrase cible permet d'améliorer la qualité des traductions en termes de scores automatiques et de jugements humains. En pratique, les travaux de Cherry [2008] et de Fox [2002] ont montré que la contrainte de cohésion n'est pas toujours respectée. Notamment, les erreurs d'analyse peuvent causer la violation de cette contrainte. Ainsi, il est nécessaire d'introduire des contraintes plus souples. Cherry [2008] propose d'introduire un trait supplémentaire dans le décodeur, indiquant le nombre des contraintes de cohésion violées, permettant ainsi au décodeur de considérer aussi des traductions ne préservant pas la structure de dépendances.

Nous proposons une autre solution pour intégrer des contraintes souples dans le processus de décodage. Notre solution consiste à entraîner le modèle en nous basant sur l'analyse syntaxique de la phrase source (l'analyse des constituants et l'analyse de dépendances) pour choisir les fragments de la phrase source tels qu'aucune distorsion n'est possible en dehors de ces segments.

Pour rendre notre approche plus claire et mieux la motiver, nous allons l'expliquer sur une phrase exemple.

Exemple :

Saddam 's power came from his control of the world 's second largest oil reserve

Nous proposons de définir les contraintes syntaxiques à partir d'une analyse syntaxique produite par XIP :

Arbre des *chunks* **:**

```
TOP{SC{NP{Saddam 's power} FV{came}} PP{from NP{his NP{control
PP{of NP{the world 's second AP{largest} oil reserve}}}}}}
```

Liste des dépendances :

```
MOD_POST(came,control)
MOD_POST(came,reserve)
MOD_PRE(reserve,largest)
MOD_PRE(power,Saddam)
MOD_PRE(reserve,world)
```

[1] Ici il s'agit de la cohésion entre les structures de dépendances source et cible, et non pas de la cohésion syntaxique au niveau de la phrase cible, comme c'était le cas dans précédemment.

```
MOD_PRE(reserve,oil)
MOD_POST(control,reserve)
DETD(control,his)
DETD(world,the)
QUANTD(reserve,second)
SUBJ_PRE(came,power)
PREPD(reserve,of)
PREPD(control,from)
PARTICLE(Saddam,'s)
PARTICLE(world,'s)
```

Une contrainte syntaxique est définie par des étiquettes XML de type <zone> ajoutées autour d'un fragment du texte source, indiquant que les mots cible correspondants doivent former un fragment connexe dans la traduction.
Prenons quelques exemples de contraintes de distorsion et regardons l'impact de ces contraintes sur la qualité de la traduction.

- Sans contraintes : Saddam's power came from his control of the world 's second largest oil reserve

- Contraintes C_1 : <zone> Saddam 's power </zone> came from <zone> his control of the world 's second largest oil reserve </zone>

- Contraintes C_2 : <zone> <zone> Saddam 's power </zone> came </zone> <zone>from <zone>his control <zone> of <zone> the world 's second largest oil reserve </zone> </zone> </zone> </zone>

Les traductions suivantes proposées par le modèle de référence pour chaque phrase sont :

- Sans contraintes : de son pouvoir de saddam hussein est le contrôle de la deuxième plus grande réserve de pétrole

- Contraintes C_1 : le pouvoir de saddam hussein est venue de son contrôle de la deuxième plus grande réserve de pétrole

- Contraintes C_2 : le pouvoir de saddam hussein est venue de son contrôle de la deuxième plus grande réserve de pétrole [*idéntique à la précédente*]

Nous voyons que l'introduction de contraintes syntaxiques de distorsion nous a permis d'obtenir une meilleure traduction finale par rapport à la traduction de base. De plus, si on regarde la taille du graphe des hypothèses pour chaque type de contraintes (tableau 6.1) nous voyons que nous avons exploré deux fois moins d'hypothèses, dont seulement 30% ont été abandonnées (par rapport à 60% pour le système de référence) en introduisant les contraintes sur la distorsion.

Nous comparons les scores internes donnés par le modèle de base. Le score interne attribué par le décodeur à la traduction du modèle de base sans contraintes est -7,84435, tandis que le score de la nouvelle traduction obtenue après l'introduction de contraintes est -7,92442. L'introduction de contraintes autour d'un segment "*his control of the world 's second largest oil reserve*" ne permet pas au mot *son* (traduction de *his*) d'être placé en dehors de ce segment dans la traduction. Ainsi, une nouvelle (meilleure) traduction est générée. Grâce aux contraintes de la traduction, la structure de cette traduction est améliorée, même si le score interne du modèle est plus bas.
Il s'agit ici d' une erreur du modèle : la solution optimale selon le modèle log-linéaire à fragments n'est pas en réalité une bonne traduction.

Le modèle de contraintes de distorsion peut également dans certains cas corriger les erreurs de recherche. Il s'agit de cas où la solution optimale selon le modèle de TA à fragments n'est pas trouvée à cause des heuristiques utilisées lors du parcours du graphe des hypothèses. Comme nous l'avons vu, l'introduction de contraintes réduit l'espace de recherche de la solution. Ces contraintes, donc, permettent

TAB. 6.1 – Nombre des hypothèses explorées et des hypothèses abandonnées en décodant avec et sans contraintes de distorsion.

	Sans contraintes	C_1	C_2
nb total hypothèses considérées	269651	127839	115923
nb hypothèses abandonnées	180939	46753	40050
nb hypothèses recombinées	76096	75669	71283
nb hypothèses élaguées	9898	3214	2459

dans certains cas de trouver une solution meilleure que celle trouvée par un modèle de base (nous allons en montrer quelques exemples dans la section 8.1).

Nous voyons ainsi sur cet exemple que l'introduction de contraintes syntaxiques sur la distorsion peut être bénéfique : non seulement une meilleure traduction a été générée, mais aussi la taille du graphe des hypothèses a été réduite. Une telle réduction du nombre des hypothèses explorées nous permet d'obtenir une liste des N meilleures traductions plus intéressante que celle du système de référence. Il est possible que la nouvelle liste contienne une bonne traduction, même si cette dernière n'était pas présente dans la liste générée par le modèle de référence. Ainsi, l'introduction de contraintes syntaxiques peut aussi avoir un impact sur le résultat final du reclassement avec les traits supplémentaires introduits dans la partie précédente (section 3.1).

6.1.2 Modélisation des contraintes syntaxiques de distorsion

Nous introduisons des contraintes de distorsion en ajoutant des balises XML qui indiquent la zone de distorsion possible. Ainsi, l'ensemble des fragments pour lesquels les contraintes de distorsion peuvent être appliquées forme une structure arborescente.
Nous allons représenter une phrase source S par un ensemble de fragments $C(S) = \{s_1, ..., s_k\}$. Pour simplifier la modélisation, nous allons restreindre les fragments de $C(S)$ uniquement à ceux respectant la structure arborescente :

$$\forall s_i, s_j \in C(S) : s_i \subset s_j \vee s_j \subset s_i \vee s_i \cap s_j = \varnothing \tag{6.1}$$

Le problème du modèle introduisant les contraintes est de choisir dans $C(s)$ de "*bons*" fragments qui doivent respecter les contraintes de distorsion. Cela peut être vu comme un problème de classification binaire qui a pour but de trouver une fonction de classification F qui mette en correspondance les entrées $s_i \in C(s)$ avec $y_i \in \{-1, +1\}$, où y_i indique si l'introduction d'une contrainte de distorsion autour de s_i est bénéfique ou non à la qualité de la traduction finale.

La décision d'introduire ou non des contraintes se base sur une fonction $\phi : C(S) \rightarrow \mathbb{R}^K$ calculant le vecteur de traits. La fonction de classification F est paramétrée par un vecteur de poids $w \in \mathbb{R}^K$ et par le coefficient $b \in \mathbb{R}$. Nous nous plaçons dans un cas de classificateur linéaire où la fonction de classification est définie comme suit :

$$F(s_i, w, b) = w^T \phi(s_i) + b = \sum_{k=1}^{K} w_i \phi_k(s_i) + b \tag{6.2}$$

La décision est prise en fonction du signe de la valeur de F. Ainsi, si $F(s_i) \geq 0$, la contrainte de distorsion autour de s_i est ajoutée, et sinon elle ne l'est pas.

6.1.2.1 Représentation d'une phrase source

Pour limiter l'ensemble des fragments à explorer, nous nous limitons dans un premier temps aux constituants syntaxiques proposés par XIP. L'ensemble des constituants syntaxiques forme une structure arborescente. Ainsi, des balises XML peuvent être introduites pour indiquer les limites de distorsion.

Arbre des *chunks* **:**

```
TOP{SC{NP{Saddam} FV{a dû}} NP{son pouvoir} PP{au NP{fait}}
SC{BG{qu'} NP{il} FV{contrôlait}} NP{la AP{deuxième} réserve}
AP{mondiale} PP{de NP{pétrole}} .}
```

Liste des dépendances :

```
SUBJ(dû,Saddam)
SUBJ(contrôlait,il)
OBJ(dû,pouvoir)
OBJ(contrôlait,réserve)
VMOD_POSIT2(dû,fait)
NMOD_POSIT1(réserve,mondiale)
NMOD_POSIT1(réserve,deuxième)
NMOD_POSIT1(réserve,pétrole)
PREPOBJ(pétrole,de)
PREPOBJ(fait,au)
DETERM_DEF(réserve,la)
DETERM_POSS(pouvoir,son)
AUXIL(dû,a)
CONNECT(contrôlait,qu')
```

FIG. 6.1 – Exemple d'analyse syntaxique fournie par XIP pour la phrase française : *"Saddam a dû son pouvoir au fait qu' il contrôlait la deuxième réserve mondiale de pétrole ."*

Rappelons que les constituants générés par XIP sont des structures minimales non récursives (*chunks*). Cela donne une structure relativement plate avec des *chunks* relativement courts. Si l'on introduisait des contraintes de distorsion uniquement autour des *chunks*, les distorsions à courte distance seraient privilégiées, mais nous n'aurions aucun moyen de tenir compte des relations à longue distance (qui sont pourtant fournies par l'analyse de dépendances) lors de l'introduction des contraintes de distorsion.

Nous avons ajouté une couche supplémentaire de règles aux grammaires de XIP (anglais, français, espagnol) qui regroupent certains *chunks* dans un nouveau constituant à condition qu'une relation de dépendance existe entre ces *chunks*. Un exemple d'une phrase avant et après l'application de cette nouvelle couche est donné dans la figure 6.2. Sur cette figure nous voyons qu'une structure arborescente plus profonde a été obtenue après l'introduction de cette couche supplémentaire.

6.1.2.2 Traits du modèle

Le modèle des contraintes de distorsion choisit parmi les fragments d'une phrase source $C(s)$ ceux qui doivent respecter les contraintes de distorsion. Nous proposons de le faire en affectant chaque fragment à une des deux classes : positif ou négatif. L'affectation d'un fragment à une classe se fait par un modèle linéaire combinant des traits de nature différente :

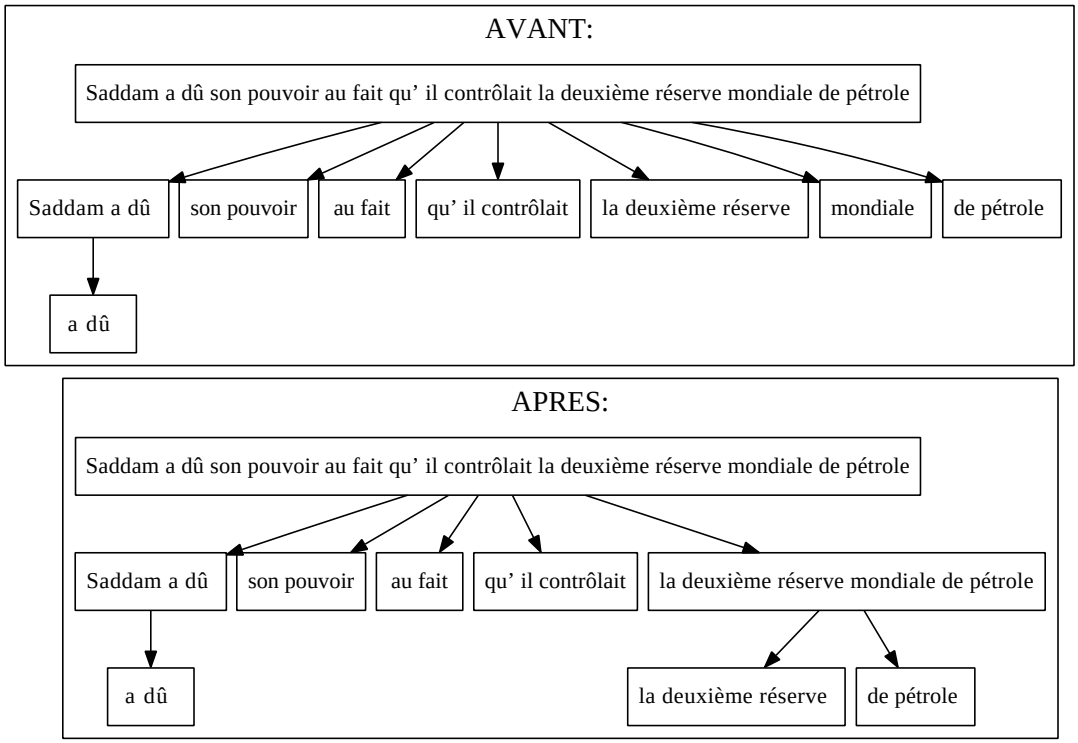


FIG. 6.2 – Analyses données par XIP avant et après avoir ajouté une couche de règles de regroupement des chunks. Une analyse de XIP est transformée en un arbre de *chunks*. Nous ne tenons pas compte des *chunks* ne contenant qu'un mot, car il ne sont pas intéressants pour le modèle de contraintes de distorsion.

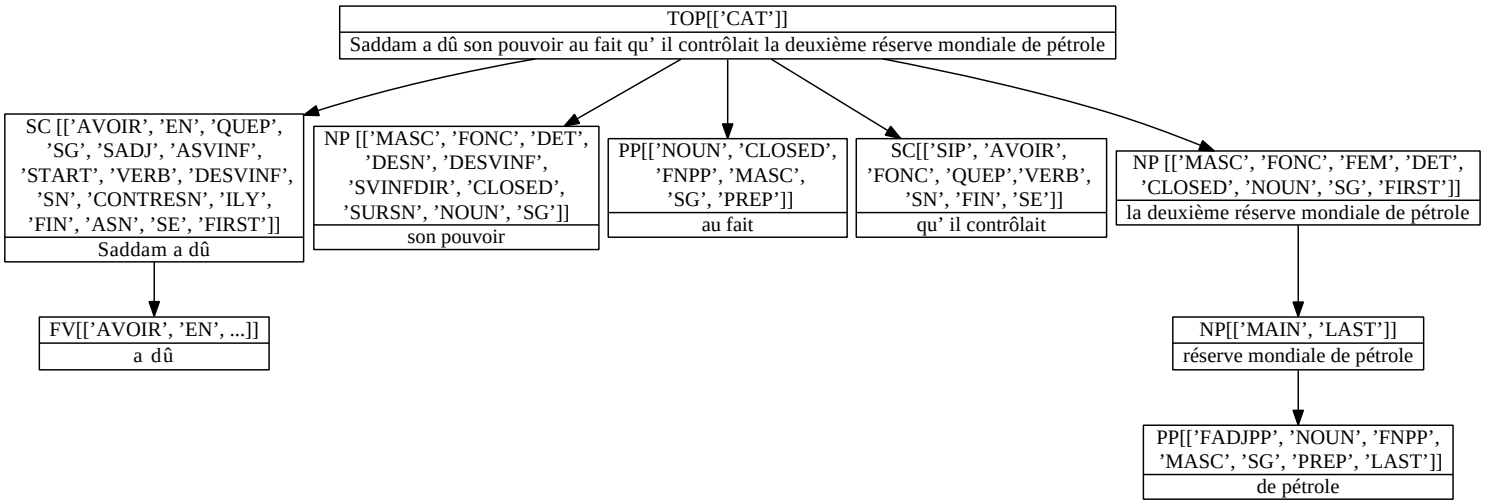


FIG. 6.3 – Les chunks produits par XIP avec des traits et catégories syntaxiques pour la phrase : “*Saddam a dû son pouvoir au fait qu’ il contrôlait la deuxième réserve mondiale de pétrole*”. Les traits attribués par XIP à chaque chunk contiennent des traits syntaxiques et morphologiques (MASC, SG, QUEP), ainsu que des traits relatifs à la position du chunk dans la phrase (START, END), et à la position du chunk fils dans le chunk parent (FIRST, LAST).

1. les traits définis par l'analyse des constituants ϕ_{ac} :
   - catégorie syntaxique du constituant,
   - attributs syntaxiques du constituant s_i : ce sont des traits morphologiques (genre, nombre), des traits spécifiques au rôle et à la position du constituant dans la phrase, etc.

   Un exemple d'analyse avec des traits pour chaque *chunk* est présenté sur la figure 6.3. Notre motivation est que certains traits de cette catégorie servent d'indicateurs pour introduire des contraintes de distorsion : ainsi, il sera naturel de dire que les mots source formant un syntagme nominal doivent rester ensemble dans la phrase cible.

2. les traits définis par l'analyse de dépendances ϕ_{ad} :
   - les étiquettes des arcs qui connectent le fragment avec son contexte ;
   - les étiquettes des arcs à l'intérieur du fragment ;

   Ces traits complètent les traits définis par les constituants. Ils doivent aussi avoir la capacité de gérer les cas où les règles de regroupement des *chunks* n'ont pas pu être appliquées. Si nous reprenons l'exemple de la figure 6.1, nous voudrions éviter d'introduire les contraintes de distorsion au niveau du *chunk* `la deuxième réserve`, mais plutôt d'introduire ces contraintes au niveau du fragment `la deuxième réserve mondiale de pétrole`. La relation de dépendance `NMOD(réserve, pétrole)` servira d'indicateur pour le modèle introduisant des contraintes.
3. des traits indépendants de l'analyse syntaxique ϕ_{plain} :
   - modèle de langage n-gramme pour un fragment ;
   - modèle de langage n-gramme pour un fragment avec son contexte ;
   - $\frac{NbMots(s_i)}{NbMots(S)}$, où $NbMots(s)$ est le nombre des mots contenus dans un segment ou fragment s.

   D'après les études faites par Fox [2002] la plupart des incohérences d'analyse entre l'anglais et le français viennent des erreurs d'analyse syntaxique. En reprenant l'exemple précédent, le fragment *réserve mondiale de pétrole* est probablement un fragment très fréquent et par conséquent les fragments *la deuxième réserve* , *mondiale* et *de pétrole* doivent être traités ensemble.

Ces réflexions motivant chaque trait que nous avons introduit ne sont données qu'à titre illustratif : en pratique cela peut être vrai dans certains cas et faux dans les autres. L'importance réelle de chaque trait sera apprise lors de l'apprentissage des paramètres.

6.2 Entraînement du modèle de traduction avec des contraintes de distorsion

En pratique, l'analyse S_1 d'une phrase source peut être cohérente (avec l'analyse cible) lors de la traduction vers une langue L_1, mais ne pas l'être pour une traduction vers une langue L_2. Ainsi, l'introduction des contraintes de distorsion dépend non seulement de la structure syntaxique source, mais aussi de la langue vers laquelle la traduction est faite.[2]

Nous souhaitons entraîner les paramètres de ce modèle sur un corpus parallèle en maximisant la *qualité*[3] de la traduction finale. Grâce à ce type d'entraînement, nous espérons apprendre des poids des traits du modèle qui équilibrent l'importance des traits syntaxiques et des autres traits, et rendent le modèle souple et robuste par rapport aux erreurs d'analyse syntaxique.

L'introduction de contraintes syntaxiques dans le décodeur peut améliorer la traduction, mais elle réduit surtout la taille de la liste des hypothèses, et même si la meilleure traduction proposée par le système avec des contraintes syntaxiques reste la même que celle de système de référence, il est possible que la liste des N meilleures hypothèses contiennent une meilleure traduction non trouvée dans la liste générée

[2] Cette idée n'est pas nouvelle en TA experte : dans les premiers systèmes de TA, l'analyse était faite en connaissant la langue cible. Cette architecture permet de simplifier les étapes d'analyse, de transfert et de génération, mais en même temps, rend les modules d'analyse et de génération peu ou pas réutilisables pour d'autres paires de langues. Dans la TA empirique, enrichie par la syntaxe, cette idée est souvent ignorée (Cherry [2008]; Mellebeek et al. [2006]). Nous souhaitons attirer l'attention sur ce type d'architecture (analyse spécifique à une paire de langues, et pas seulement à une langue source). Dans le contexte empirique, le réapprentissage du module d'analyse (et de génération si nécessaire) est automatique.

[3] La notion de qualité est à définir dans ce contexte.

par le système de référence. Nous allons étudier l'impact des contraintes de distorsion non seulement sur la meilleure traduction mais aussi sur la liste des N meilleures traductions.

6.2.1 Entraînement du modèle des contraintes

Le but de l'apprentissage est d'apprendre les valeurs de w et b (équation 6.2) qui décrivent le mieux les données d'apprentissage. Supposons qu'à cette étape nous ayons un ensemble d'exemples annotés $T(S_1, ..., S_N) = \{(s_i, y_i)\}_{s_i \in \{C(S_1),...,C(S_N)\}}$. Pour les phrases source $S_1, ..., S_N$, chaque fragment s_i de $C(S_n)$ (pour chaque phrase S_n) est annoté comme positif ou négatif. Nous souhaitons apprendre les paramètres w et b tels que les exemples de $T(S_1, ..., S_N)$ soient classifiés le plus correctement possible.
.

6.2.1.1 Génération de l'ensemble d'entraînement

L'apprentissage des paramètres nécessite un ensemble d'exemples annotés. Nous allons dire qu'une contrainte est négative si elle dégrade la qualité de la traduction, et positive si la traduction reste au moins aussi bonne après l'introduction de la contrainte. Lors de l'apprentissage des paramètres, nous devons définir la notion de qualité. Nous allons soit utiliser les mesures automatiques, soit des jugements humains pour comparer les deux traductions et annoter les fragments.

Nous proposons de générer l'ensemble d'entraînement annoté comme suit. Pour chaque fragment $s_i \in C(s)$, nous ajoutons des contraintes de distorsion autour de ce fragment, ce qui nous donne une phrase source avec une contrainte autour du fragment s_i, dénoté $s|_{s_i}$. Nous comparons ensuite la traduction de la phrase sans contraintes $t(s)$ avec la traduction de cette même phrase avec la contrainte s_i, $t(s|_{s_i})$: si la contrainte s_i a dégradé la qualité de la traduction, nous l'annotons comme négative, autrement, comme positive.

La procédure décrite génère un ensemble d'entraînement dénoté par $GEN_1(s)$. Ce type d'ensemble d'entraînement mène à un apprentissage de paramètres qui vise à améliorer (ou ne pas dégrader) *la première traduction* proposée par le système de base. En pratique, l'introduction de contraintes de distorsion peut ne pas changer la qualité de cette première traduction dans la plupart des cas. En revanche, les contraintes auront peut être plus d'impact sur la liste des N meilleures traductions, et, par conséquent, sur le reclassement des traductions. Nous proposons donc une seconde procédure, qui cherche à adapter l'ensemble d'entraînement visant à améliorer la traduction obtenue après le reclassement.

Nous procédons comme précédemment, en générant l'ensemble des phrases source avec des contraintes $s|_{s_i}$ pour chaque fragment $s_i \in C(s_i)$. Contrairement au scénario précédent, nous générons une liste des N meilleures traductions $t_N(s|_{s_i})$ pour chaque $s|_{s_i}$. Nous comparons les traductions choisies par le modèle de reclassement $t_r(s)$ et $t_r(s|_{s_i})$: si la traduction $t_r(s)$ est meilleure que la traduction $t_r(s|_{s_i})$, le fragment s_i est annoté comme négatif, et sinon il est annoté comme positif. Ce type d'ensemble sera dénoté par $GEN_r(s)$.

6.2.2 Entraînement d'une chaîne de traduction globale

6.2.2.1 Étapes principales de la traduction

Rappelons que, dans notre modèle actuel, la traduction se fait en plusieurs étapes listées ci-dessous :

1. Introduction des contraintes syntaxiques (à partir de l'analyse syntaxique de la phrase source) ;
2. Génération de la liste des N meilleures traductions ;
3. Choix d'une meilleure traduction avec le modèle de reclassement.

Précédemment, nous avons fixé le modèle de traduction qui intervenait dans l'entraînement du modèle des contraintes. Nous faisons l'hypothèse que le réapprentissage des paramètres du modèle de traduction en tenant compte des contraintes nous mènera à de nouveaux paramètres qui seront probablement plus adaptés à la traduction avec des contraintes de distorsion. De même manière, les paramètres du modèle

de contraintes pourront être mis à jour avec le nouveau modèle de traduction, etc.

Le même type de réflexion est valable dans le cas où le modèle de reclassement interviendrait dans l'apprentissage des paramètres du modèle des contraintes : le réentraînement du modèle de reclassement avec le modèle de contraintes et le nouveau modèle de traduction changerait le résultat final.

Nous proposons un modèle d'apprentissage global des paramètres d'une séquence de modèles. En général, en rajoutant un modèle après un autre, il existe un risque d'accumuler les erreurs d'une étape à l'autre. En apprenant une séquence de paramètres, nous espérons apprendre à éviter une telle accumulation des erreurs, et trouver une bonne traduction globale.

Détaillons maintenant la procédure d'apprentissage que nous proposons sur l'exemple d'une traduction en plusieurs étapes, où chaque étape est un modèle linéaire. La traduction (supposée) optimale est générée comme suit :

1. Introduction des contraintes : $\hat{t}_1(s)$ est une phrase source avec des contraintes choisies par le modèle de contraintes de distorsion (modèle de classification). Un fragments s_i est dans $\hat{t}_1(s)$, si $\Lambda_1\Phi_1(s_i) > 0$. Si aucun des fragments n'est choisi (classé comme négatif par le modèle) $\hat{t}_1(s) = s$;
2. Traduction : $\hat{t}_2(s, \hat{t}_1) = \arg\max_{t_2}(\Lambda_2\Phi_2(s, \hat{t}_1, t_2))$; $\hat{t}_2$ est une traduction intermédiaire (avec des contraintes de distorsion) optimale ;
3. Reclassement : à cette étape, nous obtenons la traduction finale $\hat{t}(s, \hat{t}_1, \hat{t}_2) = \arg\max_t(\Lambda_3\Phi_3(s, \hat{t}_1, \hat{t}_2, t))$.

6.2.2.2 Apprentissage d'une suite de paramètres

Le but de l'apprentissage est alors d'apprendre les paramètres $\Lambda_1, \Lambda_2, \Lambda_3$ afin de minimiser le taux d'erreur pour un ensemble d'entraînement formé de phrases parallèles $\{s_i, r_i\}$. Le taux d'erreur est calculé en comparant $\hat{t}(s_i)$ avec sa traduction de référence r_i.

Procédure d'apprentissage. Nous allons suivre le scénario proposé par Och [2003], en itérant le processus d'apprentissage avec le processus de retraduction.

L'ensemble d'entraînement de chaque étape GEN^e est toujours une liste de traductions finales : cela nous permet de dire que même les paramètres des étapes intermédiaires visent à maximiser la qualité de la traduction finale. La génération d'une traduction dépend des paramètres de chaque étape. GEN^e est utilisé pour l'optimisation des paramètres Λ_e. Il est généré en fixant tous les paramètres, sauf Λ_e.

La procédure générale que nous nous proposons de suivre pour l'apprentissage des paramètres est décrite par l'algorithme 4 (informel).

A chaque itération t, nous avons les paramètres $\Lambda_1^t, ..., \Lambda_E^t$ et pour chaque étape $e \in \{1..E\}$ nous exécutons la suite d'actions suivante :

1. Générer l'ensemble d'entraînement pour l'étape e : GEN_t^e ;
2. Mettre à jour les paramètres de l'étape e, Λ_e^t, en les apprenant à partir de l'ensemble d'entraînement GEN_t^e ;

Détails pour chaque étape. La nature des fonctions $GENERATE_e$ et $TRAIN_e$ est spécifique à l'étape concernée.

Pour une étape d'introduction de contraintes, $GENERATE_e$ correspond à la procédure de génération de l'ensemble dénoté GEN_r (si notre entraînement inclut l'étape de reclassement) ou GEN_1 (sinon). La fonction $TRAIN_e$ correspond à l'apprentissage d'un classificateur binaire entraîné sur GEN_r ou GEN_1.

Algorithme 4 *Entraînement d'une suite des paramètres*

ENTRÉES : Exemples d'entraînement $(X, Y) = \{(x_i, y_i)\}_{i \in [1..N]}$
SORTIES : Paramètres $\Lambda_1, ..., \Lambda_E$

```
Initialiser Λ_1, ..., Λ_E
pour t = 1..T faire
  pour e = 1..E faire
    GEN_t^e = GENERATE_e(X, Λ_1, .., Λ_{e-1}, Λ_{e+1}, .., Λ_E)
    Λ_e = TRAIN_e(Λ_e, GEN_t^e, Y)
  fin pour
fin pour
Retourner Λ_1, ..., Λ_E
```

L'entraînement du modèle de traduction demande la génération d'une liste exhaustive. Ainsi, la génération de l'ensemble d'entraînement est faite en plusieurs itérations (algorithme 1). Chaque itération consiste à générer une liste des n traductions choisies par le modèle de reclassement dans la liste des N ($N > n$) meilleures traductions selon le modèle de traduction courant. Cette liste de traductions est fusionnée ensuite avec les traductions générées aux étapes précédentes. L'optimisation des paramètres donne un nouveau modèle de traduction.

Le nombre d'itérations T de cette étape est défini soit par un critère d'arrêt (par exemple, non-changement des paramètres), ou prédéfini par un nombre fixe.

La mise à jour des paramètres du modèle de traduction peut se faire par MERT ou par Perceptron Structuré.

Le modèle de reclassement est entraîné comme précédemment avec un Perceptron Structuré, en utilisant la liste des n meilleures traductions comme ensemble d'entraînement.

Les expériences avec ce modèle (1 itération) sont présentées au chapitre 8.

Chapitre 7

Traduction par décomposition

Dans ce chapitre, nous décrivons une extension de la méthode de traduction avec des contraintes de distorsion. L'idée de cette extension est de simplifier une phrase source en remplaçant ces fragments par des fragments plus simples. Ainsi, la tâche de traduction est divisée en plusieurs sous-tâches plus simples.

7.1 Motivation

Au lieu d'introduire les contraintes de distorsion pour un fragment, il peut être plus efficace de remplacer ce fragment par un fragment plus simple. Par exemple, dans la phrase *"Il habite dans une petite maison blanche sur la colline"*, le fragment *"une petite maison blanche sur la colline"* peut être remplacé par le fragment *"une maison"*. Ce remplacement n'aura pas d'influence sur la traduction du contexte, mais simplifiera la traduction de la phrase initiale. La traduction du fragment remplacé sera faite de façon autonome. La traduction finale est obtenue par une recomposition des traductions des fragments et de la traduction de la phrase simplifiée.[1]

Cette approche est similaire à celle de Mellebeek et al. [2006] qui propose de simplifier la phrase source, en remplaçant certains fragments par des fragments plus simples statiques (prédéfinis) ou dynamiques (extraits à partir du fragment en question). Le processus de traduction est décomposé en :

(1) la traduction d'une phrase simplifiée (squelette de la phrase source),
(2) la traduction des fragments simplifiés,
(3) la recomposition de la phrase cible à partir de la traduction simplifiée et des sous-traductions.

La décomposition de la phrase source est faite à partir d'une analyse de dépendances : le fragment *pivot* est défini par la racine de l'arbre de dépendances (avec les nœuds voisins éventuellement), les fragments *satellites* sont les arguments à gauche et à droite du *pivot*.

La différence principale de notre approche avec celle de Mellebeek et al. [2006] est le modèle de décomposition que nous proposons. Ce modèle n'est pas fixe et dépend non seulement de la langue source (comme c'est le cas de Mellebeek et al. [2006]), mais aussi de la langue cible. Nous définissons notre modèle de décomposition en utilisant l'analyse syntaxique de la phrase source, et en l'apprenant sur la base d'un corpus parallèle afin de le rendre compatible avec la langue cible.

7.2 Description de notre approche

7.2.1 La procédure de décomposition

L'algorithme 5 de décomposition comporte deux étapes principales :

[1] Notons que dans certains cas, cette approche peut poser des problèmes : en fonction des sous-fragments, l'ordre des mots dans une phrase peut changer. Exemple : Dans la phrase simplifiée *J'ai vu Paul hier* et la phrase complète *J'ai vu hier le monsieur que tu m'a présenté, qui s'appelle Paul* le mot hier est placé en fonction de contexte. Ainsi, dans le cas général, la traduction d'une phrase simplifiée d'une façon complètement indépendante n'est pas toujours possible.

1. identifier des fragments qui peuvent être simplifiés ;
2. simplifier ces fragments en les remplaçant par des fragments plus simples.

Étape 1. Le modèle d'identification des fragments est similaire au modèle de contraintes. La structure arborescente des fragments est générée par XIP ; chaque fragment se voit attribuer une étiquette $+$ ou $-$, où $+$ indique que le fragment peut être simplifié. La structure obtenue après cette étape pour la phrase *Saddam's power came from his control of the world 's second largest oil reserve* est illustrée à la figure 7.1 .

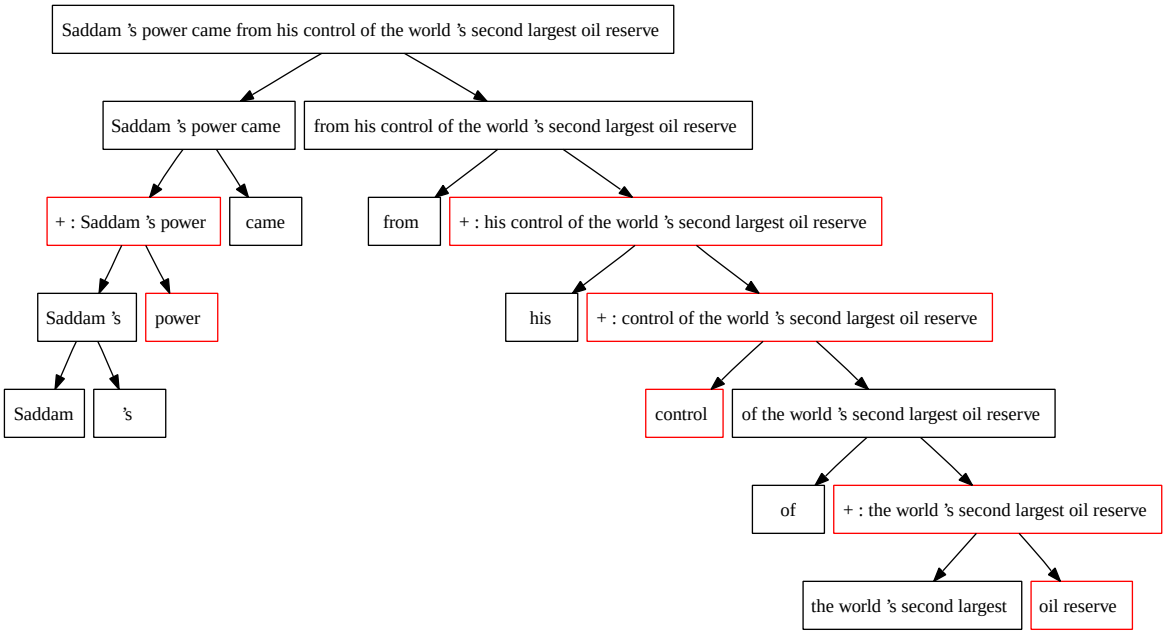


FIG. 7.1 – Choix des fragments à simplifier dans la structure arborescente de sous-phrases : $C_r(S)$. Le caractère binaire de cet arbre est accidentel.

Étape 2. La simplification du fragment s_i se fait par une substitution de ce fragment par un de ses fils. Le fragment de substitution est choisi parmi les fils directs du fragment s_i dans la structure arborescente générée par l'analyse syntaxique de XIP. Le fragment de s_i annoté comme un constituant principal par XIP est choisi pour remplacer s_i, et noté $sub(s_i)$.
Le processus de substitution est récursif : si le substitut de s_i ($s_j = sub(s_i)$) est annoté comme positif, il est remplacé à son tour par son substitut $sub(s_j)$, etc.
La structure de simplification obtenue par cette procédure est illustrée par la figure 7.2.

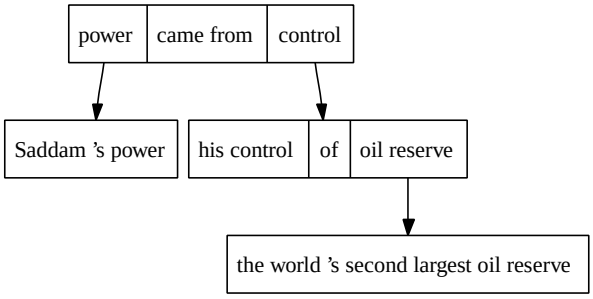


FIG. 7.2 – L'arbre de simplification S_r obtenu à partir de l'ensemble des fragments "positifs" $C_r(S)$.

Algorithme 5 *fonction* Décomposition(S)

ENTRÉES : phrase source S
SORTIES : Simplification de la phrase source S_r ;
Générer la structure de fragments-candidats par XIP : $C(S) = XIP(S)$
 # Choisir les fragments qui peuvent être remplacés
Initialiser $C_r(S) = \varnothing$
pour $s_i \in C(S)$ **faire**
si $\lambda_r \phi_r(s_i, sub(s_i)) \geq 0$ **alors**
ajouter s_i à $C_r(S)$
finsi
fin pour
Initialiser $S_r = S$
Trier les fragments de $C_r(S)$: commencer par le fragment le plus proche de la racine.
 # générer une phrase simplifiée en remplaçant les fragments choisis par leurs substituts
pour $s_i \in C_r(S)$ **faire**
Remplacer le fragment s_i par son substitut dans S_r : $S_r = S_r|_{s_i \leftarrow sub(s_i)}$
Chaque fragment S_r garde un lien avec le fragment qu'il a remplacé.
fin pour
Retourner S_r

Traits du modèle. Les traits du modèle de décomposition, contrairement au modèle introduisant les contraintes de distorsion, dépendent non seulement du fragment lui-même, mais aussi du fragment de substitution. Nous allons redéfinir les traits de la section 6.1.2.2, pour obtenir des traits plus adaptés au modèle de décomposition. Ce sont :

1. les traits définis par l'analyse des constituants $\phi_{ac}(s_i, sub(s_i))$; ce sont :
   – les traits définis par des attributs syntaxique des constituants s_i et $sub(s_i)$;
   – le trait défini par la paire des catégories syntaxiques de s_i et de $sub(s_i)$.
   Il est intuitivement clair, que certains type de constituants ne devraient pas être substitués (par exemple constituant de type PP). Ce trait doit distinguer également entre les cas où un constituant du type NP est substitué par un fragment du type NP ou du type ADJ (une situation possible suite aux erreurs de recombinaison de *chunks* par la couche supplémentaire de règles que nous avons introduite).
2. traits définis par l'analyse de dépendances $\phi_{ad}(s_i, sub(s_i))$; ce sont :
   – les étiquettes des arcs qui connectent $sub(s_i)$ avec son contexte ;
   – les étiquettes des arcs qui connectent le fragment s_i avec son contexte et sont à l'extérieur de $sub(s_i)$;
   – les étiquettes des arcs à l'intérieur de s_i, et connectant $sub(s_i)$ avec le reste du fragment s_i.
   Le but de ces traits et de comparer le contexte du fragment à substituer et de son substitut dans une phrase initiale. Par exemple, le fait qu'un substitut du fragment ne soit pas du tout relié au contexte (tandis que le fragment même l'était) peut servir d'indicateur qu'il n'est pas un bon substitut.
3. des traits indépendants de l'analyse syntaxique $\phi_{plain}(s_i, sub(s_i))$; ce sont :
   – le modèle de langage n-gramme pour s_i ;
   – le modèle de langage n-gramme pour s_i avec son contexte ;
   – le modèle de langage n-gramme pour un $sub(s_i)$ avec son contexte après substitution ;
   – $\frac{NbMots(s_i)}{NbMots(S)}$, où $NbMots(s)$ est le nombre des mots contenus dans le fragment s.

7.2.2 Le modèle de traduction

La traduction est générée à partir de la structure de simplification S_r. Pour chaque fragment $s_i \in S_r$, nous produisons une sous-traduction t_i. Les sous-traductions t_i forment un arbre de sous-traductions. L'information sur les alignements entre les mots de s_i et de t_i (fournie par le modèle de traduction ou trouvée autrement) permet de recomposer la traduction finale à partir de l'arbre des sous-traductions.

Algorithme 6 *fonction* TRANSLATE(S, TM)

ENTRÉES : Phrase source S, modèle de traduction de base TM
SORTIES : Traduction finale T
  S_r = Décomposition(S)
  //*Former l'arbre des traductions T_r à partir de l'arbre de simplification S_r*
  $T_r = \varnothing$
  pour $s_i \in S_r$ **faire**
    $t_i = TM(s_i)$ *# sous-traduction avec modèle de base*
    ajouter t_i à T_r
  fin pour
  Retourner T = GetFinalTranslation(T_r)

La procédure de traduction est décrite par l'algorithme 6. La fonction *GetFinalTranslation* (algorithme 7) définit la procédure de composition de la traduction finale à partir des sous-traductions.

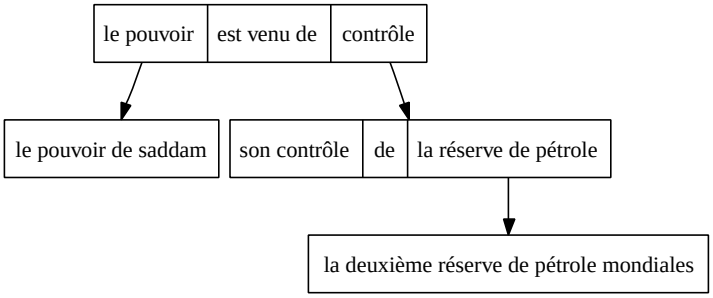


FIG. 7.3 – Arbre des sous-traductions T_r obtenu à partir de l'arbre de simplification S_r.

7.3 Entraînement

7.3.1 Modèle de décomposition

La procédure d'entraînement du modèle de décomposition est semblable à la procédure d'entraînement du modèle de contraintes. L'ensemble d'entraînement est généré en faisant une substitution pour chaque fragment. Ensuite, la traduction obtenue par recomposition est comparée avec la traduction de base : le fragment substitué est annoté comme positif si la traduction recomposée est de meilleure qualité que la traduction proposée par le modèle de base, négatif dans le cas contraire.

7.3.2 Adaptation du modèle de traduction

Notons que la structure des fragments que nous traduisons dans le cadre du modèle par décomposition est différente des phrases sur lesquelles le modèle de traduction a été entraîné. Par exemple, le modèle de langage aura tendance à pénaliser les traductions de fragments qui ne sont pas des phrases bien formées, car le modèle de langage a été entraîné sur un corpus de phrases complètes.

Deux types d'adaptation sont possibles : soit générer un corpus adapté à cette tâche, soit adapter les paramètres du modèle de traduction.

Génération du corpus adapté

Pour créer notre modèle de traduction adapté à la traduction réduite, nous utilisons l'analyse syntaxique du côté source et cible du corpus d'entraînement. Ainsi, chaque bi-phrase du corpus est transformée en une paire d'arbres de *chunks* (comme dans la figure 7.1). Nous établissons ensuite l'alignement

Algorithme 7 *fonction* GetFinalTranslation(S_r, T_r, $align$)

ENTRÉES : Une décomposition de la phrase source S_r, un arbre des sous-traductions T_r, une fonction $align(s, t, s^1)$ qui trouve un fragment $t^1 \subset t$ correspondant à $s^1 \subset s$

SORTIES : Traduction finale T

//Racine de T_r correspond à la traduction réduite d'une phrase initiale

t_0 =racine(T_r)

s_0 =racine(S_r)

$T = t_0$

pour $t_i \in$ChildNodes(t_0) **faire**

//A partir des alignements lexicaux entre s_0 et t_0, détecter le fragment cible correspondant à un substitut source de s_i dans s_0

Trouver la position de $sub(s_i)$ dans s_0

$sub(t_i) = align(s_0, t_0, sub(s_i))$

//Remplacer le fragment cible $sub(t_i)$ par une traduction finale de t_i

T_i = sous-arbre de T_r avec la racine dans t_i

$T = T|_{sub(t_i) \leftarrow \text{GetFinalTranslation}(T_i)}$

fin pour

Retourner T

entre ces deux arbres à partir des alignements lexicaux produits par GIZA++.

Définition. Nous allons dire qu'une paire de *chunks* source et cible forme un ***bi-chunk compatible avec l'alignement*** si chaque mot du *chunk* source est aligné avec un mot du *chunk* cible et inversement. De plus, aucun mot en dehors du *chunk* source ne peut être aligné à un mot à l'intérieur du *chunk* cible et inversement.

Nous créons l'ensemble B de tous les *bi-chunks* compatibles avec l'alignement. Dans cet ensemble, nous choisissons un sous-ensemble $\tilde{B}$ de *bi-chunks* formant une structure arborescente. Cela implique que, pour chaque paire de *bi-chunks* $b_k, b_l \in \tilde{B}$, les deux conditions suivantes doivent être respectées (où $src(b_i)$/$tgt(b_i)$ est la partie source/cible du *bi-chunk* b_i) :

- $src(b_k) \subset src(b_l) \Leftrightarrow tgt(b_k) \subset tgt(b_l)$;
- $src(b_k) = src(b_l) \Leftrightarrow tgt(b_k) = tgt(b_l)$.

Définition. Nous dirons que les *bi-chunks* b_k et b_l sont ***compatibles*** quand les deux conditions ci-dessus sont respectées.

Afin de choisir un sous-ensemble optimal $\tilde{B} \subset B$ de *bi-chunks* compatibles, nous procédons comme suit :

- une mesure de qualité d'alignement α_i est attribuée à chaque *bi-chunk* $b_i \in B$, $i \in \{1, .., K\}$;
- nous créons une matrice de compatibilité des *bi-chunks* $C = \{c_{ij}\}_{i,j \in \{1,..,K\}}$, où $c_{ij} = 1$ si les *bi-chunks* b_i et b_j sont compatibles, $c_{ij} = 0$ sinon ;
- un sous-ensemble $\tilde{B} \subset B$ est défini par un vecteur $x \in \mathbf{R}^K$, où chaque $x_k \in \{0, 1\}$ indique si le *bi-chunk* $b_i \in B$ appartient à $\tilde{B}$ ou non.
- x est solution de l'équation suivante :

$$\begin{aligned} \max L(x) \quad &= \quad \sum_{k=1}^{K} \alpha_i x_i \\ &s.c. \quad (1 - c_{kl})(x_k + x_l) \leq 1, k, l \in \{1, .., K\} \end{aligned}$$

La contrainte de cette équation exprime le fait que $\tilde{B}$ ne peut contenir que des bi-chunks compatibles. Si les bi-chunks b_k et b_l ne sont pas compatibles ($c_{kl} = 0$), au maximum un de ces deux

bi-chunks peut être choisi.

Notons que le problème exprimé par cette équation est un problème NP-complet[2]. Cependant, si la taille de B n'est pas très grande, il est possible de trouver une solution de ce problème en temps raisonnable. Le fait que nous n'explorons que des *bi-chunks* compatibles avec l'alignement nous permet de réduire le nombre des *bi-chunks* d'une manière importante[3].

Une fois que la structure arborescente de *bi-chunks* a été construite, nous créons un corpus de bi-phrases simplifiées (réduites). Nous parcourons l'arbre des bi-chunks en commençant par la racine. Pour chaque bi-chunk $b_i \in \tilde{B}$, nous remplaçons $src(s_i)$ par $sub(src(s_i))$ dans la phrase source, et de même pour la phrase cible. Pour chaque bi-chunk suivant, le remplacement est fait uniquement s'il n'est pas en conflit avec des remplacements déjà faits (c'est-à-dire s'il n'est pas inclus dans un bi-chunk qui a déjà été remplacé). Une fois que tous les remplacement possibles ont été faits, une paire de phrases réduites est obtenue.

Si un bi-chunk est en conflit avec un remplacement déjà fait, il est remplacé non pas dans la bi-phrase initiale, mais dans le bi-chunk (bi-fragment) qui est son parent dans l'arbre des bi-chunks.

Un tel corpus avec des bi-phrases réduites forme un corpus parallèle pour l'entraînement du modèle de la traduction réduite. La partie cible de ce corpus sert à entraîner le modèle de langage n-gramme réduit.

Adaptation des paramètres du modèle de traduction

L'entraînement d'une suite de modèles, comme décrit par l'algorithme 4, permet d'adapter les paramètres du modèle de la traduction. Ce type d'adaptation peut être fait soit après avoir entraîné un modèle de traduction sur un corpus adapté, soit en adaptant les paramètres du modèle de traduction de base.

7.3.3 Intégration du modèle de reclassement dans le modèle de traduction par décomposition

Même après l'adaptation des paramètres du modèle de traduction, il est possible qu'une sous-traduction soit erronée à cause du manque de contexte. De plus, les erreurs peuvent être cumulées d'une étape à l'autre, dégradant la qualité de la traduction finale.

L'application du modèle de reclassement au niveau des sous-traductions permet d'avoir un contrôle supplémentaire sur la qualité de la sous-traduction (fluidité, adéquation).

De plus les listes de sous-traductions sont plus variées que les listes des traductions complètes. Nous avons donc plus d'espoir de produire une bonne traduction finale.

[2]Si chaque bi-chunk x_i correspond à un nœud du graphe, si α_i est le poids du nœud x_i, et si les nœuds correspondant aux bi-chunks non compatibles sont connectés par un arc, ce problème permet de coder le problème du stable (ensemble indépendant) maximum (pondéré).

[3]En effet, après avoir éliminé des *bi-chunks* non compatibles avec l'alignement, les seuls *bi-chunks* non compatibles peuvent être de type suivant : $src(b_1) > src(b_2), tgt(b_1) = tgt(b_2)$. Par exemple, on peut avoir la paire de bi-chunks suivante : b_1 =(*la France, France*) et b_2 =(*France, France*).

Chapitre 8

Expériences avec ces nouveaux modèles

8.1 Expériences avec le modèle de contraintes de distorsion

8.1.1 Description des expériences

Dans ces expériences, qui restent préliminaires, nous avons entraîné seulement le modèle de distorsion, sans avoir adapté les autres modèles à la traduction pour tenir compte des contraintes de distorsion. Nous avons ensuite appliqué ce modèle de distorsion à la traduction avec le modèle de traduction de base. Cela nous a permis d'avoir une idée des améliorations et des dégradations possibles après l'introduction des contraintes de distorsion dans le modèle de TA à fragments.

8.1.1.1 Modèle de distorsion

La création de l'ensemble annoté nécessite un critère indiquant si la traduction a été améliorée ou non après l'introduction des contraintes. Nos expériences de reclassement ont montré que les mesures automatiques ne sont pas de bons indicateurs de la qualité (compréhension pour le juge) de la traduction individuelle. Nous avons vu, pourtant, que la mesure wlpBLEU est un peu meilleure que les mesures classiques (BLEU et NIST) dans le cas de la traduction vers l'anglais.

Suite à ces observations, nous nous sommes placée dans le cadre de la traduction vers l'anglais, en utilisant les critères suivants pour comparer les traductions avant et après l'introduction de contraintes :
- mesure wlpBLEU,
- jugements humains.

L'entraînement du modèle des contraintes de distorsion est fait soit par la version moyenne du Perceptron (Rosenblatt [1958]; Freund and Schapire [1999]), soit par SVM (Vapnik [1998]).

8.1.1.2 Autres modèles

Les autres modèles ne sont pas adaptés au modèle de distorsion dans cette série d'expériences. Nous prenons le modèle de traduction de base (section 2.2.3.1) pour effectuer la traduction d'une phrase avec les contraintes de distorsion.

8.1.2 Analyse des résultats

8.1.2.1 Introduction des contraintes de distorsion

Pour évaluer l'impact des contraintes de distorsion que nous avons introduites, nous effectuons les analyses suivantes de résultats :
- comparaisons des scores automatiques d'évaluation ;

– comparaison manuelle des traductions de chacun des modèles avec les traductions de référence pour estimer et qualifier les améliorations et les dégradations réelles apportées par l'introduction des contraintes ;[1]
– comparaison du score donné par le décodeur avant et après l'introduction de contraintes afin de voir si l'introduction des contraintes a permis de trouver une meilleure solution que celle proposée par le modèle de base.

TAB. 8.1 – Résultats d'évaluation automatique des modèles de traduction avec les contraintes de distorsion. Un modèle de contraintes de distorsion est défini par le critère utilisé lors de la création d'un ensemble d'exemples annotés (wlpBLEU, évaluation manuelle), et par l'algorithme de classification utilisé pour apprendre le modèle des contraintes de distorsion (SVM, Perceptron).

	NIST	BLEU	wlpBLEU
système de base	6,9933	0,2674	0,3712
Perceptron, wlpBLEU	6,9543	0,2646	0,3685
SVM, wlpBLEU	6,9591	0,2649	0,3687
SVM, man éval	6,9775	0,2665	0,3705

TAB. 8.2 – Évaluation humaine des traductions générées avec des contraintes de distorsion. + : nombre de traductions jugées meilleures que la traduction de base, - : nombre de traductions jugées moins bonnes que la traduction de base, total diff : nombre de traductions différentes de la traduction de base (sur 1063 phrases du corpus de test).

	+	-	total diff
Perceptron wlpBLEU	89	79	231
SVM wlpBLEU	86	75	229
SVM man éval	92	76	203

Les résultats de l'évaluation manuelle (tableau 8.2) ont montré que l'introduction de contraintes de distorsion a un impact seulement sur environ 15% des traductions. Cela explique les faibles variations des scores automatiques (tableau 8.1).

L'évaluation manuelle montre que l'introduction des contraintes de distorsion a généralement un impact positif sur la compréhension (critère de qualité utilisé lors de l'évaluation manuelle) de la traduction. Nous pouvons voir également que l'apprentissage du modèle de distorsion avec SVM donne des résultats un peu meilleurs que l'apprentissage avec Perceptron. L'utilisation des jugements humains lors de la création de l'ensemble de l'entraînement a un meilleur impact sur la qualité de la traduction que l'utilisation de la mesure wlpBLEU.

Nous avons comparé d'abord les scores attribués par le modèle log-linéaire (décodeur) avant et après l'introduction des contraintes de distorsion. Nous avons observé que, dans certains cas, le score attribué par le décodeur a augmenté après l'introduction des contraintes. Cela signifie qu'une solution de meilleur score de l'équation 1.6 a été trouvée grâce aux contraintes de distorsion. En effet, il est probable que le décodeur ne trouve pas de solution exacte de l'équation 1.6, car de nombreuses heuristiques sont appliquées lors du décodage, afin d'obtenir une solution en temps raisonnable. Dans ce cas, nous parlons d'*erreur de recherche*. Après l'introduction des contraintes de distorsion, la taille du graphe des hypothèses est réduite, et une solution de meilleur score peut donc être trouvée.

Des exemples de phrases pour lesquelles une solution de meilleur score a été proposée par le décodeur après l'introduction des contraintes de distorsion sont donnés dans le tableau 8.3. Notons que

[1]Cette série d'évaluations a été faite en comparant chaque traduction avec une traduction de base (à l'aveugle : sans savoir quelle traduction est générée par quel modèle). Faute de temps et de moyens, ces évaluations ont été faites sans intervention de juges exterieurs (indépendants). Elles peuvent cependant servir d'indicateurs supplémentaires (en plus des scores automatiques) sur les changements apportés par le modèle.

la solution jugée meilleure par le décodeur ne correspond pas nécessairement à un meilleur jugement humain. De plus, dans certains cas, la solution plus optimale peut correspondre à exactement la même solution qu'avant (exemples 2, 3 du tableau 8.3). La différence des scores attribués par le décodeur s'explique par un processus différent de génération de cette traduction (choix des bi-fragments différents et l'ordre dans lequel ces bi-fragments ont été choisis).

Quand la traduction jugée meilleure par le décodeur ne correspond pas à une bonne traduction en termes de jugements humains, nous parlons d'*erreur du modèle*. Nous observons que, dans la plupart des cas, les scores attribués par le décodeur ont diminué après l'introduction des contraintes de distorsion. Pour un certain nombre de traductions, cette diminution de score du décodeur a mené à de meilleures traductions en termes de jugements humains. Il s'agit donc de cas où les erreurs du modèle ont été corrigées (au moins en partie). Quelques exemples sont données dans le tableau 8.4.

Quelques exemples. Analysons maintenant des types de dégradation et d'amélioration apportées par l'introduction de contraintes de distorsion[2]. Quelques exemples de traductions sont présentés dans les tableaux 8.4, 8.5. Nous avons mis en gras la partie de la traduction qui a changé après l'introduction de contraintes, ainsi que les contraintes qui ont causé ce type de changement.

Le tableau 8.4 montre des exemples de traductions où l'introduction de contraintes a amélioré la traduction. La traduction a été jugée meilleure si elle est devenue plus adéquate (et souvent, par conséquent, plus fluide) que la traduction de base.

Nous avons observé 2 sources principales de dégradation de la traduction après l'introduction des contraintes (tableau 8.5) :

– les distorsions locales ne sont plus faites après l'introduction de contraintes, alors qu'elles étaient faites avant (deux premiers exemples du tableau 8.5) ;
– un choix lexical moins adapté est fait après l'introduction des contraintes (dernier exemple du tableau 8.5).

Un point positif est que la raison principale de la dégradation n'est pas une mauvaise contrainte imposée suite à une analyse erronée. Les deux types cités de dégradation peuvent probablement être corrigés en adaptant le modèle de traduction. En fait, les paramètres du modèle de traduction sont optimisés pour une traduction standard, sans contraintes de distorsion syntaxiquement motivées. Il est possible que le réapprentissage de ces paramètres dans le cadre de l'introduction de contraintes permette de corriger certaines erreurs que nous avons rencontrées.

Le tableau 8.6 donne quelques exemples de traductions jugées équivalentes avant et après l'introduction de contraintes.

Nous voyons que c'est typiquement l'ordre des mots qui pose des problèmes dans ces exemples. Ainsi, dans le premier exemple, le modèle de contraintes de distorsion a regroupé les mots du fragment *peu importante, stables et rurales*, ce qui interdit la traduction proposée par le modèle de base : *low populations , stable and rural*. La traduction proposée après l'introduction de contraintes (*populations unimportant , stable and rural*) n'a pas amélioré l'ordre des mots dans la traduction. La bonne traduction nécessite une permutation locale de *populations* et du fragment *unimportant , stable and rural*. Ce type d'erreur a déjà été cité précédemment, comme une source de dégradation des traductions. Ces erreurs peuvent être corrigées en adaptant le modèle de traduction.

Notons que la dégradation de la traduction dans certains cas (en minorité) est liée à l'introduction de contraintes de distorsion indésirables (derniers exemples des tableaux 8.5, 8.6). Ce problème vient principalement de l'analyse proposée par XIP : les contraintes sont introduites autour de *chunks* minimaux, non complets. L'extension de règles de regroupement dans une couche supplémentaire, ou l'utilisation d'un autre analyseur syntaxique améliorerait probablement ces résultats, en évitant ce type de problèmes.

[2] Quand nous parlons des problèmes typiques du modèle, il s'agit des problèmes liés à l'introduction des contraintes de distorsion. Nous évaluons donc, non pas les traductions elles-mêmes, mais la différence entre les deux traductions avant et après l'introduction des contraintes de distorsion. Nous ne parlons pas des erreurs qui sont communes aux deux traductions, et qui sont des erreurs du modèle de base, dont certaines ont été citées dans 2.2.4.

TAB. 8.3: Exemples de phrases pour lesquelles le nombre d'erreurs de recherche a été réduit après l'introduction des contraintes de distorsion.

source	système de base	score de base	traduction avec contraintes	score avec contraintes
Ils ont réaffirmé leur engagement envers le Pacte de stabilité et de croissance , **qui exige des pays de la zone euro des augmentations d' impôts et des baisses de dépenses publiques** , accentuant la pression sur leur économie .	they have reaffirmed their commitment to the stability and growth pact , **which requires the euro area countries increases taxes and cuts in public spending** , widening the pressure on their economy .	-28.2023	they have reaffirmed their commitment to the stability and growth pact , **which calls for the entire eurozone tax increases and public spending cuts** , widening the pressure on their economy .	-28.1838
Mais ce qu' ils font , en fait , revient à gérer les problèmes les plus importants du moment au détriment de la disparition d' institutions et de politiques mises en place par leurs prédécesseurs pour contrôler des problèmes qu' ils percevaient comme les plus urgents .	but what they are , in fact , is the most important problems of time at the expense of the disappearance of institutions and policies implemented by their predecessors to control problems that they received as the more urgent .	-48.6638	but what they are , in fact , is the most important problems of time at the expense of the disappearance of institutions and policies implemented by their predecessors to control problems that they received as the more urgent .	-48.4977
Les analyses élaborées des coûts et des bénéfices de grands projets ont constitué la routine du département de la Défense et d' autres départements pendant presque un demi-siècle .	the analysis process of the costs and benefits of major projects were routine department of defence and other départements for almost half-a-century .	-42.0359	the analysis process of the costs and benefits of major projects were routine department of defence and other départements for almost half-a-century .	-41.9816

TAB. 8.4: Exemples de phrases pour lesquelles l'introduction des contraintes de distorsion permet d'améliorer la traduction.

source	contraintes	système de base	traduction avec contraintes
Mes hôtes étaient d' anciens dirigeants communistes qui assumaient désormais le rôle de présidents **élus plus ou moins démocratiquement** .	<zone> mes hôtes étaient </zone> d' anciens dirigeants communistes qui assumaient désormais <zone> le <zone> rôle <zone> de présidents </zone> </zone> </zone> élus plus ou moins démocratiquement .	my hosts were former communist leaders , who assumaient now the role of **democratically elected** presidents **more or less** .	my hosts were former communist leaders , who assumaient now the role of presidents **more or less democratically elected** .

Ainsi , ne reste que la solution keynésienne : utiliser **les politiques monétaires (baisse des taux d' intérêt) et fiscales (inflation des dépenses publiques et réduction des impôts)** pour maintenir l' économie loin de ce gouffre qui rend la déflation possible .	<zone> ainsi , <zone> ne reste </zone> </zone> que <zone> la <zone> solution keynésienne </zone> </zone> : utiliser <zone> **les** <zone> **politiques monétaires** </zone> </zone> **(** <zone> **baisse** <zone> **des taux** </zone> </zone> <zone> **d' intérêt** </zone> **) et fiscales** <zone> **(** <zone> **inflation** <zone> **des** <zone> **dépenses publiques** </zone> </zone> </zone> **et** <zone> **réduction** <zone> **des impôts** </zone> </zone> **)** </zone> <zone> pour <zone> maintenir <zone> l' économie </zone> </zone> </zone> <zone> <zone> loin de </zone> <zone> ce gouffre </zone> </zone> qui rend <zone> la <zone> déflation possible </zone> </zone> .	so , is that the solution keynésienne : use **monetary policy (lower interest rates and inflation tax) (public spending and tax cut)** to maintain the economy far from this chasm that makes deflation possible .	so , is that the solution keynésienne : use **monetary policy (lower interest rates) and fiscal (inflation public spending and tax cut)** to maintain the economy far from this chasm that makes deflation possible .
La mondialisation des marchés , **le commerce sans entrave , l' islam militant** , le réveil de la Chine : ce sont là les choses que les historiens et les stratèges décrivent habituellement comme les forces essentielles qui façonnent notre destinée .	<zone> la <zone> mondialisation <zone> des marchés </zone> </zone> </zone> , <zone> **le** <zone> **commerce** <zone> **sans entrave** </zone> </zone> </zone> **,** <zone> **l'** <zone> **islam militant** </zone> </zone> , <zone> le <zone> réveil <zone> de <zone> la chine </zone> </zone> </zone> </zone> <zone> : ce <zone> sont là </zone> </zone> <zone> les choses </zone> <zone> que <zone> <zone> les historiens </zone> et <zone> les stratèges </zone> </zone> <zone> décrivent habituellement </zone> </zone> <zone> comme <zone> les <zone> forces essentielles </zone> </zone> </zone> <zone> qui <zone> façonnent <zone> notre destinée </zone> </zone> </zone> .	globalization of markets , **trade , without impediment militant islam** , the awakening china : these are things that historians and strategists usually described as the essential that shape our destiny .	globalization of markets , **trade without impediment , militant islam** , the awakening china : these are things that historians and strategists usually described as the essential that shape our destiny .
des institutions de stabilisation du marché **(gestion monétaire et fiscale)**	<zone> des <zone> institutions <zone> de stabilisation </zone> </zone> </zone> <zone> du marché </zone> <zone> **(** <zone> **gestion** <zone> **monétaire et fiscale** </zone> </zone> **)** </zone>	stabilization institutions market **management (monetary and fiscal)**	stabilization institutions market **(monetary and fiscal management)**

TAB. 8.5: Exemples des phrases pour lesquelles l'introduction des contraintes de distorsion degrade la traduction.

source	contraintes	système de base	traduction avec contraintes
En Europe , la Banque centrale européenne pense que le danger d' une **inflation incontrôlée** à la suite d' une perte de la confiance du public dans son engagement envers une faible inflation l' emporte sur les coûts du chômage qui reste trop élevé .	<zone> <zone> en europe </zone> , <zone> la <zone> banque <zone> centrale européenne </zone> </zone> </zone> pense </zone> <zone> que <zone> le <zone> danger <zone> d' <zone> une **<zone> inflation incontrôlée </zone>** </zone> </zone> </zone> </zone> <zone> <zone> à la suite d' </zone> <zone> une <zone> perte <zone> de <zone> la confiance </zone> </zone> </zone> </zone> </zone> <zone> du <zone> public <zone> dans <zone> son engagement </zone> </zone> </zone> </zone> <zone> envers <zone> une faible inflation </zone> </zone> <zone> l' emporte </zone> </zone> <zone> sur <zone> les <zone> coûts <zone> du chômage </zone> </zone> </zone> </zone> qui reste trop élevé .	in europe , the european central bank believe that the danger of **uncontrolled inflation** following a loss of public confidence in its commitment to low inflation exceeds the cost of unemployment which remains too high .	in europe , the european central bank believe that the danger of **inflation uncontrolled** following a loss of public confidence in its commitment to low inflation exceeds the cost of unemployment which remains too high .
L' idéalisme messianique qui a délivré l' Europe du nazisme et a protégé l' Europe occidentale du communisme est désormais dirigé vers d' autres ennemis .	<zone> <zone> l' **<zone> idéalisme messianique </zone>** </zone> <zone> <zone> qui <zone> a délivré </zone> </zone> <zone> l' <zone> europe <zone> du nazisme </zone> </zone> </zone> <zone> et <zone> a protégé </zone> </zone> </zone> <zone> l' <zone> europe occidentale <zone> du communisme </zone> </zone> </zone> <zone> est désormais dirigé </zone> </zone> <zone> vers <zone> d' <zone> autres ennemis </zone> </zone> </zone> .	**the messianic idealism** which has issued europe of nazism and protected western europe from communism is now led to other enemies .	**idealism messianic** who has issued europe of nazism and protected western europe from communism is now led to other enemies .
La croissance de la masse monétaire est **bien supérieure aux niveaux cible** depuis un certain temps , ce qui indique une liquidité excessive .	**<zone>** <zone> la <zone> croissance <zone> de <zone> la <zone> masse monétaire </zone> </zone> </zone> </zone> </zone> <zone> est bien </zone> **</zone>** supérieure <zone> aux niveaux </zone> cible <zone> depuis <zone> un certain temps </zone> </zone> , ce <zone> qui indique </zone> <zone> une <zone> liquidité excessive </zone> </zone> .	the growth of the money supply is **far superior to target levels** for some time , which indicates a liquidity excessive .	the growth of the money supply is **far lower levels target** for some time , which indicates a liquidity excessive .

Le projet européen représente **la réponse d' un réaliste** à la mondialisation et à ses défis .	<zone><zone>le <**zone**> projet européen</zone> </zone> <zone> représente <zone> la réponse</zone> </zone> **</zone> <zone>** d' <zone> un <zone> réaliste <zone> <zone>à <zone> la mondialisation </zone></zone> et <zone> à <zone> ses défis </zone> </zone></zone> **</zone>**</zone> </zone> .	the european project is **a realistic response** to globalization and its challenges .	the european project is **the answer to a realistic** globalization and its challenges .

TAB. 8.6: Exemples de phrases pour lesquelles l'introduction de contraintes n'a ni amélioré, ni dégradé l'adéquation de la traduction.

source	contraintes	système de base	traduction avec contraintes
Le système politique de ces pays était adapté pour des **populations peu importante , stables et rurales** .	<zone> <zone> le système </zone> politique </zone> <zone> de <zone> ces pays </zone> </zone> <zone> était adapté </zone> <zone> pour <zone> des **<zone>** populations **<zone>** peu importante , <zone> stables et rurales </zone> **</zone> </zone>** </zone> </zone> .	the political system of these countries was suitable for **low populations , stable and rural** .	the political system of these countries was suitable for **populations unimportant , stable and rural** .
Toutes les menaces contre **les intérêts nationaux de l' Amérique** ne peuvent pas se résoudre par la force militaire .	<zone> <zone> <zone>t outes les </zone> <zone> menaces <zone> contre **<zone>** les <zone> intérêts nationaux </zone> **</zone>** </zone> </zone> </zone> **<zone>** de <zone> l' amérique </zone> **</zone>** <zone> ne <zone> peuvent pas </zone> </zone> </zone> <zone> se résoudre </zone> <zone> <zone> par la force </zone> militaire </zone> .	all the threats against **america 's national interests** cannot be solved by military force .	all the threats against **the national interests of america** cannot be solved by military force .
Un degré sans précédent de solidarité traverse désormais l' Europe , se manifestant par exemple dans **le deuil collectif et les effusions de compassion** envers l' Espagne ; nous devons nous appuyer sur ce vaste potentiel pour créer une logique de solidarité dans le monde .	<zone> **<zone>** un degré </zone> **<zone>** sans précédent **</zone> <zone>** de solidarité </zone> <zone> traverse désormais </zone> </zone> <zone>l' europe </zone> , <zone> se <zone> manifestant <zone> par exemple </zone> </zone> </zone> <zone> dans <zone> le **<zone>** deuil collectif **</zone>** </zone> </zone> <zone> et <zone> les <zone> effusions <zone> de compassion </zone> </zone> </zone> <zone> envers <zone> l' espagne </zone> </zone> ; nous <zone> devons <zone> nous appuyer </zone> </zone> </zone> <zone> sur <zone> ce vaste potentiel </zone> </zone> <zone> pour <zone> créer <zone> une <zone> logique <zone> de solidarité </zone> </zone> </zone> </zone> </zone> <zone> dans <zone> le monde </zone> </zone> .	**an unprecedented degree of solidarity** across europe now , signaling for example in **mourning and collective effusions of sympathy** for spain ; we must rely on this vast potential for creating a logic of solidarity in the world .	**a degree unprecedented solidarity** now in europe , signaling for example in the **collective mourning and effusions of sympathy** for spain ; we must rely on this vast potential for creating a logic of solidarity in the world .[3]

[3]Nous avons jugé cette traduction équivalente à la traduction de base car, malgré les dégradations dans la première partie de la traduction, la deuxième partie a été améliorée.

8.1.2.2 Reclassement d'une liste de traductions obtenue après l'introduction des contraintes

Le tableau 8.7 donne les résultats d'évaluation automatique appliquée à une liste reclassée, par utilisation de traits de couplage et de traits génériques de cohésion, de traductions produites en tenant compte des contraintes de distorsion. Les différences des scores automatiques entre les modèles différents de contraintes après reclassement sont plus ou moins du même ordre qu'avant le reclassement (faible).

Nous avons comparé manuellement le modèle de base et le modèle de contraintes de distorsion (entraîné avec SVM, en utilisant les jugements humains lors de la création de l'ensemble d'entraînement) dans le cadre du reclassement. Pour le faire, nous avons extrait un échantillons de 300 phrases du corpus de test, et nous avons comparé les traductions proposées pour ces phrases par chacun des modèles. Les résultats de ces comparaisons sont présentés dans le tableau 8.8. Notons tout d'abord qu'il semble y avoir moins de différences entre les deux modèles qu'avant le reclassement (40% après le reclassement, 80% avant le reclassement).

En regardant les traductions, nous avons remarqué que, après le reclassement, la plupart des traductions jugées comme améliorées par rapport au modèle de base étaient déjà améliorées avant le reclassement.

Toutes les dégradations trouvées dans l'échantillon que nous avons extrait sont liées aux erreurs du reclassement. En fait, en regardant la liste des traductions, nous voyons qu'une meilleure traduction pourrait être trouvée, mais qu'elle n'a pas été choisie par le modèle de reclassement. Ces erreurs pourraient être évitées en adaptant le modèle de reclassement.

Notons également que les dégradations observées avant le reclassement, liées aux distorsions locales, ne sont plus présentes après le reclassement.

TAB. 8.7 – Résultats du reclassement des traductions générées avec les contraintes de distorsion. Le modèle de reclassement utilise des traits de couplage et des traits génériques de cohésion et non-cohésion. Le modèle de reclassement est entraîné avec une pseudo-référence wlpBLEU.

	NIST	BLEU	wlpBLEU
baseline	7,0728	0,2626	0,3777
Perceptron wlpBLEU	7,0332	0,2594	0.3778
SVM wlpBLEU	7,0418	0,2597	0,3784
SVM man éval	7,0510	0,2611	0,3781

TAB. 8.8 – Évaluation manuelle des traductions reclassées avec et sans contraintes de distorsion. + : nombre de traductions jugées meilleures que la traduction de base, - : nombre des traductions jugées moins bonnes que la traduction de base. Cette évaluation est faite sur un échantillon de 300 phrases extraites à partir du corpus de test.

	+	-	total diff
SVM man éval	22	10	80

8.2 Expériences de traduction avec le modèle de décomposition

8.2.1 Description des expériences

Dans cette série d'expériences, nous nous sommes concentrée seulement sur l'effet de la décomposition sur la qualité de la traduction finale, sans appliquer le modèle de reclassement.
Ainsi, la traduction est faite par les étapes suivantes :

– décomposition d'une phrase source qui donne une phrase source simplifiée et ses fragments substitués ;
– traduction de la phrase source simplifiée (traduction simplifiée) et de ces fragments substitués (sous-traductions) ;
– composition de la phrase cible finale à partir de la traduction simplifiée et des sous-traductions.

Les paramètres du modèle de chaque étape ont été optimisés sur le corpus parallèle *dev1* (section 5.1.1.1).

8.2.1.1 Modèle de décomposition

À cause de contraintes de temps, nous n'avons pas utilisé les jugements humains comme critère lors de la création de l'ensemble d'entraînement annoté. Ainsi, le seul critère utilisé dans ces expériences a été la mesure wlpBLEU. Le modèle de décomposition a été entraîné par SVM.

8.2.1.2 Modèles de traduction

Modèle de sous-traduction. Pour traduire des fragments, nous utilisons le modèle de traduction de base. Afin de pouvoir l'adapter mieux au contexte de la traduction par décomposition, le modèle de base génère la liste des N meilleures sous-traductions. C'est le modèle de recomposition de la traduction finale qui choisit ensuite dans cette liste des meilleures sous-traductions pour compléter la traduction réduite.

Modèle de traduction simplifiée. L'adaptation du modèle de traduction a été faite en adaptant seulement les paramètres du modèle log-linéaire. La création d'une bibliothèque des bi-fragments adaptés, et l'entraînement d'un modèle de langage adapté n'ont pas pu être réalisés faute de temps.

Pour adapter les paramètres du modèle de traduction, nous avons utilisé un entraînement du type MERT. Pour chaque phrase source s, la liste des N meilleures traduction à chaque itération est générée comme suit :

1. le modèle de décomposition appris précédemment choisit les bi-fragments à simplifier dans la phrase source ;
2. la liste des N meilleures traductions est générée pour la phrase source simplifiée ;
3. pour chaque traduction simplifiée de cette liste, une traduction finale est générée par un modèle de recomposition ; cette liste est ajoutée à l'ensemble d'entraînement du MERT.

Recomposition de la traduction finale. Le but du modèle de recomposition est de choisir des meilleures sous-traductions de la liste générée par un modèle de base.
Nous attribuons les traits suivants pour chaque sous-traduction :
– traits attribués par le modèle de traduction ;
– trait d'intersection : le nombre de mots en commun entre une sous-traduction et le fragment de la traduction simplifiée qu'elle remplace. La motivation de ce trait est la suivante : le fragment dans la traduction simplifiée a été traduit dans le contexte, tandis que la sous-traduction est obtenue en dehors du contexte. Si une sous-traduction contient exactement le même fragment, il est plus probable qu'elle soit mieux adaptée au contexte.
– trait du modèle de langage local : la valeur du modèle de langage 3-gramme calculé pour une sous-traduction dans le contexte local de la traduction simplifiée (les 2 mots qui se trouvent avant et après le fragment correspondant dans une traduction simplifiée).

Le modèle de recomposition est un modèle linéaire qui recombine les traits décrits ci-dessus. L'optimisation des paramètres de ce modèle est faite par l'algorithme de MERT.

8.2.2 Résultats

Faute de temps, nous n'avons fait qu'une itération de la procédure d'entraînement, en mettant à jour tout de même les paramètres de chaque étape.

TAB. 8.9 – Résultats de traduction avec le modèle de décomposition

	NIST	BLEU
baseline	6.9933	0.2674
modèle de décomposition	6.9536	0.2613

TAB. 8.10 – Évaluation manuelle des traductions générées avec le modèle de décomposition. + : nombre de traductions jugés meilleures par rapport à la traduction de base, - : nombre des traduction jugées moins bonnes par rapport à la traduction de base. Cette évaluation est faite sur un échantillon de 100 phrases extraites à partir du corpus de test

	+	-	total diff
décomposition	16	16	52

Analyse des résultats. Les tableaux 8.9, 8.10 donnent des résultats d'évaluation de la traduction (par des mesures automatiques et par des jugements humains) en appliquant le modèle de décomposition. Les mesures automatiques sont légèrement dégradées, comme c'était le cas du modèle de contraintes de distorsion. Les évaluations humaines ne montrent pas d'amélioration ni de dégradation significatives. Ainsi, les résultats de ces expériences sont moins concluants que les résultats des expériences précédentes.

Dans le tableau 8.11, nous montrons quelques exemples. Ce sont les exemples les plus démonstratifs ; ils ont pour but de montrer quelques problèmes principaux rencontrés lors de la traduction avec le modèle de décomposition.

Les deux premiers exemples sont des exemples d'améliorations obtenues en appliquant la traduction par décomposition. Notons que ces améliorations n'avaient pas été trouvées par l'introduction de contraintes de distorsion.

Les trois exemples qui suivent sont des exemples de trois sources différentes d'erreurs de la traduction par décomposition. Pour chaque phrase, nous donnons la phrase réduite obtenue par le modèle de décomposition, et la traduction réduite de cette phrase. Le fragment substitué est mis en gras dans la phrase réduite et dans sa traduction.

Dans le premier exemple, après l'introduction de la sous-traduction dans une traduction réduite, les deux articles apparaissent ensemble (the a). Cette erreur est due à l'analyse erronée dès le départ : une bonne analyse syntaxique devrait attacher *l'* et *un* dans la phrase source.

Dans le deuxième exemple, c'est la sous-traduction qui est à l'origine de la dégradation obtenue. Comme nous l'avons remarqué plus tôt, la traduction des fragments substituts est faite en dehors du contexte. De plus, ni le modèle de traduction, ni le modèle de langage ne sont adaptés à la traduction des fragments (qui ne forment pas nécessairement des phrases). Ainsi, le modèle de langage a préféré la sous-traduction *the map is fiscal*. L'introduction de cette sous-traduction dans le contexte forme une traduction erronée.

Dans le troisième exemple, l'erreur de la traduction finale vient de la traduction réduite. Cette erreur, comme précédemment, est due au manque du contexte et au modèle de traduction qui n'est pas adapté à la traduction réduite.

Conclusion. Les résultats initiaux sur la traduction par un modèle de décomposition donnent des résultats moins encourageants que ceux du modèle utilisant les contraintes de distorsion. Notons que ce ne sont ici que des expériences préliminaires, et elles ne permettent pas d'obtenir une conclusion définitive sur la démarche.

Nous nous sommes servies de ces résultats initiaux pour analyser des problèmes typiques du modèle. Nous citons ci-dessous certaines sources possibles de ces problèmes, et proposons quelques extensions du modèle qui, nous pensons, peuvent l'améliorer.

– Une analyses syntaxique complète est cruciale quand il s'agit de la décomposition de la phrase initiale. Par exemple, si un mauvais replacement a été fait, cette erreur aura un impact sur la qualité de la traduction simplifiée, et par conséquent, sur la qualité de la traduction finale. Lors de la traduction simplifiée, nous n'avons pas de contexte total pour certains fragments, et le manque du contexte (ou le contexte erroné) peut mener à des erreurs irréparables aux étapes suivantes. Pour un contrôle supplémentaire de la qualité de la traduction réduite, il serait intéressant d'appliquer un modèle de reclassement à cette étape, afin d'assurer que la traduction simplifiée conserve la bonne structure.

– Une phrase simplifiée est différente d'une phrase standard du corpus. La simplification suppose l'élimination des adjectifs et des adverbes, en gardant uniquement le squelette de la phrase initiale. Ni le modèle de traduction, ni le modèle de langage ne sont entraînés sur ce type de corpus. Ces modèles ne sont probablement pas adaptés à la traduction d'une phrase réduite. La création d'un corpus adapté comme proposé dans la section 7.3.2 peut être une étape importante.

– Le modèle de décomposition est une première étape de notre modèle, et les erreurs faites à cette étape ne peuvent pas être corrigées aux étapes suivantes. Ainsi, la notion de "qualité" utilisée lors de l'entraînement du modèle de décomposition peut être cruciale. Comme nous l'avons vu dans le cas du modèle de contraintes de distorsion, les jugements humains de la qualité de traduction lors de l'entraînement de la première étape donnent des meilleurs résultats. Cette étape est encore plus importante pour la décomposition, car après un mauvais remplacement, un contexte important peut être perdu.

TAB. 8.11: Exemples de traductions obtenues par le modèle de décomposition. La colonne *source réduite* montre une décomposition obtenue par notre modèle : les fragments substitué sont en italique. Nous avons mis en gras les parties alignés entre la phrase réduite et sa traduction réduite. La traduction finale est obtenue en remplaçant un substitut (fragment en gras dans la traduction réduite) par une traduction complète de fragment correspondant (sous-traduction). La différence entre la traduction de base et nouvelle traduction est mis en gras.

source	source réduite	trad réduite et sous-traduction	système de base	trad recomposée
Mais toute tentative visant à limiter la discrétion fiscale du gouvernement fédéral américain à la manière du Pacte de stabilité (par exemple, la loi Gramm/Rudman tristement célèbre de l'ère Clinton) s'écroule toujours à la fin devant **la pression du président et du Congrès.**	**écroule toujours** {*Mais toute tentative visant à limiter la discrétion fiscale du gouvernement fédéral américain à la manière du Pacte de stabilité (par exemple, la loi Gramm/Rudman tristement célèbre de l'ère Clinton) s'écroule toujours*} à la fin devant la **pression** {*la pression du président et du Congrès.*} .	Trad réduite : **écroule always** at the end to the **pressure** . .Sous-traduction : but any attempt to limit the fiscal discretion of the us federal government in the manner of the stability pact (for example , the law gramm / rudman notorious clinton era) if **écroule always** Sous-traduction : **pressure** of the president and congress	but any attempt to limit the fiscal discretion of the us federal government in the manner of the stability pact (for example , the law gramm / rudman notorious clinton era) if écroule always at the end **to the pressure and president of congress .**	but any attempt to limit the fiscal discretion of the us federal government in the manner of the stability pact (for example , the law gramm / rudman infamous of the clinton era) if écroule always at the end **to the pressure of the president and congress** .

De façon réaliste , la Commission peut uniquement demander aux **gouvernements nationaux européens** de faire en sorte que leurs comptes fiscaux soient transparents et clairs .	De façon réaliste , la Commission peut uniquement demander aux **gouvernements** {*gouvernements nationaux européens*} de faire en sorte que leurs comptes fiscaux soient transparents et clairs .	Trad réduite : realistically , the commission can only ask the **governments** to ensure that their fiscal accounts are transparent and clear . .Sous-traduction : european national governments	realistically , the commission can only ask the **national governments europeans** to ensure that their fiscal accounts are transparent and clear . it is impossible to prevent the use of fiscal policy .	realistically , the commission can only ask the **european national governments** to ensure that their fiscal accounts are transparent and clear . it is impossible to prevent the use of fiscal policy .
Avec 80 millions d'habitants, la Turquie **deviendrait** l'un des plus grands Etats membres de l'Union européenne.	**deviendrait** {*avec 80 millions d' habitants , la turquie deviendrait l'*} un des plus grands etats membres de l' union européenne . [4]	Trad réduite : **would** a major member states of the european union . Sous-traduction : with 80 million people , turkey **would become the**	with 80 million inhabitants , turkey **becoming** one of the largest member states of the european union .	with 80 million people , turkey **would become** *the a* major member states of the european union .
La carte fiscale **joue** un rôle critique dans le choix des groupes d'électeurs à acheter.	**joue** {*La carte fiscale joue*} **un** rôle critique dans le choix des groupes d' électeurs à acheter .	Trad réduite : **a** critical role in the choice of voters to buy Sous-traduction : the map is fiscal	the fiscal card a critical role in the choice of voters to buy .	the map is fiscal critical role in the choice of voters to buy .
Les autorités fiscales indépendantes (ou les règles fiscales telles que le Pacte de stabilité) gênent cet usage politique de la politique fiscale, qui s'avère coûteux.	**gênent** {*les autorités fiscales indépendantes (ou les règles fiscales telles que le pacte de stabilité) gênent*} cet usage politique de la politique fiscale , qui s' avère coûteux .	Trad réduite : this policy **ineffective** use fiscal policy , which is expensive . Sous-traduction : independent fiscal authorities (or fiscal rules such as the stability pact) ineffective	independent fiscal authorities (or fiscal rules such as the stability pact) this policy ineffective use fiscal policy , which is expensive .	this policy the fiscal authorities independent (or fiscal rules such as the stability pact) ineffective use fiscal policy , which is expensive .

[4] C'est typiquement un exemple où une phrase décomposée n'est pas de même nature que le corpus d'entraînement. Ce type de décomposition représente un problème lors de la traduction avec un modèle de base.

Conclusion et perspectives

Nous allons maintenant conclure l'ensemble des travaux présentés dans ce mémoire. Ces travaux sont concentrés autour du modèle de TA à fragments. C'est un modèle de traduction probabiliste, qui représente une base de référence forte parmi les systèmes de TA probabiliste aujourd'hui. La performance des modèles probabilistes est directement liée à la taille du corpus. Ainsi, si un corpus de taille suffisante n'est pas disponible, la performance du système est réduite.

Motivation

Le but de nos travaux était d'étudier les possibilités d'enrichissement du modèle de TA probabiliste à fragments à l'aide d'informations syntaxiques (bilingues et monolingues). Nous nous sommes essentiellement placée dans le cas d'un corpus parallèle relativement petit (News Commentary, 55 000 phrases parallèles). Notre but principal était d'étudier si l'enrichissement du modèle de TA à fragments à l'aide de connaissances syntaxiques expertes peut améliorer la performance du système dans le cas d'un tel corpus. Un tel enrichissement vise à résoudre certains problèmes spécifiques que nous avons observés avec le modèle à fragments (ces problèmes, bien que moins fréquents, peuvent être observés même dans le cas d'un corpus parallèle de grand taille). Le but principal des informations syntaxiques est d'améliorer la structure syntaxique de la traduction finale, et de rendre cette traduction plus compréhensible (plus adéquate et, par conséquent, souvent plus fluide).

Reclassement des traductions

Nous avons tout d'abord proposé d'étendre le modèle de TA à fragments à l'aide d'un modèle de reclassement utilisant des traits syntaxiquement motivés. Les traits que nous avons introduits sont basés sur les analyses de dépendances produites par XIP, et nous partons de l'hypothèse que l'adéquation de la traduction est proportionelle à la similarité entre la structure de dépendances source et la structure de dépendances cible. Nous avons introduit également des traits dépendant de l'analyse de la phrase cible, uniquement pour avoir un contrôle supplémentaire sur la fluidité.

Nous avons validé notre hypothèse, en effectuant des évaluations humaines (subjectives) détaillées pour des traductions de l'anglais vers le français, et vice-versa. Ainsi, ces évaluations ont montré que, grâce à l'application du modèle de reclassement enrichi par des traits syntaxiques, l'adéquation a été améliorée pour 55% des phrases dans le cas de la traduction vers l'anglais, et pour 42% dans le cas de la traduction vers le français. Notons que les améliorations en termes de fluidité sont encore plus élevées : une traduction plus adéquate est très souvent jugée plus fluide, tandis que l'inverse n'est pas nécessairement vrai.

D'après les résultats d'évaluation humaine, les meilleurs résultats de reclassement sont obtenus par des modèles entraînés avec de vraies traductions de référence, et non avec des pseudo-références, choisies par une des mesures automatiques. Il est difficile d'interpréter le rapport entre les scores attribués au modèle par les mesures automatiques et par les jugements humains : nous observons très peu de corrélation entre ces mesures lorsqu'on compare différents modèles de reclassement. La nature de la traduction de référence (peu littérale) en est sûrement une des raisons. L'utilisation de traductions post-éditées en tant que références permettrait probablement d'améliorer non seulement cette corrélation, mais aussi les

résultats du reclassement.

Les expériences sur le reclassement nous ont montré que les traits les plus intéressants pour l'adéquation sont des traits basés sur des relations de dépendance étiquetées. Cette observation est intéressante et peut être utile si nous souhaitons élaborer un nouveau modèle de traduction, basé sur les analyses de dépendances (par exemple, le modèle de traduction par des graphelets ébauché dans la section 2.3.1).

Simplification du processus de la traduction à l'aide de l'analyse syntaxique source

La seconde façon d'enrichir un modèle à fragments que nous avons proposée est basée sur l'idée suivante. La connaissance de la structure syntaxique source peut être utile pour décomposer le processus de traduction en étapes plus simples. En effet, le traducteur humain, quand il effectue la traduction, commence par détecter la structure générale de la phrase source, en détectant le sujet, le verbe, les objets. Une fois que la structure est détectée, certains fragments peuvent être traduits de façon autonome, et ensuite intégrés dans la structure cible, pour obtenir la traduction finale.

Nous avons proposé un modèle qui pourrait imiter (à un niveau très primitif) cette démarche. Ce modèle, à partir d'une analyse syntaxique source (en constituants), propose des fragments qui peuvent être traduits de façon autonome. Le choix de ces fragments est fait par un modèle log-linéaire combinant des traits basés sur l'analyse syntaxique et sur des statistiques calculées sur le corpus monolingue. Les paramètres de ce modèle sont optimisés sur la base du corpus parallèle, en maximisant la qualité de la traduction finale.

Une première réalisation de cette idée consiste seulement à limiter la permutation des mots à l'intérieur des fragments en se basant sur la structure syntaxique de la phrase source. D'après les premières évaluations humaines, ces contraintes de permutation des mots permettent d'éviter certaines erreurs de traduction (erreurs du modèle, et erreurs de recherche).

Nous avons analysé certaines erreurs typiques faites par ce modèle. Bien qu'une partie des erreurs soit due à l'analyse syntaxique incomplète proposée par XIP, ce ne sont pas les sources majeures du problème. Les erreurs les plus fréquentes viennent du manque de permutations locales, et des choix lexicaux moins bons que ceux proposés par le modèle de base. Ainsi, la plupart des erreurs introduites par le modèle de contraintes de distorsion peuvent être corrigées en adaptant le modèle de traduction à la traduction avec des contraintes de distorsion (plutôt que d'utiliser le modèle de base). De plus, certaines de ces erreurs peuvent déjà être résolues en appliquant le modèle de reclassement avec des traits de couplage.

Une deuxième réalisation consiste à simplifier la phrase source, en remplaçant certains fragments complexes par des fragments plus simples. Ces fragments plus simples sont choisis dans l'arbre de constituants produit par XIP. Ainsi, la tâche de traduction se décompose en les sous-tâches suivantes :

- traduction d'une phrase simplifiée ;
- sous-traductions : traduction des fragments qui ont été remplacés dans la phrase d'origine ;
- composition de la traduction finale à partir de la traduction de la phrase simplifiée et des sous-traductions.

Nos premières expériences avec le modèle de décomposition ont montré des améliorations de même nature que celles obtenues avec le modèle introduisant des contraintes de distorsion. Par contre, nous avons aussi observé plus de dégradations liées principalement au manque de contexte des sous-traductions, et au fait que le modèle de traduction originel est mal adapté au problème de la traduction simplifiée. Notons qu'il ne s'agit ici que d'expériences préliminaires, et que certains aspects (tels que l'adaptation du corpus parallèle, par exemple) n'ont pas encore été pris en compte à ce stade. Aussi, d'autres facteurs, tels que l'analyse incomplète de constituants fournie par XIP, ou l'utilisation de mesures automatiques lors de l'entraînement du modèle de décomposition, sont sûrement intervenus dans ce résultat. Il reste donc plusieurs pistes prometteuses à explorer dans cette direction.

Extensions de nos approches et perspectives

Nos travaux ont montré que le modèle de TA à fragments enrichi par des connaissances sur la structure syntaxique source et/ou cible permet souvent de rendre la traduction finale plus adéquate et fluide (dans les limites de ses possibilités) dans l'ensemble. Nous avons proposé différentes techniques pour intégrer des types d'information syntaxique différents. L'analyse des résultats nous a permis de détecter les avantages et les défauts de chacune de ces techniques, ainsi que les types d'information les plus intéressants à intégrer dans le modèle de traduction.

Les approches proposées dans cette thèse peuvent être étendues à d'autres modèles que le modèle à fragments sur lequel nous nous sommes concentrée. Ainsi, toutes les méthodes que nous avons proposées peuvent être facilement généralisées à des systèmes de TA différents du modèle probabiliste à fragments. Plus généralement, l'application de ces techniques à la combinaison de traductions provenant de systèmes différents pourrait être une piste intéressante.

Aussi, le modèle de décomposition a été proposé comme extension du modèle à fragments. En particulier, dans nos expériences, les sous-traductions sont produites par un modèle à fragments. Dans le cas général, cette approche peut être appliquée à n'importe quel autre système de TA. De plus, un analyseur différent de XIP peut être utilisé (les traits du modèle doivent être adaptés en fonction de l'analyseur). Notons tout de même que l'intérêt de ce modèle est d'introduire implicitement une analyse syntaxique source dans un modèle qui n'en utilise pas. Ainsi, il serait intéressant de l'appliquer à des systèmes ayant une architecture linguistique directe ou semi-directe. Cette approche peut également servir pour combiner des systèmes de TA différents.

Une autre variante du modèle de décomposition est la suivante. Nous pouvons créer une bibliothèque des bi-fragments dynamique, en fonction de la décomposition initiale. La génération de cette bibliothèque se fait par une pré-traduction de fragments choisis par le modèle de décomposition. Cette variante ne nécessite pas de génération de la phrase source simplifiée. Par conséquent, le contexte perdu lors de la traduction réduite, est gardé par cette variante. De plus, cette variante est plus résistante aux erreurs de l'analyse syntaxique : la traduction de base reste toujours accessible. En même temps, ce modèle perd la capacité de simplifier la phrase initiale. De plus, l'extension de la bibliothèque de bi-fragments avec des bi-fragments dynamiques augmente la taille du graphe des hypothèses. Il est tout de même possible que, grâce aux bi-fragments dynamiques, les hypothèses plus intéressantes soient explorées d'abord.

Pour conclure ce mémoire, nous souhaitons rappeler que l'introduction de connaissances syntaxiques dans le cadre des modèles de TA probabiliste est un problème à la fois intéressant et complexe. Beaucoup de facteurs interviennent dans la qualité finale de la traduction : la qualité de l'analyse syntaxique, la performance du modèle de base, la qualité de l'apprentissage des paramètres, la notion de "qualité de traduction" (mesure automatique ou jugement humain) adoptée lors d'apprentissage. Tous ces facteurs rendent très complexe l'analyse des erreurs de traduction, et la détection de leur origine. Ainsi, lors de l'analyse des traductions finales, il est difficile de dire sûrement à quelle étape une erreur s'est produite. C'est un problème de recherche ouvert que de développer des techniques permettant de détecter ces erreurs ou au moins d'atténuer leur propagation dans des systèmes hybrides tels que les modèles que nous avons présentés.

Bibliographie

Aït-Mokhtar, S., Chanod, J.-P. and Roux, C. [2002], 'Robustness beyond shallowness : incremental deep parsing', *Natural Language Engineering* **8**(3), 121–144.

Avramidis, E. and Koehn, P. [2008], Enriching morphologically poor languages for statistical machine translation, *in* 'Proceedings of ACL-08 : HLT', Association for Computational Linguistics, Columbus, Ohio, pp. 763–770.
URL: *http://www.aclweb.org/anthology/P/P08/P08-1087.pdf*

Banerjee, S. and Lavie, A. [2005], METEOR : an automatic metric for MT evaluation with improved correlation with human judgments, *in* 'Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization', Association for Computational Linguistics, Ann Arbor, Michigan, pp. 65–72.
URL: *http://www.aclweb.org/anthology/W/W05/W05-0909*

BinLio and Hancock, E. R. [2001], 'Structural graph matching using EM algorithm and singular value decomposition', *IEEE Transaction on pattern analysis and Machine Intelligence* **23**(10), 1120 – 1136.

Blanchon, H. and Boitet, C. [2007], 'Pour l'évaluation des systèmes de TA par des méthodes externes fondées sur la tâche', *TAL* **48**(1), 33–65.

Boitet, C. [1993], 'Ta et tao à grenoble... 32 ans déjà !', *TAL (revue semestrielle de l'ATALA)* **33**(1-2), 45–84.

Boitet, C. [2008], Les architechtures linguistiques et computationnelles en traduction automatique sont indépendantes, *in* 'TALN', Avignon, France.

Brown, P. F. and Della Pietra, S. A. [1993], 'The mathematics of statistical machine translation : parameter estimation.', *Computational Linguistics* **19(2)**, 263–311.

Callison-Burch, C., Koehn, P., Monz, C. and Schroeder, J. [2009], Findings of the 2009 Workshop on Statistical Machine Translation, *in* 'Proceedings of the Fourth Workshop on Statistical Machine Translation', Association for Computational Linguistics, Athens, Greece, pp. 1–28.
URL: *http://www.aclweb.org/anthology/W/W09/W09-0x01*

Callison-Burch, C. and Osborne, M. [2006], Re-evaluating the role of BLEU in Machine Translation research, *in* 'Proceedings of EACL', Vol. 2006, pp. 249–256.

Cancedda, N., Dymetman, M., Mahe, P., Rousu, J., Saunders, C. and Vuorinen, M. [2008], SMART deliverable 3.1 : String and rational kernels for language modelling., Technical report.
URL: *http://www.smart-project.eu/files/D31.pdf*

Cherry, C. [2008], Cohesive phrase-based decoding for statistical machine translation, *in* 'Proceedings of ACL-08 : HLT', Association for Computational Linguistics, Columbus, Ohio, pp. 72–80.
URL: *http://www.aclweb.org/anthology/P/P08/P08-1009*

Chiang, D. [2005], A hierarchical phrase-based model for statistical machine translation, *in* 'ACL'05 : Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, Ann Arbor, Michigan, pp. 263–270.

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

Collins, M. [2001], Ranking algorithms for named-entity extraction : boosting and the voted perceptron, *in* 'ACL'02 : Proceedings of the 40th Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 489–496.

Collins, M. [2002], Discriminative training methods for Hidden Markov Models : theory and experiments with perceptron algorithms, *in* 'EMNLP'02 : Proceedings of the ACL-02 conference on Empirical methods in natural language processing', Association for Computational Linguistics, pp. 1–8.

Collins, M. [2003], 'Head-driven statistical models for natural language parsing', *Comput. Linguist.* **29**(4), 589–637.

Collins, M., Koehn, P. and Kucerova, I. [2005], Clause restructuring for statistical machine translation, *in* 'ACL'05 : Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, Ann Arbor, Michigan, pp. 531–540.

Collins, M. and Roark, B. [2004], Incremental parsing with the perceptron algorithm, *in* 'ACL'04 : Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, Barcelona, Spain, pp. 111–118.

Cowan, B., Kucerova, I. and Collins, M. [2006], A discriminative model for tree-to-tree translation, *in* 'Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)', Association for Computational Linguistics, Sydney, Australia, pp. 232–241.
URL: *http://acl.ldc.upenn.edu//W/W06/W06-1628.pdf*

Crego, J. M. and Habash, N. [2008], Using shallow syntax information to improve word alignment and reordering for SMT, *in* 'Proceedings of the Third Workshop on Statistical Machine Translation', Association for Computational Linguistics, Columbus, Ohio, pp. 53–61.
URL: *http://www.aclweb.org/anthology/W/W08/W08-0307*

DeNeefe, S., Knight, K., Wang, W. and Marcu, D. [2007], What can syntax-based MT learn from phrase-based MT ?, *in* 'Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)', Association for Computational Linguistics, Prague, Czech Republic, pp. 755–763.
URL: *http://www.aclweb.org/anthology/D/D07/D07-1079.pdf*

Doddington, G. [2002], Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics, *in* 'Proceedings of the second international conference on Human Language Technology Research', Morgan Kaufmann Publishers Inc., San Diego, California, pp. 138–145.

Dolan, W. B., Pinkham, J. and Richardson, S. D. [2002], MSR-MT : The Microsoft Research Machine Translation System, *in* 'AMTA '02 : Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation : From Research to Real Users', Springer-Verlag, London, UK, pp. 237–239.

Dugast, L., Senellart, J. and Koehn, P. [2007], Statistical post-editing on SYSTRAN's rule-based translation system, *in* 'Proceedings of the Second Workshop on Statistical Machine Translation', Association for Computational Linguistics, Prague, Czech Republic, pp. 220–223.
URL: *http://www.aclweb.org/anthology/W/W07/W07-0232*

Fossum, V., Knight, K. and Abney, S. [2008], Using syntax to improve word alignment precision for syntax-based machine translation, *in* 'Proceedings of the Third Workshop on Statistical Machine Translation', Association for Computational Linguistics, Columbus, Ohio, pp. 44–52.
URL: *http://www.aclweb.org/anthology/W/W08/W08-0306*

Fox, H. J. [2002], Phrasal cohesion and statistical machine translation, *in* 'In Proceedings of EMNLP-02', pp. 304–311.

Frank, R. [2002], *Phrase Structure Composition and Syntactic Dependencies*, Current studies in linguistics - Series, MIT PRESS, Great Britain.

Freund, Y. and Schapire, R. [1999], 'Large margin classification using the perceptron algorithm.', *Machine Learning* **37(3)**, 277–296.

Germann, U., Jahr, M., Knight, K., Marcu, D. and Yamada, K. [2004], 'Fast and optimal decoding for Machine Translation', *Artif. Intell.* **154**(1-2), 127–143.

Goldwater, S. and McClosky, D. [2005], Improving statistical MT through morphological analysis, *in* 'Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Vancouver, British Columbia, Canada, pp. 676–683.
URL: *http://www.aclweb.org/anthology/H/H05/H05-1085*

Goutte, C., Yamada, K. and Gaussier, E. [2004], Aligning words using matrix factorisation, *in* 'ACL'04 : Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, Barcelona, Spain, p. 502.

Gu, L. and Gao, Y. [2004], On feature selection in maximum entropy approach to statistical concept-based speech-to-speech translation, *in* 'Proc. of the International Workshop on Spoken Language Translation', Kyoto, Japan, pp. 115–121.

Habash, N. Y. [2003], Generation-heavy hybrid machine translation, PhD thesis, College Park, MD, USA. Director-Dorr, Bonnie J.

Hewavitharana, S., Lavie, A. and Vogel, S. [2007], Experiments with a noun-phrase driven statistical machine translation system, *in* 'Proceedings of MT Summit XI', Copenhagen, Denmark.

Jelinek, F. [1969], 'Fast sequential decoding algorithm using a stack', *j-IBM-JRD* **13**(6), 675–685.

Kääriäinen, M. [2009], Sinuhe – statistical machine translation using a globally trained conditional exponential family translation model, *in* 'Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Singapore, pp. 1027–1036.
URL: *http://www.aclweb.org/anthology/D/D09/D09-1107*

Koehn, P. [2003], Noun phrase translation, PhD thesis, Los Angeles, CA, USA. Adviser-Knight, Kevin.
URL: *http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/thesis-readable.ps*

Koehn, P. [2004], Pharaoh : A beam search decoder for phrase-based statistical Machine Translation models., *in* 'Proceedings of AMTA'.

Koehn, P. [2005], Europarl : A multilingual corpus for evaluation of machine translation, MT Summit.
URL: *http://www.statmt.org/europarl/*

Koehn, P. and Hoang, H. [2007], Factored translation models, *in* 'Proceedings of the EMNLP-CoNLL 2007', Association for Computational Linguistics, Prague, Czech Republic, pp. 868–876.
URL: *http://www.aclweb.org/anthology/D/D07/D07-1091*

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. [2007], Moses : open source toolkit for statistical machine translation, *in* 'ACL'07 : Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions', Association for Computational Linguistics, Prague, Czech Republic, pp. 177–180.

Koehn, P., Och, F. J. and Marcu, D. [2003], Statistical phrase-based translation, *in* 'NAACL'03 : Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology', Association for Computational Linguistics, Edmonton, Canada, pp. 48–54.

Landis, J. R. and Koch, G. G. [1977], 'The measurement of observer agreement for categorical data.', *Biometrics* **33**(1), 159–174.
URL: *http://view.ncbi.nlm.nih.gov/pubmed/843571*

Lavie, A. [2008], Stat-XFER : A general search-based syntax-driven framework for machine translation, *in* 'Proceedings of CICLing-2008', pp. 362–375.

Lepage, Y. and Denoual, E. [2005], 'Purest ever Example-Based Machine Translation : detailed presentation and assessment', *Machine Translation* **19**(3-4), 251–282.

Lewis, P. M. and Stearns, R. E. [1966], Syntax directed transduction, *in* 'SWAT'66 : Proceedings of the 7th Annual Symposium on Switching and Automata Theory (SWAT 1966)', IEEE Computer Society, Washington, DC, USA, pp. 21–35.

Liang, P., Bouchard-Côté, A., Klein, D. and Taskar, B. [2006], An end-to-end discriminative approach to machine translation, *in* 'ACL-44 : Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Sydney, Australia, pp. 761–768.

Lopez, A. [2008], 'Statistical machine translation', *ACM Comput. Surv.* **40**(3), 1–49.

Ma, Y., Ozdowska, S., Sun, Y. and Way, A. [2008], Improving word alignment using syntactic dependencies, *in* 'Proceedings of the ACL-08 : HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)', Association for Computational Linguistics, Columbus, Ohio, pp. 69–77.
URL: *http://www.aclweb.org/anthology/W/W08/W08-0409*

Mahé, P. and Cancedda, N. [2008], Factored sequence kernels, *in* 'ESANN 2008, 16th European Symposium on Artificial Neural Networks', pp. 409–414.
URL: *http://www.dice.ucl.ac.be/Proceedings/esann/esannpdf/es2008-53.pdf*

Marcu, D., Wang, W., Echihabi, A. and Knight, K. [2006], SPMT : Statistical machine translation with syntactified target language phrases, *in* 'Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Sydney, Australia, pp. 44–52.
URL: *http://www.aclweb.org/anthology/W/W06/W06-1606*

Mellebeek, B., Owczarzak, K., Groves, D., Van Genabith, J. and Way, A. [2006], A syntactic skeleton for statistical machine translation, *in* '11th Annual Conference of the European Association for Machine Translation', Oslo, Norway, pp. 195–202.

Nagao, M. [1984], A framework of a mechanical translation between Japanese and English by analogy principle, *in* 'Proc. of the international NATO symposium on Artificial and human intelligence', Elsevier North-Holland, Inc., New York, NY, USA, pp. 173–180.

Niessen, S. and Ney, H. [2000], Improving SMT quality with morpho-syntactic analysis, *in* 'Proceedings of the 18th conference on Computational linguistics', Association for Computational Linguistics, Saarbrücken, Germany, pp. 1081–1085.

Nießen, S., Och, F. J., Leusch, G. and Ney, H. [2000], An evaluation tool for machine translation : fast evaluation for MT research, *in* 'LREC-2000 : Second International Conference on Language Resources and Evaluation.', Athens, Greece, pp. 39–45.

Nikoulina, V. and Dymetman, M. [2008], Experiments in discriminating phrase-based translations on the basis of syntactic coupling features, *in* 'Proceedings of the ACL-08 : HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)', Association for Computational Linguistics, Columbus, Ohio, pp. 55–60.
URL: *//www.aclweb.org/anthology-new/W/W08/W08-0407*

Och, F. J. [2003], Minimum error rate training in statistical machine translation, *in* 'ACL'03 : Proceedings of the 41st Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, Sapporo, Japan, pp. 160–167.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z. and Radev, D. [2003], Syntax for Statistical Machine Translation : Final report of John Hopkins 2003 Summer Workshop, Technical report, John Hopkins University.

Och, F. J. and Ney, H. [2003], 'A systematic comparison of various statistical alignment models', *Computational Linguistics* **29**(1), 19–51.

Och, F. J. and Ney, H. [2004], 'The alignment template approach to statistical machine translation', *Comput. Linguist.* **30**(4), 417–449.

Och, F. J., Ueffing, N. and Ney, H. [2001], An efficient A* search algorithm for statistical machine translation, *in* 'Proceedings of the workshop on Data-driven methods in machine translation', Association for Computational Linguistics, Toulouse, France, pp. 1–8.

Och, F. and Ney, H. [2002], 'Discriminative training and maximum entropy models for statistical machine translation.', *In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* pp. 295–302.

Oepen, S., Velldal, E., Lønning, J. T., Meurer, P., Rosén, V. and Flickinger, D. [2007], Towards hybrid quality-oriented machine translation — on linguistics and probabilities in MT, *in* 'Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)', Skövde, Sweden.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. [2002], BLEU : a method for automatic evaluation of machine translation, *in* 'ACL'02 : Proceedings of the 40th Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 311–318.

Popovic, M. and Ney, H. [2009], Syntax-oriented evaluation measures for machine translation output, *in* 'Proceedings of the Fourth Workshop on Statistical Machine Translation', Association for Computational Linguistics, Athens, Greece, pp. 29–32.
URL: `http://www.aclweb.org/anthology/W/W09/W09-0x02`

Quirk, C., Menezes, A. and Cherry, C. [2005], Dependency treelet translation : Syntactically informed phrasal SMT, *in* 'Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)', Association for Computational Linguistics, Ann Arbor, Michigan, pp. 271–279.
URL: `http://www.aclweb.org/anthology/P/P05/P05-1034`

Roark, B., Saraclar, M., Collins, M. and Johnson, M. [2004], Discriminative language modeling with conditional random fields and the perceptron algorithm, *in* 'Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)'.

Rosenblatt, F. [1958], 'The perceptron : a probabilistic model for information storage and organization in the brain', *Psychological Review* **65**(6), 386–408.

Sadat, F. and Habash, N. [2006], Combination of Arabic preprocessing schemes for statistical machine translation, *in* 'ACL-44 : Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Sydney, Australia, pp. 1–8.

Sakamoto, Y., Ishikawa, T. and Satoh, M. [1986], Concept and structure of semantic markers for machine translation in Mu-project., *in* 'Proceedings of the 11th coference on Computational linguistics', Association for Computational Linguistics, Bonn, Germany, pp. 13–19.

Sánchez-Martínez, F. [2008], Using unsupervised corpus-based methods to build rule-based machine translation systems, PhD thesis, Universitat d'Alicante.

Senellart, J., Yang, J. and Rebollo, A. [2003], SYSTRAN intuitive coding technology, *in* 'MT Summit IX'.

Shen, L. [2004], Discriminative reranking for Machine Translation, *in* 'HLT NAACL 2004', pp. 177–184.

Shen, L., Xu, J. and Weischedel, R. [2008], A new string-to-dependency machine translation algorithm with a target dependency language model, *in* 'Proceedings of ACL-08 : HLT', Columbus, Ohio, pp. 577–585.
URL: *http://www.aclweb.org/anthology-new/P/P08/P08-1066.pdf*

Simard, M., Cancedda, N., Cavestro, B., Dymetman, M., Gaussier, É., Goutte, C., Yamada, K., Langlais, P. and Mauser, A. [2005], Translating with non-contiguous phrases, *in* 'Proceedings of the HLT/EMNLP'.

Simard, M., Ueffing, N., Isabelle, P. and Kuhn, R. [2007], Rule-based translation with statistical phrase-based post-editing, *in* 'Proceedings of the Second Workshop on Statistical Machine Translation', Association for Computational Linguistics, Prague, Czech Republic, pp. 203–206.
URL: *http://www.aclweb.org/anthology/W/W07/W07-0228*

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. [2006], A study of translation edit rate with targeted human annotation, *in* 'Proceedings of the 7th Conference of AMTA', pp. 223–231.
URL: *http://www.mt-archive.info/AMTA-2006-Snover.pdf*

Specia, L., Turchi, M., Cancedda, N., Dymetman, M. and Cristianini, N. [2009], Estimating the sentence-level quality of machine translation systems, *in* 'Proceedings of the 13th Annual Conference of the EAMT', Barcelona, Spain, p. 28–35.

Stymne, S. [2008], Processing of Swedish compounds for Phrase-Based Statistical Machine Translation, *in* 'Proceedings of EAMT08, European Machine Translation Conference', Hamburg, Germany, pp. 180–189.
URL: *http://www.mt-archive.info/EAMT-2008-Stymne.pdf*

Stymne, S. [2009], Definite noun phrases in Statistical Machine Translation into Danish, *in* 'Proceedings of the Workshop on Extracting and Using Constructions in NLP', Odense, Denmark, pp. 4–9.
URL: *http://www.sics.se/~mange/papers/constructions_workshop.pdf*

Surcin, S., Lange, E. and Senellart, J. [2007], Rapid development of new language pairs at SYSTRAN, *in* 'Proceedings of the Second Workshop on Statistical Machine Translation', MT Summit XI, Copenhagen, Denmark, pp. 443–449.

Tillmann, C. and Ney, H. [2003], 'Word reordering and a dynamic programming beam search algorithm for Statistical Machine Translation', *Comput. Linguist.* **29**(1), 97–133.

Tillmann, C., Vogel, S., Ney, H., Sawaf, H. and Zubiaga, A. [1997], Accelerated DP-based search for statistical translation, *in* 'European Conference on Speech Communication and Technology', Vol. 5, Rhodes, Greece, pp. 2667–2670.
URL: *http://www-i6.informatik.rwth-aachen.de/publications/downloader.php?id=203&row=pdf*

Toutanova, K., Suzuki, H. and Ruopp, A. [2008], Applying morphology generation models to machine translation, *in* 'Proceedings of ACL-08 : HLT', Association for Computational Linguistics, Columbus, Ohio, pp. 514–522.
URL: *http://www.aclweb.org/anthology/P/P08/P08-1059*

Uchida, H., Zhu, M. and Della Senta, T. [2005], *UNL : Universal Networking Language*, UNDL Foundation, Japan.

Ueffing, N. and Ney, H. [2003], Using POS information for Statistical Machine Translation into morphologically rich languages, *in* 'EACL'03 : Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics', Association for Computational Linguistics, Budapest, Hungary, pp. 347–354.
URL: *http://www.mt-archive.info/EACL-2003-Ueffing.pdf*

Vapnik, V. N. [1998], *Statistical Learning Theory*, John Wiley & Sons, Inc., New York.

Vauquois, B. and Boitet, C. [1985], 'Automated translation at Grenoble university', *Compitational Linguistics* **11**(1), 28–36.

Vuorinen, M., Kääriäinen, M., Rousu, J., Wang, Z., Mahé, P. and Cancedda, N. [2008], SMART deliverable 3.2 : Learning methods for discriminative language models., Technical report.
URL: *http://www.smart-project.eu/files/D32.pdf*

Wang, Y.-Y. and Waibel, A. [1997], Decoding algorithm in statistical machine translation, *in* 'Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics', Association for Computational Linguistics, Madrid, Spain, pp. 366–372.

Yamada, K. and Knight, K. [2001], A syntax-based statistical translation model, *in* 'ACL'01 : Proceedings of the 39th Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, Toulouse, France, pp. 523–530.

Yamada, K. and Knight, K. [2002], A decoder for syntax-based Statistical MT, *in* 'ACL'02 : Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 303–310.

Annexe A

Exemples des traductions du système de référence

A.1 Exemples du modèle de TA probabiliste entraîné sur NC enrichi avec Europarl

Notons que les textes source ont suivi un prétraitement (*tokenisation*) :
- les majuscules ont été supprimées ;
- les ponctuations collées ont été décollées ;
- les formes avec élision ont été separées (n'est → n' est).

A.1.1 Anglais - français

id	source	traduction
2	despite occasional grumbles from politicians , no one seriously doubts the european central bank 's independence , or that monetary policy within the euro zone is therefore well insulated from political pressures .	malgré les récriminations des politiciens sérieux , personne ne remet en cause l' indépendance de la banque centrale européenne , ou que la politique monétaire au sein de la zone euro n' est donc bien à l' abri des pressions politiques .
5	the us federal reserve board is , of course , very independent .	la réserve fédérale américaine est , bien sûr , très indépendant .
11	yet monetary policy is not a unique species of economic policymaking .	pourtant , la politique monétaire unique n' est pas une espèce de la politique économique .
12	aspects of fiscal policy could also be delegated to independent bureaucrats in order to keep politicians in line .	les aspects de la politique fiscale pourrait également être délégué à des bureaucrates indépendant , afin de maintenir les politiciens en ligne .
25	however , monetary policy is a much more awkward , indirect and imprecise way of choosing winners and losers than fiscal policy .	cependant , la politique monétaire est une bien plus délicat , et indirects imprecise moyen de choisir les gagnants et perdants que la politique fiscale .
39	will the eu keep faith with turkey ?	l' union européenne de garder la foi avec la turquie ?
49	negotiations with turkey would be lengthy and accession may not occur before 2015 .	les négociations avec la turquie serait long et à l' adhésion ne peuvent pas se produire avant 2015 .

54	because of its weak economy , turkey would be eligible for substantial eu assistance .	en raison de sa faiblesse de l' économie , la turquie serait autorisée substantielle à l' aide de l' union européenne .

A.1.2 Français-anglais

id	source	traduction
1	apprivoiser les politiciens des deux côtés de l' atlantique	apprivoiser politicians on both sides of the atlantic
5	la réserve fédérale américaine est bien entendu extrêmement indépendante .	the us federal reserve is of course highly .
8	mais l' utilisation irréfléchie de l'un de ces deux mécanismes afin d' atteindre des objectifs à court terme risque de créer des coûts à long terme , justifiant certaines restrictions imposées à la latitude qu' ont les politiciens pour influencer les politiques macroéconomiques .	but the use irréfléchie of one of these two mechanisms to short-term purposes may create long-term costs , for restrictions imposed on the freedom that have politicians to inform macroeconomic policies .
11	la politique monétaire ne constitue toutefois pas un mode unique d' élaboration de politiques économiques .	monetary policy is not a single way of conducting economic policies .
22	c' est une tâche bien plus difficile !	it is a task more difficult !
25	la politique monétaire constitue toutefois un moyen bien plus maladroit , indirect et imprécis de choisir les gagnants et les perdants que la politique fiscale .	monetary policy is , however , a far more clumsy way , and indirect imprécis to choose winners and losers that fiscal policy .
28	il est désormais tout à fait clair que les gouvernements nationaux , et tout particulièrement ceux des grands pays , n' accepteront jamais des limites conséquentes sur leur discrétion fiscale .	it is now quite clear that national governments , particularly those of the big countries , will never significant limits on their fiscal discretion .
32	il est impossible d' empêcher l' utilisation politique de la politique fiscale .	it is impossible to prevent the use of fiscal policy .
36	les règles qui régissent le comportement des politiciens ne doivent pas être rédigées comme si des fonctionnaires désintéressés et bien intentionnés allaient les appliquer .	the rules governing politicians ' behavior should not be rédigées as if civil servants désintéressés and intentionnés would enforce them .
42	pendant toutes ces années , les gouvernements européens n' ont cessé de réaffirmer que la turquie pourrait devenir un etat membre à part entière , à condition de respecter des critères d' adhésion .	for all these years , european governments have been reaffirm that turkey could become a full member state , to membership criteria .

Annexe B

Resultats d'évaluation automatique du reclassement

TAB. B.1 – Résultats du reclassement avec les différents traits de couplage pour Moses (français - anglais)

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Moses Français - anglais					
Ligne de référence	7,0045	0,2679	6,7757	0,3420	0,3419
Oracle	8,5113	0,3854	8,6945	0,5602	0,4528
Pseudoréférence NIST					
générique	6,8171	0,2538	6,4384	-	-
lexicale	6,8928	0,2518	6,4806	-	-
étiqueté	6,8973	0,2522	6,4944	-	-
base, générique	6,9928	0,2652	6,8035	-	-
base, lexicale	6,9949	0,2642	6,8107	-	-
base, étiqueté	**7,0960**	**0,2700**	**6,8451**	-	-
Pseudoréférence BLEU					
générique	6,8541	0,2538	-	0,3402	-
lexicale	6,8855	0,2445	-	0,3443	-
étiqueté	6,9047	0,2550	-	0,3404	-
base, générique	7,0373	0,2669	-	0,3401	-
base, lexicale	7,0354	0,2653	-	0,3404	-
base, étiqueté	**7,0953**	**0,2712**	-	**0,3447**	-
Pseudoréférence wplBLEU					
générique	6,9029	0,2503	-	-	**0,3471**
lexicale	6,8864	0,2500	-	-	0,3416
étiqueté	6,8270	0,2446	-	-	0,3408
base, générique	7,0672	0,2639	-	-	0,3432
base, lexicale	7,0771	0,2632	-	-	0,3445
base, étiqueté	**7,1180**	0,2652	-	-	0,3467
Traduction de référence					
générique	6,9043	0,2371	-	-	-
lexicale	6,7893	0,2309	-	-	-
étiqueté	6,8708	0,2488	-	-	-

TAB. B.2 – Résultats du reclassement avec les différents traits de couplage pour Moses (anglais - français)

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Moses anglais-français					
Ligne de référence	6,8526	0,2619	6,4339	0,3196	0,3707
Oracle	8,3465	0,3621	8,4031	0,5161	0,5324
Pseudoréférence NIST					
générique	6,7922	0,2570	6,2648	-	-
lexicale	6,8567	0,2559	6,2788	-	-
étiqueté	6,7982	0,2516	6,2548	-	-
base, générique	6,9615	0,2539	6,4498	-	-
base, lexicale	6,9629	0,2542	6,4642	-	-
base, étiqueté	**6,9909**	0,2583	**6,4728**	-	-
Pseudoréférence BLEU					
générique	6,7935	0,2568	-	0,3203	-
lexicale	6,8547	0,2520	-	0,3183	-
étiqueté	6,8578	0,2527	-	0,3192	-
base, générique	6,9749	0,2556	-	0,3263	-
base, lexicale	6,9635	0,2547	-	0,3270	-
base, étiqueté	**6,9804**	0,2568	-	**0,3288**	-
Pseudoréférence wplBLEU					
générique	6,7874	0,2569	-	-	0,3675
lexicale	6,8631	0,2541	-	-	0,3688
étiqueté	6,7758	0,2501	-	-	0,3632
base, générique	6,9625	0,2538	-	-	0,3750
base, lexicale	6,9592	0,2529	-	-	0,3734
base, étiqueté	**6,9777**	0,2567	-	-	**0,3758**
Traduction de référence					
générique	6,5626	0,2350	-	-	-
lexicale	6,3876	0,2297	-	-	-
étiqueté	6,6275	0,2349	-	-	-

TAB. B.3 – Résultats du reclassement avec les différents traits de couplage pour Sinuhe (anglais - espagnol),

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Sinuhe anglais - espagnol					
Ligne de référence	7,6413	0,3465	6,9830	0,3915	0,3399
Oracle	8,2261	0,3795	7,7772	0,4528	0,3929
Pseudoréférence NIST					
générique	7,4213	0,3282	6,7060	-	-
lexicale	7,4286	0,3250	6,6879	-	-
étiqueté	7,5092	0,3316	6,7523	-	-
base, générique	**7,6903**	0,3452	**6,9953**	-	-
base, lexicale	7,6769	0,3448	6,9828	-	-
base, étiqueté	7,6768	0,3453	6,9820	-	-
Pseudoréférence BLEU					
générique	7,3818	0,3205	-	0,3902	-
lexicale	7,4213	0,3250	-	0,3907	-
étiqueté	7,4257	0,3284	-	0,3882	-
base, générique	7,5941	0,3319	-	0,3984	-
base, lexicale	7,6142	0,3343	-	0,3970	-
base, étiqueté	**7,6307**	0,3350	-	**0,3986**	-
Pseudoréférence wplBLEU					
générique	7,3799	0,3193	-	-	0,3417
lexicale	7,3306	0,3168	-	-	0,3436
étiqueté	7,4203	0,3240	-	-	0,3418
base, générique	7,5295	0,3199	-	-	0,3500
base, lexicale	7,5064	0,3196	-	-	0,3496
base, étiqueté	7,5200	0,3200	-	-	**0,3507**

TAB. B.4 – Intégration des traits à étiquette directe et générique (R - entraîné avec une vraie référence, PR - entraîné avec une pseudoréférence tenant compte des traits du modèle de base) avec d'autres traits

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Moses Français - anglais					
Ligne de référence	7,0045	0,2679	6,7757	0,3420	0,3419
Oracle	8,5113	0,3854	8,6945	0,5602	0,4528
Pseudoréférence NIST					
base, étiqueté	**7,0960**	0,2700	**6,8451**	-	-
base, étiqueté gén (R)	7,0208	0,2652	6,8170	-	-
base, étiqueté gén (PR, base)	7,0695	**0,2711**	6,8182	-	-
base, générique, lexicale, étiqueté	7,0519	0,2698	**6,8566**	-	-
base, générique, lexicale, étiqueté gén (R)	7,0438	0,2681	6,8479	-	-
base, générique, lexicale, étiqueté gén (PR, base)	**7,0790**	**0,2718**	6,8314	-	-
Pseudoréférence BLEU					
base, étiqueté	**7,0953**	**0,2712**	-	**0,3390**	-
base, étiqueté gén(R)	7,0451	0,2663	-	0,3411	-
base, étiqueté gén(PR, base)	7,0842	0,2702	-	**0,3437**	-
base, générique, lexicale, étiqueté	7,0540	0,2681	-	0,3396	-
base, générique, lexicale, étiqueté gén(R)	**7,0724**	0,2691	-	**0,3400**	-
base, générique, lexicale, étiqueté gén(PR, base)	7,0669	**0,2695**	-	0,3391	-
Pseudoréférence wplBLEU					
base, étiqueté	7,0957	0,2669	-	-	0,3478
base, étiqueté gén (R)	7,0915	0,2629	-	-	0,3435
base, étiqueté gén (PR,moses)	**7,1557**	0,2573	-	-	**0,3497**
base, générique, lexicale, étiqueté	6,9187	0,2560	-	-	0,3438
base, générique, lexicale, étiqueté gén(R)	7,1094	0,2651	-	-	0,3443
base, générique, lexicale, étiqueté gén(PR,moses)	**7,1723**	0,2591	-	-	**0,3483**
Traduction de référence					
générique, lexicale, étiqueté	6,7995	0,2424	-	-	-
générique, lexicale, étiqueté gén (R)	6,7991	0,2440	-	-	-

TAB. B.5 – Intégration des traits à étiquette directe et générique (R - entraîné avec une vraie référence, PR - entraîné avec une pseudo-référence tenant compte des traits du modèle de base) avec d'autres traits

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Moses anglais-français					
Ligne de référence	6,8526	0,2619	6,4339	0,3196	0,3707
Oracle	8,3465	0,3621	8,4031	0,5161	0,5324
Pseudoréférence NIST					
base, étiqueté	**6,9909**	0,2583	6,4728	-	-
base, étiqueté gén(R)	6,9305	0,2509	6,4298	-	-
base, étiqueté gén(PR,base)	6,9738	0,2559	**6,4739**	-	-
base, générique, lexicale, étiqueté	**7,0088**	0,2601	6,4901	-	-
base, générique, lexicale, étiqueté gén(R)	6,9696	0,2546	6,4601	-	-
base, générique, lexicale, étiqueté gén(PR,base)	6,9972	0,2579	**6,4999**	-	-
Pseudoréférence BLEU					
base, étiqueté	6,9804	0,2568	-	0,3288	-
base, étiqueté gén(R)	6,9737	0,2562	-	0,3265	-
base, étiqueté gén(PR,base)	**6,9962**	0,2592	-	**0,3302**	-
base, générique, lexicale, étiqutté	6,9842	0,2593	-	0,3282	-
base, générique, lexicale, étiqueté gén(R)	6,9899	0,2584	-	0,3275	-
base, générique, lexicale, étiqueté gén(PR,base)	**6,9970**	0,2588	-	**0,3303**	-
Pseudoréférence wplBLEU					
base, étiqueté	**6,9777**	0,2567	-	-	0,3758
base, étiqueté gén(R)	6,9502	0,2534	-	-	0,3760
base, étiqueté gén(PR,base)	6,9175	0,2524	-	-	**0,3768**
base, générique, lexicale, étiqueté	6,9532	0,2560	-	-	0,3755
base, générique, lexicale, étiqueté gén(R)	**6,9645**	0,2562	-	-	0,3775
base, générique, lexicale, étiqueté gén(PR,base)	6,9561	0,2554	-	-	**0,3783**
Traduction de référence					
générique, lexicale, étiqueté	6,5025	0,2362	-	-	-
générique, lexicale, étiqueté gén (RR)	*6,5326*	*0,2368*	-	-	-

TAB. B.6 – Résultats du reclassement avec des traits de couplage basés sur des alignements fournis par le système, Moses (français - anglais)

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Ligne de référence	7,0045	0,2679	6,7757	0,3420	0,3419
Oracle	8,5113	0,3854	8,6945	0,5602	0,4528
Moses mot à mot					
Pseudoréférence NIST					
base, générique	7,0199	0,2659	6,8216	-	-
base, lexicale	7,0252	0,2654	6,8286	-	-
base, étiqueté	7,1068	0,2698	6,8419	-	-
base, étiqueté gén(PR,base)	7,1114	0,2712	6,7683	-	-
base, générique, lexicale, étiqueté	7,0488	0,2701	6,8543	-	-
base, générique, lexicale, étiqueté gén (PR,base)	**7,1270**	**0,2715**	6,7754	-	-
Pseudoréférence BLEU					
base, générique	7,0558	0,2676	-	0,3390	-
base, lexicale	7,0397	0,2654	-	0,3368	-
base, étiqueté	7,0375	0,2665	-	0,3390	-
base, étiqueté gén(PR,base)	7,1015	0,2717	-	0,3402	-
base, générique, lexicale, étiqueté	7,0606	0,2679	-	0,3409	-
base, générique, lexicale, étiqueté gén (PR,base)	**7,1139**	**0,2732**	-	**0,3413**	-
Pseudoréférence wplBLEU					
base, générique	7,0958	0,2651	-	-	0,3438
base, lexicale	7,1008	0,2641	-	-	0,3444
base, étiqueté	**7,1840**	0,2661	-	-	**0,3500**
base, étiqueté gén(PR, base)	7,1359	0,2676	-	-	0,3458
base, générique, lexicale, étiqueté	7,0814	0,2656	-	-	0,3445
base, générique, lexicale, étiquette gén(PR, base)	7,1384	**0,2697**	-	-	0,3439
Moses fragment à fragment					
Pseudoréférence NIST					
base, générique	7,0078	0,2654	6,8150	-	-
base, lexicale	6,9718	0,2619	6,7809	-	-
base, étiqueté	7,0819	**0,2715**	**6,8671**	-	-
base, étiqueté gén(PR, base)	**7,1179**	**0,2715**	6,8228	-	-
base, générique, lexicale, étiqueté	7,0545	0,2695	6,8526	-	-
base, générique, lexicale, étiquette gén (PR,base)	7,1027	0,2697	6,8226	-	-
Pseudoréférence BLEU					
base, générique	7,0602	0,2669	-	0,3410	-
base, lexicale	7,0289	0,2643	-	0,3403	-
base, étiqueté	7,0910	**0,2698**	-	**0,3447**	-
base, étiqueté gén (PR, base)	**7,0980**	**0,2698**	-	0,3412	-
base, générique, lexicale, étiqueté	7,0730	0,2686	-	0,3411	-
base, générique, lexicale, étiqueté gén (PR, base)	7,0826	0,2681	-	0,3392	-
Pseudoréférence wplBLEU					
base, générique	7,0875	0,2639	-	-	0,3432
base, lexicale	7,0712	0,2619	-	-	0,3430
base, étiqueté	7,1404	0,2655	-	-	0,3441
base, étiqueté gén (PR,base)	**7,1756**	0,2623	-	-	**0,3468**
base, générique, lexicale, étiqueté	7,0450	0,2606	-	-	0,3442
base, générique, lexicale, étiquette gén (PR,base)	7,1247	0,2598	-	-	0,3455

TAB. B.7 – Résultats du reclassement avec des traits de couplage basés sur des alignements fournis par le système, Moses (anglais - français)

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Ligne de référence	6,8526	0,2619	6,4339	0,3196	0,3707
Oracle	8,3465	0,3621	8,4031	0,5161	0,5324
Moses mot à mot					
Pseudoréférence NIST					
base, générique	6,9673	0,2558	6,4627	-	-
base, lexicale	6,9654	0,2542	6,4659	-	-
base, étiqueté	7,0026	0,2591	6,4913	-	-
base, étiqueté gén (PR,base)	7,0093	0,2608	6,4958	-	-
base, générique, lexicale, étiqueté	**7,0158**	0,2604	**6,5185**	-	-
base, générique, lexicale, étiqueté gén(PR,base)	7,0127	0,2614	6,5016	-	-
Pseudoréférence BLEU					
base, générique	6,9787	0,2563	-	0,3268	-
base, lexicale	6,9554	0,2539	-	0,3262	-
base, étiqueté	6,9920	0,2570	-	0,3274	-
base, étiqueté gén (PR,base)	6,9827	0,2571	-	0,3302	-
base, générique, lexicale, étiqueté	6,9997	0,2588	-	0,3259	-
base, générique, lexicale, étiqueté gén (PR,base)	6,9861	0,2589	-	0,3277	-
Pseudoréférence wplBLEU					
base, générique	6,9193	0,2514	-	-	0,3750
base, lexicale	6,9031	0,2486	-	-	0,3737
base, étiqueté	6,9910	0,2569	-	-	0,3776
base, gén, étiqueté (PR,base)	6,9539	0,2550	-	-	0,3788
base, gén, lex, étiqueté	**6,9915**	0,2579	-	-	0,3773
base, générique, lexicale, étiqueté gén (PR,base)	6,9870	0,2577	-	-	**0,3800**
Moses fragment à fragment					
Pseudoréférence NIST					
base, générique	6,9666	0,2545	6,4529	-	-
base, lexicale	6,9430	0,2520	6,4394	-	-
base, étiqueté	**6,9748**	0,2569	**6,4581**	-	-
base, étiqueté gén (PR, base)	6,9505	0,2552	6,4400	-	-
base, générique, lexicale, étiqueté	6,9680	0,2571	6,4481	-	-
base, générique, lexicale, étiquete gén (PR,base)	6,9451	0,2542	6,4390	-	-
Pseudoréférence BLEU					
base, générique	6,9929	0,2576	-	0,3283	-
base, lexicale	6,9763	0,2557	-	0,3267	-
base, étiqueté	6,9769	0,2569	-	0,3297	-
base, étiqueté gén (PR,base)					
base, générique, lexicale, étiqueté	7,0140	0,2607	-	0,3282	-
base, générique, lexicale, étiqueté gén (PR,base)	6,9983	0,2588	-	0,3277	-
Pseudoréférence wplBLEU					
base, générique	6,8940	0,2490	-	-	0,3736
base, lexicale	6,8711	0,2463	-	-	0,3732
base, étiqueté	**6,9503**	0,2533	-	-	0,3751
base, générique, étiqueté (PR,base)	6,9213	0,2515	-	-	0,3755
base, générique, lexicale, étiqutté	6,9359	0,2524	-	-	**0,3766**
base, générique, lexicale, étiquette gén (PR,base)	6,9036	0,2486	-	-	0,3758

TAB. B.8 – Résultats du reclassement avec des traits de cohésion, Moses (français-anglais)

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Français - anglais					
Ligne de référence	7,0045	0,2679	6,7757	0,3420	0,3419
Oracle	8,5113	0,3854	8,6945	0,5602	0,4528
Pseudo-référence NIST					
base, cohésion	7,0482	0,2632	6,7401	-	-
base, incohésion	7,0702	0,2668	6,7465	-	-
base, cohésion , incohésion	7,0374	0,2667	6,8097	-	-
base, cohésion gén(PR,base)	7,0379	0,2632	6,7714	-	-
base, incohésion gén(PR,base)	7,0564	0,2651	6,7903	-	-
base, cohésion gén(PR,base), incohésion gén(PR,base)	**7,1061**	0,2652	6,7719	-	-
base, cohésion gén(RR)	7,0025	0,2647	6,8035	-	-
base, incohésion gén(RR)	7,0230	0,2660	6,8023	-	-
base, cohésion gén(RR), incohésion gén(RR)	7,0374	0,2670	**6,8174**	-	-
Pseudo-référence BLEU					
base, cohésion	7,0665	0,2656	-	0,3395	-
base, incohésion	7,0463	0,2661	-	0,3383	-
base, cohésion , incohésion	7,0230	0,2631	-	0,3385	-
base, cohésion gén(PR,base)	7,0872	0,2670	-	**0,3435**	-
base, incohésion gén(PR,base)	**7,1203**	0,2675	-	0,3410	-
base, cohésion gén(PR,base), incohésion gén(PR,base)	7,0983	0,2670	-	0,3416	-
base, cohésion gén(RR)	7,0508	0,2663	-	0,3396	-
base, incohésion gén(RR)	7,0350	0,2654	-	0,3401	-
base, cohésion gén(RR), incohésion gén(RR)	7,0416	0,2655	-	0,3396	-
Pseudo-référence wpBLEU					
base, cohésion	7,0962	0,2583	-	-	0,3472
base, incohésion	7,0938	0,2633	-	-	0,3458
base, cohésion , incohésion	7,0448	0,2594	-	-	0,3452
base, cohésion gén(PR,base)	7,1027	0,2646	-	-	0,3453
base, incohésion gén(PR,base)	7,1215	0,2479	-	-	0,3471
base, cohésion gén(PR,base), incohésion gén(PR,base)	**7,1385**	0,2550	-	-	**0,3488**
base, cohésion gén(RR)	7,0686	0,2626	-	-	0,3437
base, incohésion gén(RR)	7,1043	0,2646	-	-	0,3438
base, cohésion gén(RR), incohésion gén(RR)	7,0960	0,2642	-	-	**0,3447**
Traduction de référence					
cohésion	6,9475	0,2482	-	-	-
incohésion	6,8808	0,2522	-	-	-
cohésion , incohésion	6,8749	0,2529	-	-	-
cohésion gén(RR), incohésion gén(RR)	6,8771	0,2519	-	-	-

TAB. B.9 – Résultats du reclassement avec des traits de cohésion, Moses (anglais-français)

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Anglais - français					
Ligne de référence	6,8526	0,2619	6,4339	0,3196	0,3707
Oracle	8,3465	0,3621	8,4031	0,5161	0,5324
Pseudo-référence NIST					
base, cohésion	6,9728	0,2533	6,4632	-	-
base, incohésion	**6,9887**	0,2556	**6,4926**	-	-
base, cohésion , incohésion	6,9746	0,2541	6,4678	-	-
base, cohésion gén(PR,base)	6,9484	0,2519	6,4565	-	-
base, incohésion gén(PR,base)	6,9246	0,2512	6,4190	-	-
base, cohésion gén(PR,base) , incohésion gén(PR,base)	6,8852	0,2493	6,378	-	-
base, cohésion gén(RR)	6,9474	0,2521	6,4302	-	-
base, incohésion gén(RR)	6,9555	0,2535	6,4687	-	-
base, cohésion gén(RR) , incohésion gén(RR)	6,9560	0,2538	6,4593	-	-
Pseudo-référence BLEU					
base, cohésion	6,9916	0,2564	-	0,3279	-
base, incohésion	6,9685	0,2544	-	0,3255	-
base, cohésion , incohésion	6,9797	0,2549	-	0,3270	-
base, cohésion gén(PR,base)	6,9634	0,2554	-	0,3289	-
base, incohésion gén(PR,base)	6,9668	0,2557	-	**0,3297**	-
base, cohésion gén(PR,base) , incohésion gén(PR,base)	6,9367	0,2540	-	0,3275	-
base, cohésion gén(RR)	6,9845	0,2572	-	0,3260	-
base, incohésion gén(RR)	6,9845	0,2571	-	0,3277	-
base, cohésion gén(RR) , incohésion gén(RR)	**7,0057**	0,2591	-	0,3284	-
Pseudo-référence wpBLEU					
base, cohésion	6,9618	0,2520	-	-	0,3766
base, incohésion	6,9813	0,2544	-	-	0,3750
base, cohésion , incohésion	6,9606	0,2533	-	-	0,3762
base, cohésion gén(PR,base)	6,9537	0,2535	-	-	0,3770
base, incohésion gén(PR,base)	6,8752	0,2488	-	-	0,3751
base, cohésion gén(PR,base) , incohésion gén(PR,base)	6,8914	0,2494	-	-	0,3764
base, cohésion gén(RR)	6,9513	0,2536	-	-	0,3774
base, incohésion gén(RR)	6,9584	0,2543	-	-	**0,3776**
base, cohésion gén(RR) , incohésion gén(RR)	6,9537	0,2535	-	-	0,3770
Traduction de référence					
cohésion	6,6902	0,2518	-	-	-
incohésion	6,7077	0,2469	-	-	-
cohésion , incohésion	6,7002	0,2484	-	-	-
cohésion gén(RR), incohésion gén(RR)	6,6970	0,2487	-	-	-

TAB. B.10 – Résultats du reclassement avec le modèle de langage discriminatif.

	NIST	**BLEU**	Ind NIST	Ind BLEU	wpBLEU
Sinune : Anglais -espagnol					
Ligne de référence	7,6413	0,3465	6,9830	0,3915	0,3399
Oracle	8,2261	0,3795	7,7772	0,4528	0,3929
Pseudo-référence NIST					
sinuhe, MLF (SVM, CE)	**7,6906**	0,3446	**6,9940**	-	-
sinuhe, MLF (SVM, NIST)	7,6800	0,3446	6,9887	-	-
sinuhe, coh, incoh	7,6564	0,3437	6,9772	-	-
sinuhe, coh , incoh , MLF (CE)	7,6683	0,3438	6,9850	-	-
sinuhe, coh , incoh , MLF (NIST)	7,6605	0,3432	6,9802	-	-
Pseudo-référence BLEU					
sinuhe, MLF (CE)	7,5926	0,3309	-	**0,3986**	-
sinuhe, MLF (NIST)	7,6025	0,3317	-	0,3984	-
sinuhe, coh, incoh	7,5844	0,3331	-	0,3955	-
sinuhe, coh , incoh , MLF (CE)	7,5717	0,3313	-	0,3970	-
sinuhe, coh , incoh , MLF (NIST)	7,5601	0,3331	-	0,3975	-
Pseudo-référence wpBLEU					
sinuhe, MLF (SVM, CE)	7,4977	0,3179	-	-	**0,3516**
sinuhe, MLF (SVM, NIST)	7,5807	0,3241	-	-	0,3501
sinuhe, coh, incoh	7,5131	0,3206	-	-	0,3496
sinuhe, coh, incoh , MLF (CE)	7,5086	0,3207	-	-	0,3500
sinuhe, coh, incoh , MLF (NIST)	7,5081	0,3201	-	-	0,3489

TAB. B.11 – Moses (français - anglais), combinaison des traits bilingues et monolingues. Alignements mot à mot fournis par le système.

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Ligne de référence	7,0045	0,2679	6,7757	0,3420	0,3419
Oracle	8,5113	0,3854	8,6945	0,5602	0,4528
Moses mot à mot					
Pseudoréférence NIST					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR, base)	7,1360	0,2717	6,8132	-	-
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	7,1336	0,2722	6,7826	-	-
base, générique, lexical, étiquette gén(PR,base), coh gén(PR, base)	7,1434	**0,2732**	6,8198	-	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,1194	0,2719	6,7767	-	-
Pseudoréférence BLEU					
base, étiquette gén(PR,base), cohésion générique, incohésion générique (PR, base)	7,1161	0,2694	-	0,3437	-
base, étiquette gén(PR,base), cohésion générique, incohésion générique (RR)	7,1136	0,2712	-	0,3445	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	7,1547	**0,2732**	-	0,3444	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,1276	**0,2735**	-	0,3448	-
Pseudoréférence wplBLEU					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR, base)	7,1550	0,2569	-	-	0,3517
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	7,1591	0,2685	-	-	0,3462
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	7,1824	0,2623	-	-	**0,3516**
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,1532	0,2706	-	-	0,3487

TAB. B.12 – Moses (français - anglais), combinaison des traits bilingues et monolingues. Alignements fragment à fragment fournis par le système.

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Ligne de référence	7,0045	0,2679	6,7757	0,3420	0,3419
Oracle	8,5113	0,3854	8,6945	0,5602	0,4528
Moses fragment à fragment					
Pseudoréférence NIST					
base, étiquette gén(PR, base), cohésion gén(PR, base), incohésion gén(PR, base)	7,1327	0,2705	6,8195	-	-
base, étiquette gén(PR, base), cohésion gén(RR), incohésion gén(RR)	7,1318	0,2720	**6,8213**	-	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	7,1362	0,2691	6,7982	-	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,1259	0,2717	**6,8254**	-	-
Pseudoréférence BLEU					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR, base)	7,0955	0,2682	-	0,3410	-
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	7,1068	0,2710	-	0,3429	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	7,1154	0,2684	-	0,3425	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,0765	0,2689	-	0,3421	-
Pseudoréférence wplBLEU					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR, base)	7,1630	0,2585	-	-	0,3488
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	**7,1927**	0,2619	-	-	0,3470
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	7,1511	0,2583	-	-	**0,3502**
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,1515	0,2582	-	-	0,3469

TAB. B.13 – Moses (français - anglais), combinaison des traits bilingues et monolingues. Alignements mot à mot établis par GIZA++.

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Ligne de référence	7,0045	0,2679	6,7757	0,3420	0,3419
Oracle	8,5113	0,3854	8,6945	0,5602	0,4528
Moses GIZA++					
Pseudoréférence NIST					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR, base)	7,1130	0,2714	6,8383	-	-
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	7,0886	0,2721	**6,8425**	-	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	7,1315	0,2723	6,8363	-	-
base, générique, lexical, étiqueté gén(PR, base), coh gén(RR)	7,1192	**0,2734**	**6,8674**	-	-
Pseudoréférence BLEU					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR, base)	7,0955	0,2682	-	0,3410	-
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	7,1068	0,2710	-	0,3429	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	7,1527	**0,2737**	-	0,3392	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,0809	0,2714	-	0,3454	-
Pseudoréférence wplBLEU					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR, base)	7,1315	0,2552	-	-	**0,3501**
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	**7,1931**	0,2634	-	-	**0,3501**
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	7,1565	0,2560	-	-	**0,3507**
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,1766	0,2610	-	-	0,3483
Traduction de référence					
étiquette gén(RR), coh gén(RR)	6,9399	0,2484	-	-	-
étiquette gén(PR,base) incoh gén(RR)	6,9217	0,2522	-	-	
étiquette gén(RR), coh gén(RR), incoh gén(RR)	6,9154	0,2517	-	-	-
générique, lexical, étiqueté gén(RR), coh gén(RR)	6,8376	0,2447	-	-	-
générique, lexical, étiqueté gén(RR), incoh gén(RR)	6,8610	0,2492	-	-	-
générique, lexical, étiqueté gén(RR), coh gén(RR), incoh gén(RR)	6,8647	0,2491	-	-	-

TAB. B.14 – Moses (anglais - français), combinaison des traits bilingues et monolingues. Alignements mot à mot fournis par le système.

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Ligne de référence	6,8526	0,2619	6,4339	0,3196	0,3707
Oracle	8,3465	0,3621	8,4031	0,5161	0,5324
Moses mot à mot					
Pseudoréférence NIST					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR, base)	6,9665	0,2572	6,4523	-	-
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	7,0045	0,2601	6,4993	-	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	6,9626	0,2575	6,4508	-	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,0197	0,2611	6,5075	-	-
Pseudoréférence BLEU					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR,base)	6,9595	0,2567	-	0,3313	-
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	7,0036	0,2593	-	0,3297	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	7,0124	0,2605	-	0,3321	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,0062	0,2595	-	0,3291	-
Pseudoréférence wplBLEU					
base, étiquette gén(PR,base), cohésion gén(PR, base), incohésion gén(PR, base)	6,9314	0,2546	-	-	0,3792
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	6,9689	0,2565	-	-	0,3802
base, générique, lexical, étiqueté gén(PR, base), coh gén(PR, base)	6,9576	0,2560	-	-	0,3798
base, générique, lexical, étiqueté gén(RR), coh gén(RR)	6,9873	0,2571	-	-	0,3787

TAB. B.15 – Moses (anglais - français), combinaison des traits bilingues et monolingues. Alignements fragment à fragment fournis par le système.

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Ligne de référence	6,8526	0,2619	6,4339	0,3196	0,3707
Oracle	8,3465	0,3621	8,4031	0,5161	0,5324
Moses fragment à fragment					
Pseudoréférence NIST					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR, base)	6,9466	0,2541	6,434	-	-
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	6,9676	0,2558	6,4807	-	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	6,9431	0,2543	6,4335	-	-
base, générique, lexical, étiqueté gén(RR), coh gén(RR)	6,9530	0,2549	6,4415	-	-
Pseudoréférence BLEU					
base, étiquette gén(PR,base), coh gén(PR, base), incohésion gén(PR, base)	6,9868	0,2587	-	0,3301	-
base, étiquette gén(PR,base), cohésion générique, incohésion générique (RR)	6,9950	0,2584	-	0,3285	-
base, générique, lexical, étiquette gén(PR,base), coh gén(PR, base)	6,9726	0,2569	-	0,3283	-
base, générique, lexical, étiquette gén(PR,base), coh gén(RR)	7,0062	0,2595	-	0,3291	-
Pseudoréférence wplBLEU					
base, étiquette gén(PR,base), cohésion gén(PR, base), incohésion gén(PR, base)	6,9029	0,2507	-	-	0,3772
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	6,9267	0,2513	-	-	0,3764
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	6,8862	0,2485	-	-	0,3754
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	6,8505	0,2456	-	-	0,3724

TAB. B.16 – Moses (anglais - français), combinaison des traits bilingues et monolingues. Alignements mot à mot établis par GIZA++.

	NIST	**BLEU**	Ind NIST	Ind BLEU	wplBLEU
Ligne de référence	6,8526	0,2619	6,4339	0,3196	0,3707
Oracle	8,3465	0,3621	8,4031	0,5161	0,5324
Moses GIZA++					
Pseudoréférence NIST					
base, étiquette gén(PR,base), cohésion gén(PR, base), incohésion gén(PR, base)	6,9795	0,2580	6,4666	-	-
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	6,9961	0,2583	6,4903	-	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	6,9804	0,2573	6,4544	-	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	6,9893	0,2575	6,4730	-	-
Pseudoréférence BLEU					
base, étiquette gén(PR,base), cohésion gén(PR,base), incohésion gén(PR, base)	6,9801	0,2580	-	0,3301	-
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	7,0119	0,2613	-	0,3307	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(PR, base)	6,9781	0,2582	-	0,3304	-
base, générique, lexical, étiqueté gén(PR,base), coh gén(RR)	7,0231	0,2612	-	0,3303	-
Pseudoréférence wplBLEU					
base, étiquette gén(PR,base), cohésion gén(PR, base), incohésion gén(PR, base)	6,8909	0,2497	-	-	0,3755
base, étiquette gén(PR,base), cohésion gén(RR), incohésion gén(RR)	6,9522	0,2546	-	-	0,3768
base, générique, lexical, étiquette gén(PR,base), coh gén(PR, base)	6,9432	0,2543	-	-	0,3790
base, générique, lexical, étiquette gén(PR,base), coh gén(RR)	6,9498	0,2555	-	-	0,3784
Traduction de référence					
étiquette gén(RR), coh gén(RR)	6,6081	0,2434	-	-	-
étiquette gén(RR), incoh gén(RR)	6,6416	0,2428	-	-	
étiquette gén(RR), coh gén(RR),incoh gén(RR)	6,6512	0,2439	-	-	-
générique, lexical, étiquette gén(RR), coh gén(RR)	6,5199	0,2369	-	-	-
générique, lexical, étiquette gén(RR), incoh gén(RR)	6,5621	0,2367	-	-	-
générique, lexical, étiquette gén(RR), coh gén(RR), incoh gén(RR)	6,5554	0,2373	-	-	-

Annexe C

Jugements humains

Le modèle de reclassement est défini par :
- le type de la référence : vraie référence (RR), ou pseudo-référence (BLEU, NIST, wplBLEU) ;
- l'ensemble des traits utilisé parmi les ensemble suivants : traits du modèle de référence (base), traits de couplage générique (gen), traits de couplage lexicalisé (lex), traits de couplage étiqueté (lab), trait de couplage générique étiqueté (lab(RR) : la première phrase d'entraînement est faite avec une vraie référence, lab(PR) : la première phase d'entraînement est faite avec une pseudo-référence et avec des traits de base), traits de cohésion (coh), trait générique de cohésion (coh(RR), coh(PR)), traits de non-cohésion (incoh), trait générique de non-cohésion (incoh(RR), incoh(PR)) ;
- le type d'alignement utilisé pour établir les traits de couplage : à l'aide de GIZA++ (giza), alignements au niveau des mots fournis par le système (w2w), alignement au niveau des fragments fournis par le système (c2c).

TAB. C.1 – Correspondance entre les codes des modèles et les ensembles de fonctions de traits de chaque modèle.

coh0	incoh(RR), coh(RR)
coh1	base, incoh(PR), coh(PR)
coh2	base, incoh(PR)
coh3	base, coh(PR)
coh4	base, incoh(RR)
coh5	base, coh(RR)
cpl0	base, lab(PR)(c2c)
cpl1	base, lab(PR)(w2w)
cpl3	base, gen, lex, lab(PR)(giza)
cpl4	gen, lex, lab(RR)(giza)
cpl5	base, lab(PR)(giza)
cpl6	base, lab(RR)(giza)
cpl7	base, gen, lex, lab(PR)(w2w)
cpl8	base, gen, lex, lab(PR)(c2c)

C.1 Anglais - français

Rappelons qu'ici les juges donnent un classement de l'adéquation ou de fluidité (de 1 à N, avec ex-equo possible).

C.1.1 Jugements d'adéquation

modèle	**traduction**	**juge1**	**juge2**
	ID :121		

TAB. C.2 – Continued

modèle	traduction	juge1	juge2
	SRC : Few countries score strongly here . REF :Rares sont les pays qui peuvent marquer des points dans ces domaines .		
fB-cpl0-coh0 ;fB-cpl1-coh0 ;fB-coh1 ;fB-coh2 ;fB-coh3 ;fB-coh0 ;fB-coh4 ;fB-coh5 ;fB-cpl3 ;fB-cpl4 ;fB-cpl5 ;fB-cpl0 ;fB-cpl7 ;fB-cpl1 ;N-cpl7 ; baseline, ;	quelques pays fortement note ici .	3	1
R-coh0 ;	quelques pays note vivement ce sujet .	7	1
oracleBLEU, ;	peu de résultats dans des pays fortement .	9	2
R-cpl6-coh0 ;R-cpl6-coh4 ;	quelques pays note a fortement ce sujet .	5	1
B-cpl0-coh0 ;B-cpl1-coh0 ;B-coh1 ;B-coh2 ;B-coh3 ;B-coh0 ;B-coh4 ;B-coh5 ;B-cpl3 ;B-cpl4 ;B-cpl5 ;B-cpl6 ;B-cpl8 ;B-cpl0 ;B-cpl7 ;B-cpl1 ;fB-cpl6 ;N-cpl0-coh0 ;N-cpl1-coh0 ;N-coh1 ;N-coh2 ;N-coh3 ;N-coh0 ;N-coh4 ;N-coh5 ;N-cpl3 ;N-cpl4 ;N-cpl5 ;N-cpl6 ;N-cpl8 ;N-cpl0 ;N-cpl1 ;	quelques pays note fortement ici .	3	1
R-cpl4 ;	certains pays note fermement à ce sujet .	7	1
R-cpl6-coh5 ;	quelques pays point a fortement à ce sujet .	10	1
oraclewpBLEU, ;	dans peu de pays fortement note .	2	2
fB-cpl8 ;	quelques pays score fortement ici .	6	1
oracleNIST, ;	rares sont les pays fortement résultats dans ce domaine .	1	1
	ID :127 SRC : As the social theorist Robert Putnam has explained , " social capital " –the networks , norms , and social trust that facilitate cooperation and coordination for mutual benefit–is as much a determinant as it is a result of economic growth . REF :Comme le démontre Robert Putnam , le théoricien social , « le capital social » (les réseaux , les normes et la confiance sociale qui facilitent la coopération et la coordination au bénéfice de tous) est autant un paramètre qu' un résultat de la croissance économique .		
oracleBLEU, ;	comme l' a expliqué le théoricien robert putnam social » , « le capital social des réseaux –the normes sociales et la confiance que faciliter la coopération et de coordination mutuelle benefit–is pour autant un élément comme c' est une conséquence de la croissance économique .	4	3
baseline, ;	comme l' a expliqué le théoricien sociale , robert putnam " social " –the des réseaux , des normes sociales et la confiance que faciliter la coopération et de coordination mutuelle benefit–is pour autant un élément comme c' est une conséquence de la croissance économique .	13	3

TAB. C.2 – Continued

modèle	**traduction**	**juge1**	**juge2**
N-cpl8 ;N-cpl0 ;N-cpl1 ;	comme le théoricien sociale de robert putnam a expliqué , " social " réseaux –the normes sociales et la confiance que faciliter la coopération et de coordination mutuelle benefit–is pour autant un élément comme c' est une conséquence de la croissance économique .	13	2
R-cpl4 ;	comme l' a expliqué le théoricien robert putnam social » , le capital social " réseaux –the normes sociales et la confiance que faciliter la coopération et de coordination mutuelle benefit–is pour autant un élément que c' est une conséquence de la croissance économique .	1	3
R-cpl6-coh0 ;	le théoricien sociale , comme l' a expliqué , robert putnam " social " –the des réseaux , des normes sociales et la confiance que faciliter la coopération mutuelle et la coordination benefit–is pour autant un élément comme c' est une conséquence de la croissance économique .	13	4
oraclewpBLEU, ;	comme l' a expliqué le théoricien robert putnam social » , le capital social " –the des réseaux , des normes sociales et la confiance que faciliter la coopération et de coordination mutuelle benefit–is pour autant un élément que c' est une conséquence de la croissance économique .	1	3
B-cpl0-coh0 ;B-cpl1-coh0 ;B-coh1 ;B-coh2 ;B-coh3 ;B-coh0 ;B-coh4 ;B-coh5 ;B-cpl6 ;B-cpl8 ;B-cpl0 ;B-cpl1 ;fB-coh0 ;fB-coh4 ;fB-coh5 ;fB-cpl5 ;fB-cpl6 ;N-cpl0-coh0 ;N-coh1 ;N-coh2 ;N-coh3 ;N-coh0 ;N-coh4 ;N-coh5 ;N-cpl3 ;N-cpl5 ;N-cpl6 ;	comme le théoricien sociale robert putnam a expliqué , " social " réseaux –the normes sociales et la confiance que faciliter la coopération et de coordination mutuelle benefit–is pour autant un élément comme c' est une conséquence de la croissance économique .	6	1
R-cpl6-coh4 ;	comme l' a expliqué le théoricien robert putnam social » , le capital social " –the des réseaux , des normes sociales et la confiance que faciliter la coopération mutuelle et de coordination benefit–is pour autant un élément comme c' est une conséquence de la croissance économique .	1	3
B-cpl3 ;B-cpl5 ;	comme le théoricien sociale robert putnam a expliqué , " social " réseaux –the normes sociales et la confiance que faciliter la coopération et de coordination mutuelle benefit–is pour autant un élément que c' est une conséquence de la croissance économique .	6	1
B-cpl4 ;B-cpl7 ;fB-coh3 ;fB-cpl3 ;fB-cpl4 ;fB-cpl7 ;fB-cpl1 ;N-cpl4 ;N-cpl7 ;	comme le théoricien sociale robert putnam a expliqué , " social " réseaux –the normes sociales et la confiance que faciliter la coopération mutuelle et la coordination benefit–is pour autant un élément comme c' est une conséquence de la croissance économique .	6	1

TAB. C.2 – Continued

modèle	**traduction**	**juge1**	**juge2**
fB-cpl0-coh0 ;fB-cpl8 ;fB-cpl0 ;	comme le théoricien sociale robert putnam a expliqué , " social " réseaux –the normes sociales et la confiance que faciliter la coopération mutuelle et de coordination pour benefit–is autant un élément comme c' est une conséquence de la croissance économique .	6	1
N-cpl1-coh0 ;	comme le théoricien sociale , robert putnam a expliqué , " social " réseaux –the normes sociales et la confiance que faciliter la coopération et de coordination mutuelle benefit–is pour autant un élément comme c' est une conséquence de la croissance économique .	6	1
fB-coh1 ;fB-coh2 ;	comme le théoricien sociale robert putnam a expliqué , " social " réseaux –the normes sociales et la confiance que faciliter la coopération benefit–is mutuelle et de coordination pour autant un élément comme c' est une conséquence de la croissance économique .	6	1
R-cpl6-coh5 ;	comme l' a expliqué le théoricien sociale de robert putnam capital social » , « réseaux –the normes sociales et la confiance que faciliter la coopération et de coordination mutuelle benefit–is pour autant un élément comme c' est une conséquence de la croissance économique .	17	2
fB-cpl1-coh0 ;	comme le théoricien sociale de robert putnam a expliqué , " social " réseaux –the normes sociales et la confiance que faciliter la coopération mutuelle et de coordination pour benefit–is autant un élément comme c' est une conséquence de la croissance économique .	6	2
oracleNIST, ;	comme l' a expliqué le théoricien robert putnam social » , « le capital social des réseaux – the normes sociales et la confiance que faciliter la coopération et de coordination pour autant un élément benefit–is mutuelle , comme c' est une conséquence de la croissance économique .	4	3
R-coh0 ;	comme l' a expliqué robert putnam le théoricien sociale , " social " –the des réseaux , des normes sociales et la confiance que faciliter la coopération et de coordination mutuelle benefit–is pour autant un élément comme c' est une conséquence de la croissance économique .	13	3
	ID :63 SRC : It is also feared that Islamists may one day turn Turkey into a fundamentalist state . REF :On craint également que les islamistes fassent un jour de la Turquie un Etat fondamentaliste .		
N-cpl1-coh0 ;	il est également craint que les islamistes pourraient un jour se tourner la turquie dans un état fondamentaliste .	2	2

TAB. C.2 – Continued

modèle	**traduction**	**juge1**	**juge2**
B-coh1 ;B-coh2 ;B-coh3 ;B-coh0 ;B-coh4 ;B-coh5 ;B-cpl4 ;B-cpl5 ;B-cpl6 ;B-cpl8 ;B-cpl0 ;B-cpl1 ;fB-coh3 ;fB-coh0 ;fB-coh4 ;fB-cpl4 ;fB-cpl6 ;fB-cpl8 ;N-cpl0-coh0 ;N-coh2 ;N-coh3 ;N-coh0 ;N-coh4 ;N-coh5 ;N-cpl3 ;N-cpl4 ;N-cpl5 ;N-cpl6 ;N-cpl8 ;N-cpl0 ;N-cpl7 ;N-cpl1 ;	il est également craint que les islamistes pourraient un jour son tour la turquie dans un état fondamentaliste .	6	3
R-coh0 ;R-cpl6-coh0 ;R-cpl6-coh4 ;R-cpl6-coh5 ;	il est également craint qu' ils pourraient un jour se tourner turquie dans un état fondamentaliste .	4	4
oracleNIST, ;	il est également craint que les islamistes pourraient un jour de transformer la turquie un état fondamentaliste .	1	1
oracleBLEU, ;	il est également craint que les pourraient un jour son tour la turquie dans un état fondamentaliste .	8	6
R-cpl4 ;	il est également craint que les islamistes peut-être un jour se tourner turquie dans un état fondamentaliste .	3	3
B-cpl0-coh0 ;B-cpl1-coh0 ;B-cpl3 ;B-cpl7 ;fB-cpl1-coh0 ;fB-coh1 ;fB-coh2 ;fB-coh5 ;fB-cpl3 ;fB-cpl5 ;fB-cpl7 ;fB-cpl1 ;N-coh1 ; baseline, ;	il est également craint qu' ils pourraient un jour son tour la turquie dans un état fondamentaliste .	7	5
oraclewpBLEU, ;	il est également craint qu' ils pourraient un jour se tourner vers un état fondamentaliste la turquie .	5	5
fB-cpl0-coh0 ;fB-cpl0 ;	il est également craint que les islamistes mai un jour son tour la turquie dans un état fondamentaliste .	9	6
	ID :88 SRC : The current war in Iraq is perhaps the strongest proof of this . REF :La guerre qui se déroule actuellement en Irak en est sans doute la meilleure preuve .		
oraclewpBLEU, ;	la guerre actuelle en irak est peut-être le fait preuve la plus de cela .	10	3
fB-coh1 ;	la guerre actuelle en irak est peut-être la preuve de plus .	9	3
fB-cpl6 ;N-coh4 ;N-cpl4 ;N-cpl6 ;	la guerre actuelle en irak est peut-être le plus sérieux preuve de cela .	2	1
N-cpl0-coh0 ;N-cpl8 ;N-cpl0 ;	la guerre actuelle en irak est peut-être la plus puissantes preuve de cela .	4	1
R-cpl4 ;R-cpl6-coh0 ;R-cpl6-coh4 ;R-cpl6-coh5 ;	le contexte actuel de la guerre en irak est peut-être la plus forte fait preuve de ce point .	11	3
fB-cpl8 ;	la guerre actuelle en irak est peut-être le plus sérieux la preuve de cela .	7	2
oracleNIST, ;	la guerre actuelle en irak est sans doute la plus grande preuve de cela .	1	3
B-cpl0 ;	la guerre actuelle en irak est peut-être le plus puissantes preuve de cela .	5	1
R-coh0 ;	la guerre actuelle en irak est peut-être le plus fort des preuves de tout cela .	6	1
oracleBLEU, ;	la guerre actuelle en irak est sans doute le plus sérieux la preuve de cela .	8	3

TAB. C.2 – Continued

modèle	traduction	juge1	juge2
B-cpl0-coh0 ;B-cpl1-coh0 ;B-coh1 ;B-coh2 ;B-coh3 ;B-coh0 ;B-coh4 ;B-coh5 ;B-cpl3 ;B-cpl4 ;B-cpl5 ;B-cpl6 ;B-cpl8 ;B-cpl7 ;B-cpl1 ;fB-cpl0-coh0 ;fB-cpl1-coh0 ;fB-coh2 ;fB-coh3 ;fB-coh0 ;fB-coh4 ;fB-coh5 ;fB-cpl3 ;fB-cpl4 ;fB-cpl5 ;fB-cpl0 ;fB-cpl7 ;fB-cpl1 ;N-cpl1-coh0 ;N-coh1 ;N-coh2 ;N-coh3 ;N-coh0 ;N-coh5 ;N-cpl3 ;N-cpl5 ;N-cpl7 ;N-cpl1 ; baseline, ;	la guerre actuelle en irak est peut-être le plus fort preuve de cela .	3	1

C.1.2 Jugements de fluidité

modèle	traduction	juge1	juge2
	ID :120		
B-cpl0-coh0 ;B-coh1 ;B-coh2 ;B-coh3 ;B-coh4 ;B-cpl3 ;B-cpl4 ;B-cpl5 ;B-cpl6 ;B-cpl8 ;B-cpl0 ;B-cpl1 ;fB-coh4 ;fB-cpl3 ;fB-cpl4 ;fB-cpl5 ;fB-cpl6 ;fB-cpl0 ;fB-cpl7 ;fB-cpl1 ;N-cpl0-coh0 ;N-coh1 ;N-coh2 ;N-coh3 ;N-coh4 ;N-cpl3 ;N-cpl4 ;N-cpl5 ;N-cpl6 ;N-cpl8 ;N-cpl0 ;N-cpl7 ;N-cpl1 ;	"un " retour aux bases " approche est vital dans trois domaines inter-linked , dans lequel les gouvernements nationaux doivent prendre la tête : réduire la corruption , l' application des droits de propriété , et l' application cohérente de l' état de droit .	4	4
B-cpl1-coh0 ;B-coh0 ;B-coh5 ;fB-cpl0-coh0 ;fB-cpl1-coh0 ;fB-coh0 ;fB-coh5 ;N-cpl1-coh0 ;N-coh0 ;N-coh5 ;	un " retour aux bases « approche est vital dans trois domaines inter-linked , dans lequel les gouvernements nationaux doivent prendre la tête : réduire la corruption , l' application des droits de propriété , et l' application cohérente de l' état de droit .	4	3
oraclewpBLEU, ; oracleBLEU, ;	une approche » par « retour à inter-linked est essentielle dans trois domaines , dans lequel les gouvernements nationaux doivent prendre la tête : réduire la corruption , l' application des droits de propriété et l' application cohérente de l' état de droit .	2	4
fB-coh1 ;fB-coh3 ;fB-cpl8 ; baseline, ;	par « retour à une « approche est vital dans trois domaines inter-linked , dans lequel les gouvernements nationaux doivent prendre la tête : réduire la corruption , l' application des droits de propriété , et l' application cohérente de l' état de droit .	7	4
oracleNIST, ;	une approche » par « retour à inter-linked est essentielle dans trois domaines dans lesquels les gouvernements nationaux doivent prendre la tête : réduire la corruption , l' application des droits de propriété et l' application cohérente de l' état de droit .	1	1
fB-coh2 ;	par « retour à une « approche est vital dans trois domaines inter-linked , dans lequel les gouvernements nationaux doivent prendre la tête : réduire la corruption , l' application des droits de propriété , et application cohérente de l' état de droit .	7	4

TAB. C.3 – Continued

modèle	traduction	juge1	juge2
R-cpl4 ;	une approche » par « retour à inter-linked est vital dans trois domaines , dans lequel les gouvernements nationaux doivent prendre la direction : réduire la corruption , l' application des droits de propriété , et l' application cohérente de l' état de droit .	3	4
R-coh0 ;R-cpl6-coh0 ;R-cpl6-coh4 ;R-cpl6-coh5 ;	une approche » par « retour à inter-linked est vital dans trois domaines , dans lequel les gouvernements nationaux doivent prendre la tête : réduire la corruption , l' application des droits de propriété et l' application cohérente de l' état de droit .	4	2
B-cpl7 ;	par « retour à une « approche est vital dans trois inter-linked domaines , dans lequel les gouvernements nationaux doivent prendre la tête : réduire la corruption , l' application des droits de propriété , et l' application cohérente de l' état de droit .	7	4
	ID :125		
oracleNIST, ;	une structure claire de règles formelles est le meilleur antidote dans ce cas à ces défauts , car ils veillent les coutumes informel sur lesquelles tant dans la société dépend .	1	1
fB-coh2 ;	une structure claire de règles formelles est le meilleur antidote dans ce cas à ces défauts , car ils veillent les coutumes informel sur lequel tant dans la société dépend .	2	2
R-coh0 ;	une structure claire des règles formelles constitue le meilleur antidote dans ce cas à ces défauts , car ils veillent les coutumes informel sur lequel tant dans la société dépend .	3	2
oracleBLEU, ;	une structure claire des règles formelles est le meilleur antidote dans ce cas à ces défaillances , car ils veillent les coutumes informel sur lequel tant dans la société dépend .	3	2
R-cpl6-coh0 ;R-cpl6-coh4 ;R-cpl6-coh5 ;	une structure claire des règles formelles est le meilleur antidote dans ce cas de ces failles , pour qu' ils veillent aux coutumes informel sur lequel tant dans la société dépend .	3	3
R-cpl4 ;	une répartition claire des règles formelles est le meilleur antidote dans ce cas à ces défauts , car ils veillent à des coutumes informel sur lequel tant dans la société dépend .	3	2
N-cpl0-coh0 ;N-cpl8 ;N-cpl0 ;	une structure claire des règles formelles est le meilleur antidote dans ce cas de ces failles , car ils veillent les coutumes informel sur lequel tant dans la société dépend .	3	2

TAB. C.3 – Continued

modèle	**traduction**	**juge1**	**juge2**
B-cpl0-coh0 ;B-cpl1-coh0 ;B-coh1 ;B-coh2 ;B-coh3 ;B-coh0 ;B-coh4 ;B-coh5 ;B-cpl3 ;B-cpl4 ;B-cpl5 ;B-cpl6 ;B-cpl8 ;B-cpl0 ;B-cpl7 ;B-cpl1 ;fB-cpl0-coh0 ;fB-cpl1-coh0 ;fB-coh1 ;fB-coh3 ;fB-coh0 ;fB-coh4 ;fB-coh5 ;fB-cpl3 ;fB-cpl4 ;fB-cpl5 ;fB-cpl6 ;fB-cpl0 ;fB-cpl7 ;fB-cpl1 ;N-cpl1-coh0 ;N-coh1 ;N-coh2 ;N-coh3 ;N-coh0 ;N-coh4 ;N-coh5 ;N-cpl3 ;N-cpl4 ;N-cpl5 ;N-cpl6 ;N-cpl7 ;N-cpl1 ; baseline, ;	une structure claire des règles formelles est le meilleur antidote dans ce cas à ces défauts , car ils veillent les coutumes informel sur lequel tant dans la société dépend .	3	2
oraclewpBLEU, ;	une structure de règles formelles est le meilleur antidote dans ce cas à ces défauts , car ils renforcer les coutumes informel sur lequel tant dans la société dépend .	9	3
fB-cpl8 ;	une structure claire des règles formelles est le meilleur antidote dans ce cas à ces défauts , car ils renforcer les coutumes informel sur lequel tant dans la société dépend .	10	3
	ID :156		
B-cpl0-coh0 ;B-cpl1-coh0 ;B-coh1 ;B-coh2 ;B-coh3 ;B-coh0 ;B-coh4 ;B-coh5 ;B-cpl3 ;B-cpl4 ;B-cpl5 ;B-cpl6 ;B-cpl8 ;B-cpl0 ;B-cpl7 ;B-cpl1 ;fB-cpl0-coh0 ;fB-cpl1-coh0 ;fB-coh1 ;fB-coh2 ;fB-coh3 ;fB-coh0 ;fB-coh4 ;fB-coh5 ;fB-cpl3 ;fB-cpl4 ;fB-cpl5 ;fB-cpl6 ;fB-cpl8 ;fB-cpl0 ;fB-cpl7 ;fB-cpl1 ;N-cpl0-coh0 ;N-cpl1-coh0 ;N-coh1 ;N-coh2 ;N-coh3 ;N-coh0 ;N-coh4 ;N-coh5 ;N-cpl3 ;N-cpl4 ;N-cpl5 ;N-cpl6 ;N-cpl8 ;N-cpl0 ;N-cpl7 ;N-cpl1 ; oraclewpBLEU, ; oracleBLEU, ; oracleNIST, ; baseline, ;	absolument pas .	1	1
R-coh0 ;	il faut non pas l' union européenne .	4	3
R-cpl6-coh0 ;R-cpl6-coh5 ;	ce n' est pas absolument à ce sujet .	3	2
R-cpl4 ;R-cpl6-coh4 ;	ce n' est pas tout à fait l' union européenne .	2	1
	ID :172		
oraclewpBLEU, ;	parfois , les parlements décident à agir rapidement , et peut-être devons le faire , mais plutôt comme un état , ils prendre suffisamment de temps pour examiner les questions totalement .	14	5
R-cpl6-coh4 ;	parfois , les parlements veulent agir rapidement , et peut-être devons le faire , mais comme une règle ils prennent pas suffisamment de temps pour examiner des questions totalement .	4	4
B-coh1 ;B-coh2 ;B-coh3 ;B-cpl0 ;fB-coh3 ;fB-coh0 ;fB-cpl8 ;N-coh1 ;N-coh2 ;N-coh3 ;N-coh0 ;N-coh4 ;N-coh5 ;N-cpl4 ;N-cpl6 ;	parfois , les parlements souhaitent agir rapidement , et peut-être devons le faire , mais comme une règle ils prendre suffisamment de temps à examiner des questions totalement .	11	6
B-cpl0-coh0 ;B-cpl1-coh0 ;B-coh0 ;	parfois , les parlements décident à agir rapidement , et peut-être doivent le faire , mais comme une règle ils prendre suffisamment de temps à examiner des questions totalement .	5	5

TAB. C.3 – Continued

modèle	traduction	juge1	juge2
fB-cpl0-coh0 ;fB-cpl0 ;	parfois , les parlements souhaitent agir rapidement , et peut-être devons le faire , mais comme une règle ils prennent suffisamment de temps à examiner des questions totalement .	7	3
B-cpl4 ;B-cpl7 ;fB-cpl4 ;fB-cpl7 ;N-cpl7 ;N-cpl1 ; baseline, ;	parfois , les parlements décident à agir rapidement , et peut-être devons le faire , mais comme une règle ils prennent suffisamment de temps à examiner des questions totalement .	7	3
B-coh4 ;N-cpl1-coh0 ;	parfois , les parlements décident à agir rapidement , et peut-être doivent le faire , mais comme une règle ils prennent suffisamment de temps pour examiner des questions totalement .	2	1
oracleNIST, ;	parfois , les parlements souhaitent agir rapidement , et peut-être devons le faire , mais comme une règle qu' ils prennent suffisamment de temps pour examiner les questions totalement .	3	3
R-cpl6-coh0 ;R-cpl6-coh5 ;	parfois , les parlements veulent agir rapidement , et peut-être devons le faire , mais comme une règle qu' ils prennent pas suffisamment de temps pour examiner des questions totalement .	7	3
R-coh0 ;	parfois , les parlements décident à agir rapidement , et peut-être doivent le faire , mais comme une règle ils prennent pas suffisamment de temps pour examiner des questions totalement .	1	2
oracleBLEU, ;	parfois , les parlements souhaitent agir rapidement , et peut-être devons le faire , mais comme une règle ils prennent suffisamment de temps pour examiner les questions totalement .	7	2
fB-coh1 ;fB-coh2 ;N-cpl0-coh0 ;N-cpl3 ;N-cpl8 ;N-cpl0 ;	parfois , les parlements veulent agir rapidement , et peut-être devons le faire , mais comme une règle ils prendre suffisamment de temps à examiner des questions totalement .	11	5
R-cpl4 ;	parfois , les parlements veulent agir rapidement , et peut-être devons le faire , mais comme une règle qu' ils prennent pas suffisamment de temps pour examiner les questions totalement .	6	3
B-coh5 ;B-cpl3 ;B-cpl5 ;B-cpl6 ;B-cpl8 ;B-cpl1 ;fB-cpl1-coh0 ;fB-coh4 ;fB-coh5 ;fB-cpl3 ;fB-cpl5 ;fB-cpl6 ;fB-cpl1 ;N-cpl5 ;	parfois , les parlements décident à agir rapidement , et peut-être devons le faire , mais comme une règle ils prendre suffisamment de temps à examiner des questions totalement .	11	5
	ID :181		
oraclewpBLEU, ;	le partenariat jonathan du premier ministre , tony blair et le chancelier de l' échiquier gordon brown , est un bon exemple de cette tactique .	11	3
R-cpl6-coh4 ;	le partenariat inquiets du premier ministre , tony blair et du chancelier de l' échiquier gordon brown , est un bon exemple de cette tactique .	3	1
B-cpl6 ;fB-cpl4 ;fB-cpl6 ;fB-cpl0 ;N-coh0 ;N-coh4 ;N-coh5 ;N-cpl4 ;N-cpl6 ;	le partenariat inquiets du premier ministre tony blair et chancelier de l' échiquier gordon brown est un bon exemple de cette tactique .	9	4
fB-cpl0-coh0 ;N-coh1 ;N-coh2 ;	le partenariat difficiles du premier ministre tony blair et chancelier de l' échiquier gordon brown est un bon exemple de cette tactique .	4	3

TAB. C.3 – Continued

modèle	**traduction**	**juge1**	**juge2**
R-cpl6-coh5 ;	le partenariat difficiles du premier ministre , tony blair et du chancelier de l' échiquier gordon brown est un bon exemple de cette tactique .	1	2
B-cpl1-coh0 ;B-coh0 ;B-coh4 ;B-coh5 ;B-cpl3 ;B-cpl4 ;B-cpl5 ;B-cpl8 ;B-cpl7 ;B-cpl1 ;fB-cpl1-coh0 ;fB-coh1 ;fB-coh2 ;fB-coh3 ;fB-coh0 ;fB-coh4 ;fB-coh5 ;fB-cpl3 ;fB-cpl5 ;fB-cpl7 ;fB-cpl1 ;N-cpl1-coh0 ;N-cpl8 ;N-cpl0 ;N-cpl7 ;N-cpl1 ;	le partenariat inquiets du premier ministre tony blair et le chancelier de l' échiquier gordon brown est un bon exemple de cette tactique .	10	3
R-coh0 ;R-cpl6-coh0 ;	le partenariat difficiles du premier ministre , tony blair et du chancelier de l' échiquier gordon brown , est un bon exemple de cette tactique .	1	1
baseline, ;	le partenariat difficiles du premier ministre tony blair et le chancelier de l' échiquier gordon brown , est un bon exemple de cette tactique .	7	3
R-cpl4 ;	le partenariat précaire du premier ministre tony blair et chancelier de l' échiquier gordon brown , est un bon exemple de cette tactique .	4	2
oracleBLEU, ; oracleNIST, ;	le partenariat précaire du premier ministre tony blair et chancelier de l' échiquier gordon brown est un bon exemple de cette tactique .	4	2
B-cpl0-coh0 ;B-coh1 ;B-coh2 ;B-coh3 ;B-cpl0 ;fB-cpl8 ;N-cpl0-coh0 ;N-coh3 ;N-cpl3 ;N-cpl5 ;	le partenariat difficiles du premier ministre tony blair et le chancelier de l' échiquier gordon brown est un bon exemple de cette tactique .	7	3

Annexe D

Tests de significativité

D.1 Test des signes

Le test des signes (*sign test*) s'applique au cas de deux échantillons appariés. Il est basé uniquement sur l'étude des signes des différences observées entre les paires d'individus, quelles que soient les valeurs de ces différences.

Méthode. Supposons, que $p = P(X > Y)$. Alors, l'hypothèse nulle suppose pour une paire d'individus (x_i, y_i) que x_i est meilleur que y_i est aussi probable que son contraire ($H_0 : p = 0,50$). Nous avons n observations appariées indépendantes $\{(x_1, y_1), ..., (x_n, y_n)\}$ (uniquement les couples pour lesquels $x_i \neq y_i$ sont considérés).

Si dans la plupart des cas $y_i > x_i$, une observation w de la variable W est un nombre d'occurrences où $y_i > x_i$. Si l'hypothèse H_0 est vraie, alors $W \sim B(n, 0,5)$.

Le niveau de siginificativité est la probabilité $P(W \leq w)$ sous l'hypothèse H_0. Si cette probabilité est faible, nous rejetons l'hypothèse H_0, et acceptons l'hypothèse $H_1 : p < 0,50$.

D.2 Significativité des comparaisons des différents modèles de reclassement deux à deux

TAB. D.1 – Systèmes jugés significativement moins bons (adéquation) que le modèle de base (baseline).

Modèle	jugés +	jugés =	p-valeur
oracleNIST	0,38	0,45	0
wpBLEU, base, gén(RR) incoh	0,32	0,5	0,01
NIST, base, gén(RR) coh	0,31	0,5	0,01
NIST, base, gén(RR) lab_giza	0,31	0,51	0,01
NIST, base, gen_giza, lex_giza, gén(RR) lab_giza	0,3	0,51	0,01
NIST, base, gén(RR) coh, gén(RR) incoh	0,3	0,51	0,01
wpBLEU, base, gen_w2w, lex_w2w, gén(PR) lab_w2w	0,3	0,5	0,01
wpBLEU, base, gén(RR) coh, gén(RR) incoh	0,29	0,52	0,01
NIST, base, gén(PR) lab_w2w, gén(RR) coh, gén(RR) incoh	0,29	0,54	0,01
BLEU, base, gén(RR) coh	0,28	0,54	0,01
BLEU, base, gen_giza, lex_giza, gén(RR) lab_giza	0,28	0,56	0,01
wpBLEU, base, gén(RR) coh	0,28	0,54	0,01
wpBLEU, base, gén(PR) lab_c2c, gén(RR) coh, gén(RR) incoh	0,28	0,5	0,01
wpBLEU, base, gén(PR) lab_w2w, gén(RR) coh, gén(RR) incoh	0,28	0,5	0,01
NIST, base, gén(PR) lab_c2c	0,27	0,54	0,01
wpBLEU, base, gen_giza, lex_giza, gén(PR) lab_giza	0,27	0,54	0,01
wpBLEU, base, gén(PR) coh	0,27	0,54	0,01
wpBLEU, base, gén(PR) lab_c2c	0,27	0,5	0,01
NIST, base, gen_giza, lex_giza, gén(PR) lab_giza	0,27	0,53	0,01
NIST, base, gén(PR) lab_c2c, gén(RR) coh, gén(RR) incoh	0,27	0,54	0,01
wpBLEU, base, gén(PR) lab_w2w	0,27	0,53	0,01
BLEU, base, gén(PR) coh	0,26	0,54	0,01
BLEU, base, gén(PR) lab_c2c	0,26	0,54	0,01
wpBLEU, base, gen_c2c, lex_c2c, gén(PR) lab_c2c	0,26	0,49	0,01
BLEU, base, gén(RR) lab_giza	0,26	0,58	0,01
BLEU, base, gén(PR) lab_c2c, gén(RR) coh, gén(RR) incoh	0,26	0,59	0,01
NIST, base, gén(PR) coh	0,26	0,56	0,01
NIST, base, gén(PR) lab_giza	0,26	0,53	0,01
NIST, base, gen_c2c, lex_c2c, gén(PR) lab_c2c	0,25	0,57	0,01
NIST, base, gén(PR) lab_w2w	0,25	0,58	0,01
BLEU, base, gen_c2c, lex_c2c, gén(PR) lab_c2c	0,25	0,61	0,01
BLEU, base, gén(PR) lab_giza	0,25	0,6	0,01
BLEU, base, gén(PR) lab_w2w	0,25	0,59	0,01
BLEU, base, gén(RR) coh, gén(RR) incoh	0,24	0,61	0,01
BLEU, base, gen_giza, lex_giza, gén(PR) lab_giza	0,24	0,61	0,01
BLEU, base, gén(PR) lab_w2w, gén(RR) coh, gén(RR) incoh	0,24	0,6	0,01
BLEU, base, gén(RR) coh, gén(RR) incoh	0,24	0,61	0,01
wpBLEU, base, gén(PR) lab_giza	0,24	0,56	0,01

TAB. D.2 – Systèmes jugés significativement meilleurs (adéquation) que le modèle de base (baseline).

Modèle	jugés +	jugés =	p-valeur
RealRef, gen_giza, lex_giza, gén(RR) lab_giza	0,43	0,44	0,0000
oraclewpBLEU	0,32	0,44	0,0066
RealRef, gén(RR) coh, gén(RR) incoh	0,31	0,5	0,0069
oracleBLEU	0,29	0,44	0,0081
RealRef, gén(RR) lab_giza, gén(RR) coh	0,39	0,41	0,0139
NIST, moses, gén(PR) coh, gén(PR) incoh	0,23	0,56	0,0168
RealRef, gén(RR) lab_giza, gén(RR) incoh	0,39	0,4	0,0173
wpBLEU, moses, gén(PR) incoh	0,22	0,56	0,0199