Le rôle des inférences pour la fouille d'opinion : applications aux réseaux sociaux en langue chinoise

Liyun Yan

Institut National des Langues et Civilisations Orientales

École doctorale n°265 : *Langues, Littératures et Sociétés du monde*

Équipe de recherche textes, informatique, multilinguisme

thèse

présentée par

Liyun Yan

soutenue le 5 juillet 2021

pour obtenir le grade de docteur de l'inalco

Science du langage et langues appliquées : ingénierie linguistique

Le rôle des inférences pour la fouille d'opinion : applications aux réseaux sociaux en langue chinoise

Thèse dirigée par :

M. Mathieu Valette	Professeur des universités, INALCO, ERTIM
M. Cyril Grouin	Ingénieur de recherche, Université Paris-Saclay, CNRS, LISN

Rapporteurs :

M^me^ Farah Benamara	Maître de conférences, Université Paul Sabatier, CNRS, IRIT
M. Dominique Legallois	Professeur des universités, Université Paris 3 Sorbonne Nouvelle, CNRS, LaTTICe

---

Membres du jury :

M. Mathieu Valette	Professeur des universités, INALCO, ERTIM
M. Cyril Grouin	Ingénieur de recherche, Université Paris-Saclay, CNRS, LISN
M^me^ Farah Benamara	Maître de conférences, Université Paul Sabatier, CNRS, IRIT
M. Dominique Legallois	Professeur des universités, Université Paris 3 Sorbonne Nouvelle, CNRS, LaTTICe
M^me^ Yue Ma	Maître de Conférence, Université Paris-Saclay, CNRS, LISN
M^me^ Christine Lamarre	Professeur des universités, INALCO, EHESS, CNRS, CRLAO

Résumé

Cette thèse s'intéresse à l'inférence linguistique dans la fouille d'opinion dans un corpus des commentaires touristiques en chinois. Les techniques existantes qui sont bien développées sur des opinions courtes et explicites donnent des résultats limités dans l'interprétation des contextes implicites. De plus, l'expression de l'opinion met en œuvre des stratégies énonciatives différentes suivant les langues et les cultures. Notre hypothèse de travail consiste à étudier les inférences pour améliorer la fouille d'opinion. Dans cette perspective, notre première contribution propose une typologie des inférences pour le chinois en 5 types : logique, pragmatique, lexicale, énonciative et discursive (Rossi et Campion, 1999 ; Marin, 2004 ; Duchêne, 2008 ; Doucy et Massoussi, 2012). Nous avons appliqué cette typologie pour annoter un corpus, dans l'objectif de mener des expériences de fouille d'opinion avec et sans le traitement des inférences. Notre deuxième contribution se focalise sur la classification automatique des inférences en nous basant sur les caractéristiques linguistiques, les métadonnées du domaine et les vecteurs du plongement de mots. L'objectif d'une part est de prouver que le traitement des inférences améliore la performance de la fouille d'opinion et d'autre part de trouver une solution équilibrée entre la classification manuelle couteuse et la classification automatique. Dans ce travail de thèse, nous avons démontré l'intérêt d'étudier les inférences pour réaliser une fouille d'opinion en chinois. Toutefois, l'identification automatique des inférences reste complexe et nécessite une poursuite des travaux de recherche.

Mots-clés : Fouille d'opinion, Inférence, Langue chinoise, Réseaux sociaux

Abstract

This thesis is interested in linguistic inference in opinion mining in a corpus of tourist commentaries in Chinese. Existing techniques which are well developed on short and explicit opinions, give limited results in interpreting implicit contexts. In addition, expression of opinion implements different enunciative strategies according to languages and cultures. Our hypothesis consists in studying inferences to improve opinion mining. In this perspective, our first contribution proposes a typology of inferences for Chinese in 5 types : logical, pragmatic, lexical, enunciative and discursive (Rossi and Campion, 1999 ; Marin, 2004 ; Duchêne, 2008 ; Doucy and Massoussi, 2012). We applied this typology to annotate a corpus, with the objective of conducting opinion mining experiments with and without the processing of inferences. Our second contribution focuses on automatic classification of inferences based on linguistic characteristics, domain metadata and word embedding vectors. The objective on the one hand is to prove that the processing of inferences improves the performance of opinion mining and on the other hand to find a balanced solution between expensive manual annotation and automatic classification. In this thesis, we demonstrated the interest of studying inferences for opinion mining in Chinese. However, the automatic identification of inferences remains complex and requires further research.

Keywords : Opinion Mining, Inference, Chinese Language, Social Network

Remerciements

En premier lieu, je remercie chaleureusement mon directeur de thèse Mathieu Valette ainsi que mon co-encadrant Cyril Grouin, pour leur écoute et leurs conseils avisés dispensés tout au long de cette thèse.

Je remercie spécialement à Mathieu. Grâce à sa grande compétence, son professionnalisme, il m'a éclairée sur de nombreux domaines de la recherche scientifique, tout en m'ayant offert des conditions de travail idéales permettant de m'épanouir librement dans mes recherches.

Je remercie aussi particulièrement à Cyril, qui me guide et m'accompagne depuis cinq ans avec une inébranlable bonne humeur. Tous ses efforts se reflètent dans les réunions hebdomadaires, les retours immédiats et les environ 1000 allers-retours de courriels.

Merci aux membres de l'équipe ERTIM. Pendant mes études en France, l'opportunité la plus précieuse est de rencontrer cette équipe formidable. Merci en particulier à Jean-Michel Daube, qui a voulu résoudre tous les problèmes administratifs d'une étudiante étrangère depuis le début du Master, à François Stuck avec qui j'ai pu avoir de nombreuses discussions qui m'ont beaucoup aidé, à Damien Nouvel qui est toujours disponible pour le support technique et à tous les professeurs du Master ingénierie multilingue qui m'ont fait découvrir et m'intéresser au domaine de TAL.

Merci à Jennifer, Qinran, Manying, Manon, Bénédict, Lucie, Johanna et Amélie qui m'ont beaucoup aidée non seulement avec l'annotation manuelle et la relecture en français et en anglais, mais aussi tous les échanges passionnés. Merci à Mei, Danni et Yuning pour l'annotation manuelle de haute qualité.

Un grand merci à mon cher grand-père, de m'avoir soutenue sans rien attendre en retour jusqu'au dernier moment de sa vie. Merci à mes parents pour les échanges téléphoniques de tous les jours et la grande patience visant non seulement le retard de ma thèse, et aussi mes doutes et mon choix.

Je remerci également tous mes amis en France et en Chine, en particulier à Chantal, qui m'a fait découvrir l'INACO pour la première fois, et avec qui la plus précieuse amitié s'est maintenue pendant 10 ans, à Xiaoting pour son accompagnement quotidien et son

soutien à tout moment, à Gengjing pour chaque rencontres et les partages à distance, à Richard pour sa bienveillance et les repas toujours délicieux et à toutes les personnes qui m'ont aidée dans la réalisation de cette thèse.

Merci à ces cinq années de thèse, qui me permettent d'avoir une attitude plus tolérante d'aborder l'Autre.

Table des matières

Résumé i

Abstract iii

Remerciements v

1 Introduction 1
1.1 Problématique 1
1.2 Objectif 2
1.3 Plan de thèse 2

PREMIÈRE PARTIE : ÉTAT DE L'ART : INFÉRENCES ET FOUILLE D'OPINION 5

2 L'état de l'art 7
2.1 Les inférences 7
2.1.1 Introduction 7
2.1.2 Définitions 8
2.1.3 Modalité d'analyse des inférences 13
2.1.4 Classification des types d'inférences 16
2.2 La fouille d'opinion 18
2.2.1 Introduction 18
2.2.2 Fouille d'opinion en général 18
2.2.3 Inférence dans la fouille d'opinion 20
2.2.4 Implicite dans la fouille d'opinion 20
2.3 Conclusion 22

DEUXIÈME PARTIE : PRÉSENTATION ET CONSTRUCTION DES RESSOURCES LINGUISTIQUES 23

3 Corpus d'avis touristiques 25
3.1 Introduction 25
3.2 Identification des sites 25
3.2.1 Origine du corpus 25

3.2.2 Informations disponibles 26
3.3 Critères de sélection des messages 27
3.4 Extraction automatique du corpus 27
3.5 Prétraitement du corpus 28
3.5.1 Segmentation 28
3.5.2 Transformation du chinois traditionnel vers chinois simplifié 30
3.5.3 Standardisation des entités nommées 31
3.5.4 Partie du discours 31
3.6 Conclusion 34

4 Identification manuelle des inférences dans un corpus de commentaires 37
4.1 Introduction 37
4.2 Guide d'annotation 38
4.2.1 Objectifs 38
4.2.2 Annotation des inférences 38
4.2.3 Annotation des types d'inférences 40
4.2.4 Niveaux d'analyse 41
4.3 Processus d'annotation 42
4.3.1 Sélection des commentaires à annoter 42
4.3.2 Outil et configuration d'annotation 43
4.3.3 Accords inter-annotateurs 45
4.3.4 Difficultés rencontrées 47
4.4 Analyse des résultats d'annotation 48
4.4.1 Corrélation et distribution des catégories selon différents niveaux 48
4.4.2 Analyse textométrique du corpus 52
4.5 Conclusion 55

TROISIÈME PARTIE : IDENTIFICATION ET CLASSIFICATION DES INFÉRENCES POUR LA FOUILLE D'OPINION EN CHINOIS 57

5 Identification et classification automatiques des inférences 59
5.1 Introduction 59
5.2 Présentation des caractéristiques et des méthodes 60
5.2.1 Corpus d'entraînement 60
5.2.2 Caractéristiques généralisées 60
5.2.3 Plongement des mots 61
5.3 Expérimentation 62
5.3.1 Identification des inférences 62
5.3.2 Classification des types d'inférence 63
5.3.3 Amélioration des performances 65

5.4 Analyse des résultats 65
5.4.1 Identification des inférences 66
5.4.2 Classification des types d'inférence 67
5.4.3 Amélioration des performances du modèle 70
5.5 Conclusion 72

6 Inférences dans la fouille d'opinion 73
6.1 Introduction 73
6.2 Méthode lexicale pour la fouille d'opinion 74
6.2.1 Ressources : ontologie émotionnelle chinoise 74
6.2.2 Taux couverture de la ressource linguistique 75
6.2.3 Application de l'ontologie 76
6.2.4 Méthode insuffisante 77
6.3 Fouille d'opinion avec les inférences annotées manuellement 77
6.3.1 Prédiction automatique de la polarité avec SVM 77
6.3.2 Détection automatique du thème 82
6.4 Fouille d'opinion avec les inférences annotées automatiquement 83
6.4.1 Prédiction de la polarité 84
6.4.2 Détection des thèmes 84
6.5 Conclusion 86

7 Ajustement et application du modèle 87
7.1 Introduction 87
7.2 Équilibre entre l'annotation manuelle et automatique des inférences 87
7.3 Analyses linguistiques des résultats 89
7.3.1 Discussion des résultats par rapport aux théories linguistiques sur les inférences 90
7.3.2 Phénomène linguistique observé dans les résultats 95
7.4 Conclusion 98

8 Conclusion 99
8.1 Contribution 99
8.1.1 Corpus annoté avec les inférences, la polarité et les thèmes 99
8.1.2 Classification automatique des inférences 100
8.2 Perspectives 101
8.2.1 À court terme 101
8.2.2 À long terme 102

Bibliographie 105

Liste des tableaux 115

Liste des figures 117

Annexes 119

A Guide d'annotation (version française) 119

B Guide d'annotation (version chinoise) 135

C Exemple du planning des annotations 143

D Publications 145

D.1 Analyse des inférences pour la fouille d'opinion en chinois, CORIA-TALN-RJC, May 2018, Rennes, France 145

D.2 Inference Annotation of a Chinese Corpus for Opinion Mining, LREC, May 2020, Marseille, France 145

Chapitre 1

Introduction

Contents

1.1 Problématique 1
1.2 Objectif 2
1.3 Plan de thèse 2

1.1 Problématique

L'essor d'Internet permet aux utilisateurs d'échanger facilement leurs opinions et sentiments sur divers aspects de la vie quotidienne. Cette possibilité d'expression rapide constitue un enjeu de veille pour les entreprises (étude de la réputation, avis de satisfaction clientèle, etc.) et modifie également le mode de pensée des utilisateurs, soit par la possibilité de laisser un nombre élevé de commentaires de peu d'intérêt, soit par la possibilité de se retrancher derrière un commentaire anonyme pour exprimer un avis négatif (que l'absence d'anonymat n'aurait pas permis). Les messages laissés par d'anciens clients témoignent de différences culturelles et sociales, dans le choix des critères d'évaluation (taille des chambres, présence d'équipements et services dans l'hôtel), dans l'utilisation du vocabulaire (terme générique vs terme spécifique au domaine), et dans la manière d'exprimer une information (en particulier les éléments jugés négatifs).

La compréhension d'une opinion chez l'homme est une procédure simple. En raison de la masse de données disponibles sur Internet, il est nécessaire de disposer d'outils automatiques de fouille d'opinion pour analyser le contenu et dégager les tendances exprimées. Pourtant, la procédure de compréhension est complexe pour une machine. Il comprend la reconnaissance des caractères, la segmentation des phrases en mots, la désambiguïsation, l'extraction d'informations, la détection d'éléments d'une opinion et l'analyse des sentiments. Les techniques existantes qui sont bien développées sur des opinions courtes et explicites donnent des résultats limités dans l'interprétation des contextes implicites. De plus, l'expression d'opinion met en œuvre différentes stratégies énonciatives suivant les langues et les cultures. Ce genre d'opinions contient de nombreuses inférences qui nécessite une analyse dédiée.

1.2 Objectif

Dans le cadre de cette thèse, nous nous intéressons à l'inférence linguistique dans la fouille d'opinion dans un corpus des commentaires touristiques en chinois provenant de deux sites (Booking.com et mafengwo.cn). L'objectif de ce travail est alors d'étudier l'inférence linguistique et ensuite d'implanter l'inférence dans l'apprentissage automatique de la fouille d'opinion.

De ce fait, nous proposons une typologie des inférences basée sur l'état de l'art de la définition et de la classification des inférences (Kintsch 1998 ; Rossi et Campion 1999 ; Dufaye 2001 ; Martin 2004 ; Duchêne 2008 ; Doucy et Massoussi 2012 ; Bouquiaux et Leclercq 2017). Les cinq types d'inférences (logique, pragmatique, lexicale, énonciative, discursive) proposés permettent de caractériser et de distinguer les opinions implicites. Cette typologie est appliquée pour annoter un corpus, dans l'objectif de mener des expériences de fouille d'opinion avec et sans le traitement des inférences et de prouver que le traitement des inférences améliore la performance de la fouille d'opinion.

Le deuxième objectif de cette thèse est de réaliser la classification automatique des inférences en fonction des caractéristiques linguistiques, des métadonnées du domaine et des vecteurs du plongement de mots. Comme l'annotation manuelle est couteuse, l'objectif est de trouver une solution équilibrée entre la classification manuelle et la classification automatique.

1.3 Plan de thèse

Ce manuscrit est organisé en trois parties. La première partie porte sur l'état de l'art des inférences et de la fouille d'opinion. La deuxième partie présente la construction de ressources linguistiques qui concernent à la fois la collecte de corpus et l'annotation manuelle. Enfin, la troisième partie présente les expériences que nous avons menées et les résultats obtenus.

Première partie : Dans cette partie, nous introduisons l'état de l'art de cette thèse et comment notre travail se situer entre l'inférence et la fouille d'opinion.

- Chapitre 2 : Dans ce chapitre, nous présentons une sélection des travaux portant sur les inférences. Nous axons cette présentation autour des différents courants de recherche existants et des travaux de définition et de classification des inférences, en tenant compte de plusieurs dimensions linguistiques.
- Chapitre 3 : Dans ce chapitre, nous nous intéressons à la problématique de la fouille d'opinion que nous envisageons de lier, dans notre travail, avec celle des inférences.

Deuxième partie : Dans cette deuxième partie, nous présentons les ressources linguistiques et les corpus que nous avons constitués, les traitements appliqués sur ces corpus,

et les annotations réalisées sur ces corpus.

- Chapitre 4 : Ce chapitre présente les informations concernant le corpus : l'origine du corpus, les critères de sélection des messages, la réalisation du crawler avec Python, ainsi que les étapes de prétraitement du corpus.
- Chapitre 5 : Dans ce chapitre, nous présentons d'abord une annotation manuelle sur un sous-corpus représentatif du corpus complet. Cette annotation consiste à étiqueter d'un côté la présence des inférences et leurs types, de l'autre côté les informations de la fouille d'opinion comme la polarité et le thème. Ensuite, nous faisons des analyses sur ce corpus d'annotation.

Troisième partie :

- Chapitre 6 : Dans ce chapitre, nous nous concentrons sur l'identification et la classification automatique des inférences. Tout d'abord, nous définissons et normalisons les caractéristiques du modèle SVM. Ensuite, l'expérimentation concerne plusieurs modèles. À la fin, nous comparons les résultats des modèles afin de trouver une meilleure solution.
- Chapitre 7 : Dans ce chapitre, nous démontrons qu'une méthode lexicale n'est pas suffisante pour la fouille d'opinion. Notre méthode avec métadonnées et surtout inférences apporte une amélioration significative. Ensuite, nous comparons les résultats des inférences identifiées et classées manuellement et automatiquement afin de trouver une solution équilibrée entre l'annotation manuelle coûteuse et l'annotation automatique avec erreurs.
- Chapitre 8 : L'inférence et ses types se sont avérés efficaces pour la fouille d'opinion selon les expériences précédentes. Dans ce chapitre, nous nous concentrons sur l'amélioration du modèle. Les premières expériences visent à trouver une meilleure solution en prenant compte du prix et de la qualité de l'annotation manuelle. Ensuite, nous menons une discussion des résultats par rapport aux théories linguistiques sur les inférences. Enfin, nous présentions quelques phénomènes linguistiques mérités d'être étudiés dans les résultats.

Première partie

État de l’art : inférences et fouille d’opinion

Chapitre 2

L'état de l'art

Contents

2.1 Les inférences 7
- 2.1.1 Introduction 7
- 2.1.2 Définitions 8
- 2.1.3 Modalité d'analyse des inférences 13
- 2.1.4 Classification des types d'inférences 16

2.2 La fouille d'opinion 18
- 2.2.1 Introduction 18
- 2.2.2 Fouille d'opinion en général 18
- 2.2.3 Inférence dans la fouille d'opinion 20
- 2.2.4 Implicite dans la fouille d'opinion 20

2.3 Conclusion 22

2.1 Les inférences

Dans cette section, nous présentons une sélection de travaux portant sur les inférence. Nous axons cette présentation autour des différents courants de recherche existants et des travaux de définition et de classification des inférences, en tenant compte des différentes dimensions linguistiques.

2.1.1 *Introduction*

Etymologiquement, le terme "inférence" provient du latin *inferentia* qui a pour signification "conséquence" (Vittori 1609).

Dès son apparition, des philosophes comme Aristote ont utilisé ce terme pour désigner des syllogismes : « B appartient à A, est prédiqué de A ». La logique était introduite dans la science formelle.

La formalisation des propositions conduit à l'utilisation de la notion d'inférence en mathématiques géométriques (Cuel 2014), en statistique (Kern-Isberner et Eichhorn 2014 ;

Rodriguez et Müller 2013), et par la suite en linguistique (Denhière et Baudet 1992 ; Martin 1976), voire plus concrètement dans les enseignements bouddhistes de Xuanzang (M. Tang 2015) et la compréhension des inférences dans l'œuvre de Mengzi (Tanaka 2011).

Bien qu'il existe un grand nombre d'études sur les inférences, il n'y pas de consensus sur la définition et la classification uniforme des différents types d'inférence (Lavigne 2008), puisque tout travail de classification dépend à la fois du domaine scientifique et des objectifs spécifiques à l'étude.

2.1.2 *Définitions*

Nous distinguons les définitions présentes dans les dictionnaires, destinées aux locuteurs du français, de celles présentes dans les travaux de recherche, qui reposent sur des usages plus spécifiques et dont la description se destine à la communauté scientifique.

2.1.2.1 Définition dans les dictionnaires

Comparaison parallèle Dans différents dictionnaires, l'inférence est définie par différentes expressions :

1. Dans Larousse (2015), l'inférence est définie comme une "opération par laquelle on passe d'une assertion considérée comme vraie à une autre assertion au moyen d'un système de règles qui rend cette seconde assertion également vraie".
2. Le Robert définit l'inférence comme une opération logique par laquelle on admet une proposition en vertu de sa liaison avec d'autres propositions déjà tenues pour vraies.
3. Dans le dictionnaire en ligne de philosophie, l'inférence est une "Opération par laquelle on passe d'un ensemble de prémisses à une conclusion justifiée, rendue légitime par ces prémisses" [1].
4. Le dictionnaire en ligne du Centre National de Ressources Textuelles et Lexicales (CNRTL), l'inférence est une "opération qui consiste à admettre une proposition en raison de son lien avec une proposition préalable tenue pour vraie" [2].
5. Wiktionary propose deux définitions pour le terme "inférence": Elle est à la fois une "action d'inférer" et le "résultat de l'action d'inférer". Les quatre hyponymes de l'inférence sont "abduction", "déduction", "induction" et "rétroduction".
6. Dans le Nouveau Dictionnaire Encyclopédique des Sciences du Langage de Ducrot et Schaeffer (1999), ils définissent qu'alors que l'anaphore et la coordination sont des relations intérieures à un même texte, qui relient entre eux les énoncés d'un discours, l'inférence et la paraphrase mettent en rapport les énoncés abstraction faite des textes où ils prennent place.

Ces définitions s'accordent sur le fait que la production de l'inférence est une opération ou une action dynamique, et non statique. L'inférence permet de lier une conclusion à

1. `https://dicophilo.fr/definition/inference/`
2. `https://www.cnrtl.fr/definition/inf%C3%A9rence/substantif`

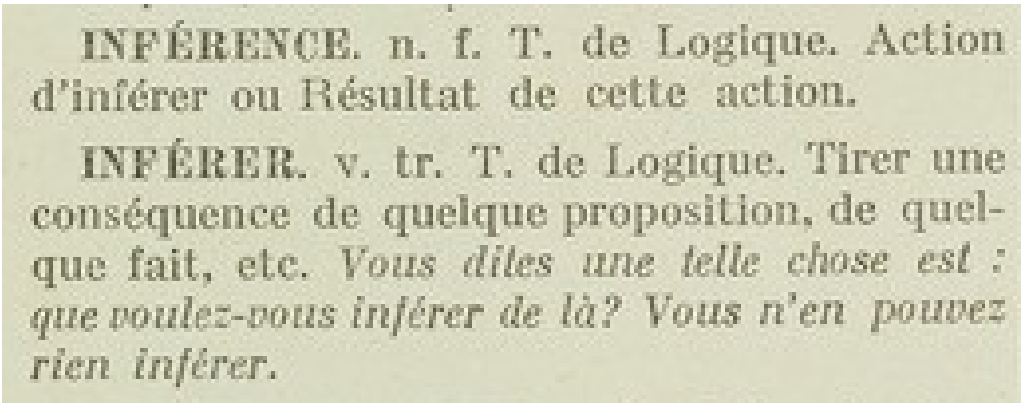
un ou plusieurs faits préalables. Certaines définitions insistent sur l'issue du processus (la vraisemblance de la conclusion) tandis que d'autres s'appuient sur l'explication du raisonnement mis en œuvre.

Par contre, la définition de ce mot n'apparaît pas dans tous les dictionnaires, par exemple Dictionnaire de linguistique et des sciences du langage de Larousse (1994) ne recense aucun mot autour du radical "infere". C'est peut-être parce qu'il s'agit d'une notion récente. Ce mot n'a pas encore été inclus dans le dictionnaire professionnel.

Évolution dans le temps La définition de l'inférence a également évolué au fil du temps. Nous avons comparé la définition dans les versions disponibles du Dictionnaire de l'Académie française, dont 8 versions papier publiées des années 1694, 1718, 1740, 1762, 1798, 1835, 1878 et 1935 ainsi qu'une version numérique récente (décembre 2019).

Le terme "inférence" est absent des sept premières versions qui ne contenaient que le verbe "inférer", défini comme le fait de"*tirer une conséquence de quelque proposition, de quelque fait, etc. Vous dites une telle chose est : que voulez-vous inférer de là ? Vous n'en pouvez rien inférer.*" (*Dictionnaire de L'Académie française* 1798).

L'inférence apparaît pour la première fois dans l'édition de 1935. Elle est alors décrite comme étant un terme de logique et renvoie à *action d'inférer ou au résultat de cette action* (Figure 2.1).


INFÉRENCE. n. f. T. de Logique. Action d'inférer ou Résultat de cette action.

INFÉRER. v. tr. T. de Logique. Tirer une conséquence de quelque proposition, de quelque fait, etc. *Vous dites une telle chose est : que voulez-vous inférer de là? Vous n'en pouvez rien inférer.*

Fig. 2.1 : Extrait du Dictionnaire de l'Académie française

Dans l'édition électronique de 2019[3], le Dictionnaire de l'Académie distingue le domaine de la logique de celui de l'informatique :

- *LOGIQUE : Opération consistant à établir qu'une proposition est vraie par le seul fait de sa liaison avec une ou plusieurs propositions dont la vérité a été précédemment établie. Raisonner par inférence. La déduction et l'induction sont deux formes particulières d'inférence.*

3. `https://www.dictionnaire-academie.fr/article/A9I1155`

– *INFORMATIQUE : Moteur d'inférence, type de programme utilisé en intelligence artificielle pour réaliser une telle opération logique.*

Quel que soit le domaine, ce concept renvoie à une opération de raisonnement qui existe entre une proposition et une ou plusieurs autres propositions. Le dictionnaire distingue deux formes d'inférence : la déduction et l'induction. Ces deux formes constituent un début de classification sur lequel nous reviendrons dans ce chapitre.

Dictionnaires européens et chinois Alors que les définitions précédentes provenaient des dictionnaires français, nous avons élargi notre recherche aux dictionnaires anglais et chinois, dans une perspective de comparaison, notamment pour vérifier l'évolution historique du concept en chinois.

Le *Cambridge Advanced Learner's Thesaurus* (2013) définit l'inférence comme étant une supposition ou une opinion réalisée sur la base des informations disponibles (*"a guess that you make or an opinion that you form based on the information that you have"*). Le *Cambridge Advanced Learner's Thesaurus* (2013) fournit comme exemples la supposition d'une attaque suite à de nombreux combats (*"They were warned to expect a heavy air attack and by inference many casualties"*) ou le changement d'avis d'une personne expliqué par l'influence supposée d'une autre personne (*"His change of mind was recent and sudden, the inference being that someone had persuaded him"*).

L'inférence en chinois est "推理" (pinyin pour phonétique : tuī lǐ). Il existe trois définitions de ce mot dans le dictionnaire 汉典[4] :

1. 推究整理 (déduire et démêler). 辞趣过诞，意旨迂阔，推理陈迹，恨为繁冗。——南朝梁肖绮《拾遗记序》(Exemple en chinois classique : Le style est trop étrange, le sens est détourné, le raisonnement et l'ordonnancement sont fastidieux. -Dynastie du Sud Liang Xiaoi)
2. 通过对一主题或材料的再三考查，而且通常不经实验证明或引入新资料而引申出概念或理论 (Dériver des concepts ou des théories en vérifiant à plusieurs reprises des sujets ou des matériaux, sans avoir besoin d'expériences scientifiques ou d'introduire de nouveaux matériaux.)
3. 逻辑学名词。通过一个或几个被认为是正确的陈述、声明或判断达到另一真理的行动，而这真理被相信是从前面的陈述、声明或判断中得出 (Terme logique. Le fait d'atteindre une autre vérité à travers une ou plusieurs déclarations ou jugements considérés comme vrais. Ce fait est considéré comme dérivée de déclarations ou de jugements antérieurs.)

Ces trois définitions considèrent que l'inférence provient soit d'une seule information, soit de la répétition de plusieurs observations. En chinois ancien, l'inférence n'avait pas seulement un sens de la déduction, mais aussi un sens de l'ordre. Mais en chinois moderne, le second sens a disparu.

Dans la section suivante, nous allons constater les différentes définitions apparues dans les travaux scientifiques.

4. `zdic.net`

2.1.2.2 Définitions dans les travaux scientifiques

Linguistique Le concept d'inférence est employé et étudié dans plusieurs travaux scientifiques, tant en termes de production que de compréhension. Il existe une distinction entre les différentes interprétations des domaines de recherche.

En linguistique descriptive, l'inférence est d'abord une procédure déterminante pour décrire l'ordre du sens entre plusieurs propositions. Une proposition *p* infère ou implique une proposition *q* si et seulement si, *p* étant vrai, *q* est nécessairement vraie. Martin (2004) donne comme exemple que : si *Elle a cueilli des roses* est vrai, alors *Elle a cueilli des fleurs* est vrai. Le sens ne sera plus valable si l'on inverse les deux propositions. La contraposition permet d'identifier et de décrire un grand nombre de phénomènes . Cette description ressemble à la relation entre les prémisses et la conclusion en logique. Cependant, elle emprunte la forme de la logique afin de décrire des faits linguistiques.

En sémantique, Doussau et Rigal (2011) rappellent la définition du Dictionnaire d'orthophonie (2004). Une inférence est un "*ajout d'informations n'étant pas explicitement données dans le message, mais que le lecteur peut déduire ou supposer à partir de ses propres connaissances générales sur le monde, établissant ainsi des liens entre les différentes parties du texte et permettant de construire sa représentation mentale intégrée.*"

Compréhension de texte En compréhension de texte, Martins et Le Bouëdec (1998) proposent une définition proche des travaux de Wagener-Wender et Wender (1990) et Calvo et al. (2001). Toute information, non explicite, construite mentalement par le lecteur pour mieux comprendre le texte est considérée comme une inférence. Les inférences constituent également un processus d'interprétation, essentiel pour la compréhension du discours, dans la mesure où elles mettent en évidence des relations qui ne sont pas directement accessibles (Fayol 2003). La notion d'inférence a été, par la ensuite, affinée et développée en prenant des connaissances personnelles des lecteurs. Gombert et al. (1992) considèrent que l'accès au sens ne provient pas directement du texte, mais qu'il est construit par le lecteur, donc variable selon les individus en fonction de leurs connaissances.

De même, Kispal (2008) indique que la compréhension des inférences est facilitée si le lecteur dispose de larges connaissances et qu'il partage le contexte culturel du texte. Dans les dialogues quotidiens, Beaupré (2009) estime qu'il n'existe aucune loi absolue, mais que les inférences reposent sur un processus de généralisation et de règles. De manière générale, "la notion d'inférence renvoie aux informations que le lecteur ajoute au contenu explicite du texte pour le comprendre" (Rossi et Campion 1999). Les informations peuvent provenir de l'intérieur ou de l'extérieur du texte. Ces définitions fournissent une partie des règles de classification dans la suite de ce chapitre.

Il est à noter que la définition du concept d'inférence ne varie pas, mais que les travaux scientifiques éclairent les raisonnements qui sont à l'œuvre dans la production et la compréhension des inférences. Ces travaux complètent donc la définition de l'inférence qui est proposée dans les dictionnaires destinés au grand public.

Logique formelle L'étude des inférences en logique formelle repose davantage sur les raisonnements mis en œuvre. Le concept de raisonnement concerne les liens entre prémisses et conclusions.

La logique formelle admet la vérité de propositions (prémisses) et que l'on tire de vraies conclusions si les prémisses sont vraies. Parfois, des inférences des prémisses aux conclusions sont nécessaires (Bouquiaux et Leclercq 2017). Dans une approche déductive, ces règles permettent d'identifier la vérité d'une proposition à partir d'une ou plusieurs propositions prises en entrée. Dufaye (2001) considère que l'opération d'inférence consiste à poser un contenu non vérifié en s'appuyant sur un contenu vérifié ou supposé vérifié.

Logique classique et inférence langagière Martin (2004) propose des critères pour distinguer les inférences langagières des inférences logiques :

- L'inférence langagière est **non monotone**. De *x est un oiseau* peut être inféré *x vole*, puisque *l'oiseau vole* est généralement vérifié. Mais les exceptions existent toujours : la poule ne vole pas ni l'autruche.
- L'inférence langagière est **présuppositionnelle**. Elle est contestable dans le dialogue. Deux propositions *Pierre veut divorcer* et *Pierre ne veut pas divorce* infèrent toutes que *Pierre est marié*. Autrement dit, une proposition *p* et *non-p* induisent que *q* est vrai (*Pierre est marié*). Si l'interlocuteur sait que Pierre n'est en fait pas marié, ce *q* n'est pas nécessairement vrai, cette inférence ne valide pas non plus dans le dialogue. Cela montre que l'inférence classique n'y suffit pas.
- L'inférence langagière est également **conversationnelle**. Supposons que je demande à mon camarade s'il a cours aujourd'hui, il a plus de possibilités de me répondre en listant tous les cours qu'il aura, au lieu de mentionner qu'un seul cours. L'assertion langagière prétend non seulement la vérité, mais aussi toute la vérité. (Grice 1975) appelle "maxime conversationnelle" cette assertion.

Sous-entendus La notion des inférences doit également être analysée dans le cadre comparatif des travaux concernant les sous-entendus.

Dans les travaux de Ducrot (1969), il souligne que les données derrière un énoncé ne concernent pas seulement l'énoncé lui-même, mais aussi les "multiples occurrences possibles de cet énoncé dans les diverses situations" où les gens l'utilise. Et il n'est pas possible d'interpréter cet énoncé hors contexte. Ce contexte peut être affiné dans la mesure où "je comprends une langue, je suis capable d'attribuer une signification, et par suite de trouver des synonymes, aux énoncés prononcés hic et nunc". Même si la notion d'inférence n'a pas été mentionnée dans cet ouvrage, l'interprétation d'un sous-entendu par un humain nécessite des inférences.

Implicite Partant de l'analyse du contenu globale de tout énoncé Bres (2013) propose une structure en profonde de l'implicite : un contenu propositionnel et une valeur illocu-

toire. Chacun de ces constituants peuvent être le lieu de l'implicite. Le contenu propositionnel est sous la forme de l'inférence, tandis que la valeur illocutoire sous la forme d'illocutoire dérivé. L'interaction toujours au risque du malentendu, induit notamment par les différences de compétence entre des locuteurs et des récepteurs et la construction des inférences.

Dans la suite de notre travail, nous ne nous intéressons qu'aux inférences produites par les locuteurs d'une langue, dans une perspective de communication. Nous excluons donc les inférences utilisées en logique formelle pour nous concentrer sur les inférences langagières. Dans cette section suivante, nous nous intéressons aux différentes modalités d'analyse des inférences langagières.

2.1.3 *Modalité d'analyse des inférences*

Dans cette section, nous nous intéressons aux différentes modalités d'analyse des inférences langagières.

2.1.3.1 Les inférences en fonction de la façon de comprendre le texte

Selon les travaux de Rossi et Campion (1999), il existe trois axes de recherches autour de l'inférence au fil du temps. Au cours des années 1970, les inférences étudiées ne reposaient que sur la compétence linguistique : les présuppositions et les implications linguistiques (Marcel Adam et Herbert 1973).

Un deuxième courant de recherches s'intéresse aux dialogues. Le principe est d'étudier "les contraintes réagissant à la situation de communication et à la reconnaissance de l'intention du locuteur" qui mettent en lumière l'inférence pragmatique (Rossi et Campion 1999).

Un troisième courant étudie les inférences logiques du texte qui prennent en compte les compétences logiques des locuteurs qui ne concernent pas directement la compréhension du texte, mais davantage la rationalité humaine. La dernière approche consiste à inventaire et identifier toutes les inférences que les lecteurs vont potentiellement rencontrer. Le problème central est de déterminer quel type d'inférence est le plus effectif lors de la compréhension du texte. Le choix d'un type d'inférence dépend des connaissances qui doivent être mobilisées. Le point essentiel de cette approche considère que la génération de l'inférence a l'accessibilité des connaissances du monde de référence du texte (Rossi et Campion 1999).

2.1.3.2 Les inférences en fonction de l'utilisation des informations en mémoire à long terme

Avant d'entrer dans le détail des typologies d'inférence, il est nécessaire de préciser ce que sont la mémoire à long terme et la mémoire de travail. La génération de l'inférence nécessite souvent les connaissances du monde (Rossi et Campion 1999), car la

compréhension est simulée par un système de production dont les règles opèrent à différents niveaux : soit les résultats sont proposés à partir de l'information linguistique fournie par le texte ; soit on retrouve des connaissances en mémoire à long terme, associées au contenu du texte. Le dernier permet de produire les inférences (Denhière et Baudet 1992). Quant à la mémoire de travail, ell désigne les connaissances activées par ce contexte actuel.

Kintsch (1998) propose une classification avec 4 catégories. Nous observons que ces quatre catégories ne sont pas nommées, mais simplement désignées par des lettres de A à D. Cette désignation souligne la difficulté de nommer les types d'inférences :

- la catégorie A : quelques éléments du texte activent des informations en mémoire à long terme qui sont disponible en mémoire de travail. Ces informations s'ajoutent à celles tirées du texte pour comprendre le texte.
- la catégorie B : aucune information n'est automatiquement activée. La compréhension du texte demande une recherche active en mémoire en long terme pour établir une chaîne causale.
- la catégorie C : les informations à l'intérieur du texte construisent automatiquement une représentation mentale pour générer des informations non explicites dans le texte, sans l'intervention de connaissances extérieures au texte.
- la catégorie D : est la seule catégorie qui fait appel à un raisonnement logique et conscient, dite l'inférence logique.

Meredith et al. (2010) mettent en évidence un lien avec le développement de la mémoire de travail comme facteur explicatif important lors de la production d'inférences.

2.1.3.3 Les inférences en fonction de la direction entre deux propositions

N'ayant pas pris en compte les interventions d'informations en mémoire ni la production d'informations nouvelles, Van den Broeck classe les inférences en fonction de la direction de texte. Par rapport à la catégorisation qui contient seulement des inférences rétrogrades de Kintsch Walter, Rossi et Campion (1999) impliquent que celles de Van den Broeck peuvent être rétrogrades et antérogrades. L'inférence rétrograde relie un événement focal à un événement antérieur suite à une relation causale avec une recherche de mémoire(Van den Broek et al. 1999). Elle permet de mettre en relation deux éléments lors de l'interprétation de la seconde information (Fayol 2003). Par exemple, dans la phrase "Maman a préparé deux jupes pour Julie : une rouge et une verte. Julie n'aime pas le rouge". L'interprétation du résultat est que "Julie allait choisir la jupe verte". Cette interprétation mettait les deux phrases en relation.

Van den Broek et al. (1999) considèrent qu'il existe trois types d'inférences rétrogrades :

- l'inférence de connexion, "dans laquelle un événement 1 , lu en premier, remplit les critères de causalité par rapport à un événement 2 lu ensuite" (Rossi et Campion 1999). Un lien inférentiel est ainsi créé. Ceci correspond à la catégorie C de Kintsch (1998).

- l'inférence de restauration. Comme l'événement 1 ne remplit pas les critères de causalité, il semble y avoir une rupture de cohérence. Le lecteur doit chercher les informations dans sa mémoire. Cela correspond à la catégorie A.
- l'inférence d'élaboration, dans laquelle l'information causale n'est pas explicite dans le texte. Le lecteur doit utiliser sa connaissance du monde. Ces inférences correspondent à la catégorie B.

En revanche, les inférences antérogrades permettent au lecteur d'anticiper ses hypothèses sur ce qui va se passer par la suite. Par exemple, "C'était une expérience horrible. J'ai pu entendre la dispute de la chambre de gauche toute la nuit. De plus, le bar ne s'est pas terminé vers 2 heures du matin." On suppose que la phrase est un commentaire négatif d'un hébergement.

Il distingue deux types d'inférences antérogrades :

- l'inférence anticipatrice est spécifique. Elle se produit si l'information donnée est suffisante pour l'occurrence d'une conséquence.
- l'inférence prédictive est générale. L'information pour satisfaire une causalité se retrouve dans la suite du texte.

2.1.3.4 Les inférences en fonction du mécanisme mental

Dufaye (2001) se fonde sur la théorie du sens élaborée par Peirce (1958) pour proposer une distinction des inférences basée sur le mécanisme mis en œuvre mentalement : déduction, induction, et rétroduction. Ces trois types sont considérés par Peirce comme les trois figures du syllogisme (Deledalle et Peirce 1994). Pour parvenir à une conclusion valide, la déduction suppose des prémisses valides alors que l'induction se fonde sur des probabilités. Selon le nombre de prémisses, la déduction peut être classifiée en :

1. l'inférence immédiate qui ne possède qu'une prémisse et que la conclusion est atteinte via cette prémisse
2. l'inférence médiale qui possède au moins 2 prémisses (Khemlani et al. 2012).

Les inférences rétroductives nécessitent la prise en compte des connaissances antérieures (Deledalle et Peirce 1994).

Duchêne (2008) distingue les inférences logiques des inférences pragmatiques. Les inférences logiques reposent sur un raisonnement formel et mettent en œuvre un processus logique tandis que les inférences pragmatiques sont basées sur un raisonnement inductif et s'appuient sur l'ensemble des connaissances acquises par un individu au cours de ses expériences passées. Par exemple, il faut savoir que le Champ-de-Mars est à côté de la tour Eiffel pour conclure positivement au commentaire "*L'hôtel est à 5 min à pieds du Champ-de-Mars*".

2.1.3.5 Les inférences en fonction de la structuration sémantique des textes

Doucy et Massoussi (2012) opèrent une distinction en fonction du niveau d'analyse sémantique des textes. Ils distinguent ainsi les inférences lexicales (la phrase en dehors

de tout cadre énonciatif), les inférences énonciatives (un énoncé actualisé en contexte) et les inférences discursives (la séquence cohérente de phrases). Les inférences lexicales et énonciatives s'inscrivent dans un continuum : les inférences lexicales construisent le sens à partir des structures prédicatives (prédicats et arguments) et les inférences énonciatives se fondent sur le sens ainsi construit pour le placer en situation énonciative. Doucy et Massoussi (2012) soulignent également le fait que des connaissances extérieures au texte permettent de moduler le sens résultant d'inférences lexicales en apportant de nouvelles significations.

2.1.4 *Classification des types d'inférences*

Nous inscrivons notre travail dans le cadre de l'étude des inférences langagières, et nous nous inspirons des travaux précédents pour proposer une classification des inférences en cinq types. Plus particulièrement, nous nous intéressons aux inférences langagières du point de vue de la réalisation sémantique d'une part, et du point de vue du mode de production d'autre part. Nous détaillons ci-dessous les cinq types d'inférences auxquels nous nous intéressons dans ce travail.

2.1.4.1 Réalisation sémantique

La réalisation sémantique s'intéresse à l'accès au sens exprimé dans l'inférence. Pour cela, nous distinguons deux types d'inférences : les inférences logiques et les inférences pragmatiques.

Les inférences logiques se réalisent en découlant du texte. Elles reposent sur un raisonnement formel et mettent en œuvre un processus logique qui contient induction ou déduction. Par exemple, "房间整理及时" (la pièce est nettoyée au moment opportun). Le mot "及时" (au moment opportun) infère que le client est satisfait du service de ménage. Nous obtenons cette conclusion par le transfert entre des deux unités lexicales synonymes de "en temps opportun" et "bon moment". Cette phrase contient donc une inférence logique. L'induction correspond à un processus qui permet de passer du particulier (faits observés, cas singuliers, données expérimentales, situations) au général (une loi, une théorie, une connaissance générale). La déduction correspond au processus presque inverse qui permet de conclure (déduire) une affirmation à partir d'hypothèses, de prémisses ou d'un cadre théorique : les conclusions résultent formellement de ces prémisses ou de cette théorie. Pour l'annotation on ne distingue pas entre les deux.

Les inférences pragmatiques sont possiblement vraies et communes à l'ensemble des lecteurs. Une inférence sera pragmatique si le lecteur moyen (comparé à son groupe d'appartenance) a tendance à la donner après avoir été incitée. Elles s'appuient sur l'ensemble des connaissances acquises par un individu lors de ses expériences passées (Duchêne 2008). Par exemple, dans un corpus de commentaires touristiques, les séquences "没有烧水壶" (pas de bouilloire) et "离老佛爷近" (proche des Galeries Lafayette) contiennent des inférences pragmatiques, car les deux exemples reflètent les habitudes culturelles

chinoises (la bouilloire pour préparer le thé et les habitudes d'achat).

2.1.4.2 Mode de production

Le mode de production renvoie à la manière dont l'émetteur du message a produit l'inférence. Nous reprenons la distinction réalisée par Doucy et Massoussi (2012) sur la sémantique textuelle pour proposer les inférences lexicales, énonciatives et discursives.

L'inférence lexicale est la phrase en dehors de tout cadre énonciatif et dépend exclusivement de l'articulation entre les prédicats et leurs arguments. Le cadre minimal pour analyse des inférences lexicales est la phrase élémentaire, définie comme une relation entre un prédicat de premier ordre et ses arguments. C'est cette relation qui permet d'interpréter les unités lexicales, et d'expliciter toutes les inférences qui s'y rattachent. Cela signifie, d'une part, que les propriétés sémantiques sont subordonnées à la syntaxe, et, d'autre part, que les unités lexicales sont étudiées strictement du point de vue de leurs propriétés linguistiques. Les inférences lexicales construisent du sens à partir de structures prédicatives (prédicats et arguments).

- Unité lexicale spécifique à l'hôtellerie relative à un sentiment ou à une opinion : si un mot qui n'est pas un mot de sentiment ou d'opinion, mais qui porte une polarité dans le domaine de l'hôtellerie, nous le considérons comme une inférence lexicale. Par exemple, "aération", "anglais" ou "cafard".
- Absence de connecteur reliant deux phrases simples (la parataxe). Cette absence est palliée par l'inférence. Très souvent, les connecteurs de cause sont absents et l'inférence prend le relais. Par exemple, "Je n'arrive pas à me connecter. Mon téléphone est parfaitement fonctionnel". Un "mais" implicite entre les deux propositions infère un connecteur de cause.

L'inférence énonciative est actualisée en contexte. Les inférences énonciatives se fondent sur le sens ainsi construit pour l'inscrire dans une situation énonciative. C'est-à-dire que le cadre d'une inférence est un énoncé au lieu d'un mot ou d'un entraînement des énoncés. Plusieurs mécanismes sont en jeu tels que la négation, la polyphonie, le savoir partagé, etc.

- Dans les interactions clients, l'interprétation d'un dysfonctionnement ou d'un événement négatif dans le parcours du client s'explique par cette connaissance partagés. Celle-ci repose très souvent sur une comparaison implicite entre une situation "normale" attendue et une situation présentée comme "perturbée" par le locuteur.
  1. Des adverbes déterminant la polarité d'une phrase : "入住三天，只有第一天提供了瓶装水" (Sur les trois jours, les bouteilles d'eau étaient offertes seulement le premier jour) implique la situation attendue est « chaque jour il y a de nouvelles bouteilles d'eau offertes ».
  2. Des phrases interrogatives : "捡到东西难道不主动联系客人?" (Ne devriez-vous pas contacter immédiatement les clients en cas d'objets trouvés ?)
  3. Des déterminants d'occurrences multiples et de fraction : "找前台要了两次吹风机无果" (J'ai demandé deux fois à l'accueil pour avoir un sèche-

cheveux, pas de résultat) ; "客房价格仅是旺季的三分之一" (Le prix des chambres est le tiers de celui appliqué en haute saison).

- Reproche ou appréciation implicite : 敲门敲了一下，还没来得及应门，员工就进来了。(Le personnel a frappé à la porte une seule fois et je n' ai même pas eu le temps de répondre, la personne est entrée.) Cette phrase ne mentionne pas directement que la qualité de service n'est pas bonne, mais elle implique un reproche du personnel de l'hôtel.

L'inférence discursive interfère avec des connaissances extralinguistiques au niveau du discours. Le niveau d'analyse de ces inférences n'est plus la phrase ni l'énoncé, mais ce que l'on peut appeler des situations et qui se définissent comme la conjonction de paramètres linguistiques directement extraits du texte et issues d'expériences personnelles. L'interprétation et la mise en situation font appel à des inférences linguistiques auxquelles s'ajoutent des connaissances extérieures au texte. Ces connaissances permettent notamment de moduler les premières inférences linguistiques et de leur ajouter de nouvelles significations. Dans notre cas, si les mots clés d'un thème ne sont pas explicitement mentionnés dans un commentaire et que l'implication du thème vient du contexte, nous le considérons comme ayant une inférence discursive. Par exemple, d'après le contenu de la phrase "酒店送了一小盒巧克力给女儿" (L'hôtel a offert à ma fille une boîte de chocolat), nous résumons que l'attitude du personnel est bonne.

2.2 La fouille d'opinion

2.2.1 *Introduction*

Dans cette section, nous nous intéressons à la problématique de la fouille d'opinion que nous envisageons de lier, dans notre travail, à celle des inférences.

2.2.2 *Fouille d'opinion en général*

Selon la définition de S.-M. Kim et Hovy (2004) et Pang et Lee (2008), une opinion contient un sujet (Topic), un porteur d'opinion (Holder), une revendication (Claim) et un sentiment (Sentiment). Le porteur exprime une revendication sur un sujet. Le but de la fouille d'opinion est de définir automatiquement chaque élément et les relations entre eux. Sur la base de cette relation, S.-M. Kim et Hovy (2006) proposent quatre sous-tâches : l'identification des sujets, l'identification des porteurs, la délimitation des revendications et l'analyse des sentiments. Elles peuvent être synthétisées par 2 aspects :

- L'extraction des informations sur tous les éléments. Les sujets et porteurs peuvent être extraits par les algorithmes ad hoc (S.-J. Wu et Chiang 2015), les caractéristiques des compositions des phrases et leurs positions (Yi et Niblack 2005) , les entités nommées (S.-M. Kim et Hovy 2004) ou la cooccurrence des indications fréquentes et non-fréquentes (Hu et B. Liu 2004). L'enjeu est alors dans l'extraction d'information subjective et la catégoriser la polarité qui est soit binaire (positif

ou négatif), soit multiclasse (mauvais/bon/excellent) (Eensoo et Valette 2012).

– 2) L'analyse des sentiments sur les éléments extraits. Cette deuxième tâche nécessite une compréhension approfondie des règles langagières explicites et implicites, régulières et irrégulières, syntaxiques et sémantiques (Cambria et al. 2013). Dans les travaux existants, ils distinguent deux approches principales pour l'analyse de la polarité : premièrement les apprentissages automatiques basés sur des corpus d'entraînement annotés et deuxièmement, des stratégies basées sur des lexiques de polarité (Rinaldi et Musdholifah 2017). Au niveau de l'apprentissage automatique, la catégorisation des sentiments peut être considérée comme une classification binaire (Pang, Lee et Vaithyanathan 2002 ; Dave et al. 2003). Cette classification peut être réalisée par plusieurs méthodes d'apprentissage automatique telles que le bayésien naïf, l'entropie maximale, ou les machines à vecteurs de support (SVM), parmi lesquels les SVM ont obtenu d'excellents résultats (Saleh et al. 2011). La principale difficulté dans l'utilisation des SVM concerne le choix des paramètres appropriés (Basari et al. 2013). Les SVM fonctionnent bien même avec un grand nombre de caractéristiques et un petit corpus d'entraînement réparti (Varghese et Jayasree 2013 ; Joshi et Papola 2017). Les méthodes d'apprentissage varient également en fonction de la granularité des éléments : l'orientation sémantique du mot ou de l'expression (Samha et al. 2014 ; Hu et B. Liu 2004 ; Du et S. Tan 2009), des phrases (Lou et T.-F. Yao 2006) et des documents (Pang, Lee et Vaithyanathan 2002). Une tendance des méthodes est d'intégrer dans l'apprentissage automatique et l'analyse linguistique du corpus, comme les n-grammes, le nombre de négations, la similarité cosinus, caractéristiques extraites et représentation textuelle variée (Rinaldi et Musdholifah 2017 ; Joshi et Papola 2017). Les résultats illustrent qu'une classification hybride peut améliorer l'efficacité de la catégorisation des polarités dans plusieurs domaines (Prabowo et Thelwall 2009 ; Basari et al. 2013 ; Saleh et al. 2011 ; Itkes et Mashal 2015).

De nombreux travaux ont été récemment entrepris dans le domaine de l'exploration d'opinion ces dernières années. À l'instar du développement des méthodes de fouille d'opinion générales, la fouille d'opinion en chinois suit un chemin similaire : des méthodes du dictionnaire ou des règles, en passant par les méthodes statistiques aux méthodes automatiques (Prabowo et Thelwall 2009 ; T.-f. Yao et al. 2008 ; Han et al. 2018). Cependant, ces méthodes ont leurs limites. Cambria et al. (2013) regroupent les approches d'analyse des sentiments existantes en quatre catégories principales : repérage de mots-clés, affinité lexicale, méthodes statistiques et techniques basées sur des concepts. Les méthodes des mots-clés se concentrent sur la propagation de la valence des mots dont la polarité est connue à des termes qui coexistent avec eux dans un texte général ou dans des dictionnaires. Cette méthode est faible à deux égards : elle ne peut pas reconnaître de manière fiable la négation et elle repose sur les caractéristiques de la surface. Cambria et al. (2013) indiquent également que les méthodes statistiques comme les SVM sont sémantiquement faibles. Par conséquent, nous considérons les vecteurs du plongement lexicaux comme faisant partie des caractéristiques de notre modèle.

2.2.3 *Inférence dans la fouille d'opinion*

Parmi les tâches de la fouille d'opinion, nos travaux s'inscrivent dans le traitement des inférences dans les opinions implicites.

Il existe plein d'outils qui analysent les opinions sur les produits, par exemple SenticNet, Factiva, Attensity et Converseon. Cependant, la plupart des outils qui reposent sur des textes explicites pour classer la polarité et les émotions sont incapables de capturer les opinions implicites (Cambria et al. 2013).

S. Li et al. (2017) a également fait valoir que si les commentaires ne contiennent aucun mot de sentiments explicites, mais expriment des émotions, leur méthode n'est pas capable de calculer le score de polarité. S. Li et al. (2017) donnent un exemple concret : 白天叫服务人员来打扫卫生, 一直也没见到人. 就凭这一点, 以后再也不会住该酒店 ! (J'ai appelé le personnel de service pour nettoyer dans la journée, mais je n'ai jamais vu personne. Sur ce point, je ne reviendrai plus jamais dans cet hôtel !) Comme défini dans le chapitre précédent, cet exemple est typiquement un commentaire avec des inférences.

Yang et Cardie (2013) présentent un modèle qui identifie conjointement les entités liées à l'opinion, y compris les expressions d'opinion, les cibles d'opinion et les détenteurs d'opinion ainsi que les relations de liaison des opinions associées. L'inférence est utilisée pour déterminer le détenteur ou la cible dans le cas où ces éléments ne sont pas explicitement exprimés dans le texte. Cependant, l'inférence dans cet article ne concerne pas la compréhension des expressions d'opinions. Cela se distingue de notre travail.

La méthode de Turney et Littman (2003) consiste à inférer l'orientation sémantique d'un mot à partir de son association statistique avec un ensemble de mots de paradigme positifs et négatifs. Pourtant, le concept de l'inférence dans ce travail ne concerne pas l'inférence linguistique.

Comme présentée dans le chapitre précédent, le contexte culturel est une dimension négligeable dans l'inférence pragmatique, car les opinions exprimées par les humains sont interprétées différemment selon les cultures. Rubin (2014) suggère aussi que la linguistique informatique s'appuie sur une recherche interdisciplinaire plus large sur les différences culturelles et l'utilisation pragmatique de la langue dans les cultures asiatiques sur la détection automatisée de la tromperie dans les langues asiatiques.

Lorsque le message d'un locuteur dépasse le sens littéral ou logique des phrases utilisées, une inférence pragmatique est nécessaire pour comprendre le sens complet d'un énoncé. Spotorno et al. (2015) étudient une inférence pragmatique dans l'utilisation de "some" (certain). Par rapport à "some", le locuteur a vraisemblablement une raison de ne pas utiliser un terme plus fort comme "all" (tous).

2.2.4 *Implicite dans la fouille d'opinion*

Les opinions sont généralement des expressions subjectives qui décrivent les sentiments, les évaluations ou les sentiments des personnes envers des entités, des événe-

ments et leurs propriétés (B. Liu 2010). Parmi les opinions subjectives, une partie des opinions sont implicitement exprimées.

D'abord, nous distinguons la notion de "l'implicature" et "l'implication" par la définition de Davis (2019). "L'implicature" désigne soit l'acte de signifier ou d'impliquer une chose en disant autre chose, soit l'objet de cet acte. Les implications peuvent être déterminées par le sens de la phrase ou par le contexte conversationnel conventionnelles ou non conventionnelles, par exemple la métaphore et l'ironie. L'implicature sert une variété d'objectifs : communication, maintien de bonnes relations sociales, tromperie sans mensonge, style et efficacité verbale. La connaissance des formes courantes d'implicature s'acquiert en même temps que sa langue maternelle. S'appuyant sur l'implicature conversationnelle de Grice, T. Wilson et al. (2005) considèrent l'évaluation implicite comme des implicatures d'opinion, qui sont "les inférences par défaut qui peuvent ne pas passer en contexte".

Les opinions sont classifiées en explicite et implicite selon la subjectivité et l'objectivité. La polarité et le thème des opinions explicites sont plus faciles à traiter car ces opinions contiennent des mots explicites (B. Liu 2010 ; B. Liu 2012). Pour des opinions implicites, nous devons l'inférer à partir des énoncés. Une implicature peut être caractérisée comme une inférence, mais l'acte ' impliquer" n'est pas pareil à l'acte "inférer". Impliquer quelque chose, c'est exprimer une opinion d'une manière particulière. Inférer quelque chose, c'est acquérir ou posséder une opinion d'une manière particulière. L'implicature est indirecte car impliquer quelque chose, c'est le signifier en disant autre chose. Même si cela nécessite une inférence, notre reconnaissance de ce qui est signifié est généralement automatique et sans effort. Tous les actes de langage doivent être inférés à partir de preuves contextuelles, y compris ce qui a été dit et quels mots ont été prononcés. La question de savoir s'il existe une différence significative dans le type d'inférence requise pour reconnaître une implicature fait l'objet d'un débat et peut dépendre du type d'implicature (Recanati 2002 ; D. Wilson et Sperber 2012 ; Levinson 2016 ; Simonin 2018 ; Davis 2019).

Les études autour des opinions implicites sont basées sur l'extraction des caractéristiques. H. Liu et Yu (2005) et Su et al. (2008) définissent des caractéristiques implicites qui ne se produisent pas explicitement, mais peuvent être déduite par des mots d'opinion qui l'entourent. H.-Y. Chen et H.-H. Chen (2016b) traite à la fois la polarité implicite et le thème implicite. Ils observent qu'une opinion implicite et son opinion explicite voisine ont tendance à avoir le même aspect et la même polarité. Mais en raison de l'ambiguïté causée par des mots d'opinion communs qui sont souvent exprimés sur divers caractéristiques, la performance est insatisfaisante. Par conséquent, Lazhar (2019) crée un ensemble de règles d'association regroupant des couples caractéristiques-opinion explicites et ensuite utiliser cet ensemble pour construire un modèle de classification capable de prédire pour chaque ensemble donné de mots d'opinion la cible appropriée.

2.3 Conclusion

Dans ce chapitre, nous avons introduit l'état de l'art concernant l'inférence et la fouille d'opinion.

Même s'il n'existe pas de consensus omnium sur la définition de l'inférence, nous pouvons en extraire quelques caractéristiques en commun : l'inférence est une opération ou une action par laquelle on passe d'une assertion à une conclusion par un système de raisonnement. L'inférence est classifiée différemment selon les courants de recherche. Sur la base des classifications existantes et dans le cadre d'étude de la fouille d'opinion menées sur les avis d'utilisateurs chinois dans le domaine du tourisme, nous avons proposé une typologie d'inférences selon la modalité de réalisation avec cinq catégories : les inférences logiques, pragmatiques, lexicales, énonciatives et discursives.

Ensuite, nous avons d'abord brièvement présenté les principales tâches et approches de la fouille d'opinion. Ensuite, nous avons introduit les travaux qui concernent à la fois un aspect du traitement automatique des langues et des inférences. Cependant, il n'existe pas assez de recherches combinant directement l'inférence et la fouille d'opinion.

Nous nous intéressons également aux travaux des opinions implicites, puisque les méthodes d'extraction des caractéristiques des opinions implicites peuvent aussi adaptées pour caractériser les opinions avec des inférences.

Pour mieux se situer dans la recherche, notre travail est défini comme une intégration d'inférences linguistiques dans la fouille d'opinion, en particulier la détection de polarité et de thème, dans le but de catégoriser et de modéliser les caractéristiques inférentielles des expressions dans l'hôtellerie.

Deuxième partie

Présentation et construction des ressources linguistiques

Chapitre 3

Corpus d'avis touristiques

Contents

3.1 Introduction 25
3.2 Identification des sites 25
3.2.1 Origine du corpus 25
3.2.2 Informations disponibles 26
3.3 Critères de sélection des messages 27
3.4 Extraction automatique du corpus 27
3.5 Prétraitement du corpus 28
3.5.1 Segmentation 28
3.5.2 Transformation du chinois traditionnel vers chinois simplifié 30
3.5.3 Standardisation des entités nommées 31
3.5.4 Partie du discours 31
3.6 Conclusion 34

3.1 Introduction

Dans cette deuxième partie, nous présentons les ressources linguistiques et les corpus que nous avons constitués, les traitements appliqués à ces corpus, et les annotations réalisées sur ces corpus.

3.2 Identification des sites

3.2.1 *Origine du corpus*

Dans ce travail, nous nous intéressons aux inférences utilisées en chinois, à des fin de fouille d'opinion. A cet effet, nous avons rassemblé un corpus de commentaires postés

sur deux sites par des touristes chinois en visite à Paris, sur la thématique de l'hébergement à Paris (en hôtel ou chez l'habitant) : Booking[1] et Mafengwo[2]. Les deux sites fournissent tous une plateforme permettant aux utilisateurs de partager leurs propres expériences de séjour dans des hôtels de Paris. Le premier représente une plateforme internationale et le second une plateforme locale.

Booking est utilisé par des utilisateurs internationaux, dans différentes langues, tandis que Mafengwo est un site chinois, utilisé par des internautes sinophones (Chine continentale, Hong Kong, Taiwan, etc). Il existe plein d'autres sites internationaux qui fournissent des services similaires comme Booking, par exemple Airbnb et TripAdvisor. Les raisons pour lesquelles nous n'avons pas choisi ces deux sites sont les suivantes :

- les messages automatiquement traduits. TripAdvisor permet aux volontaires de traduire des messages dans d'autres langues. Un commentaire en chinois peut être la source d'un commentaire français. Ce genre de messages traduits ne représentent pas les points de vue de la communauté chinoise, en particulier en terme de traditions culturelles. De plus, selon des expressions et des vocabulaires peu naturels, il est évident que les messages traduits ne sont pas rédigés par un natif.
- la validation des messages par les propriétaires de structures touristiques. Quant à l'Airbnb, tout propriétaire a la possibilité de vérifier un commentaire avant d'autoriser un utilisateur à le publier sur le site. Il est possible qu'une partie des commentaires négatifs soient contrôlés. Même selon l'observation, les commentaires sur Airbnb sont majoritairement positifs. Ce genre de corpus apportera un déséquilibre entre des opinions positives et négatives dans les échantillons. De plus, les commentaires ne sont pas subjectifs faute de réels avis négatives.

3.2.2 *Informations disponibles*

Les sites fournissent des informations complémentaires, d'une part autour de l'hôtel, d'autre part autour des utilisateurs. Les informations complémentaires sur hôtel concernent les coordonnées physiques et électroniques, le nombre d'étoiles et les services proposés. Les informations associées aux utilisateurs concernant l'âge, le score proposé, la date du commentaire. Ces informations complémentaires nous permettent de dresser une typologie des utilisateurs pour chaque site : les utilisateurs de Mafengwo sont plus jeunes. Les jeunes utilisateurs rédigent des blogs de voyage et postent des annonces pour des voyages de groupe. Ce phénomène est confirmé par la longueur moyenne des commentaires de ces deux sites : celle de Booking est deux fois plus courte que celle de Mafengwo. Les chiffres comparatifs de deux sites sont présentés dans le tableau 3.1.

Contrairement à Mafengwo qui ne propose qu'un seul espace pour publier un commentaire, Booking offre un espace pour les commentaires positifs et un autre pour les commentaires négatifs. Bien que tous les utilisateurs n'utilisent pas cette fonction, elle fournit déjà des informations sur la polarité.

1. `http://www.booking.com/`
2. `http://www.mafengwo.cn/`

Sites	N. d'hôtels	N. de commentaires	Longueur moyenne par commentaire	N. de mots
Booking	4457	12131	36,2	439 142
Mafengwo	2018	10910	82,7	902 257

Tab. 3.1 : Chiffres comparatifs de deux sites

De plus, les deux sites présentent un guide hôtelier avec une présentation générale, le règlement, ses points forts, les alentours de l'hôtel, les sites touristiques recommandés par les locaux autour de cet hôtel et le plan pour indiquer la localisation visuelle de cet hôtel.

Lors de l'extraction, nous avons conservé les commentaires et les informations complémentaires sur les hôtels et les utilisateurs.

3.3 Critères de sélection des messages

Nous avons limité les messages à ceux concernant Paris, en raison du caractère touristique de la ville. Nous estimons que cela constitue une propriété intéressante pour l'étude des inférences pragmatiques et lexicales. Par exemple, "La tour Eiffel est à 5 minutes à pieds de cet hôtel" contient une inférence pragmatique utilisant le site touristique de Paris.

En terme de langues, nous avons recueilli des commentaires majoritairement rédigés en chinois simplifié et traditionnel. Dans certains commentaires, il contient également des expressions anglaises, françaises, japonaises ou coréennes. Nous n'avons pas trié la langue lors de la collecte du corpus. Ce travail se fera dans le prétraitement du corpus dans la section *Prétraitement du corpus*.

La période couverte par les messages va de janvier 2016 à décembre 2017. Nous ne retrouvons pas d'événement touristique majeur sur la période couverte par notre corpus qui pourrait amener une catégorie de touristes plus spécifiques et moins représentatives du tourisme ordinaire.

3.4 Extraction automatique du corpus

En respectant les critères définis, nous avons réalisé un crawler en Python, adapté aux deux sites. Pour chaque commentaire, il enregistre d'abord le message lui-même et les informations associées que nous avons mentionnées dans la section *Informations disponibles des sites* .

Cette extraction automatique peut être divisée en deux étapes :

1. Il commence par trouver tous les hôtels correspondant aux critères de Booking et Mafengwo en créant un identifiant unique pour chaque hôtel. En même temps, le

crawler trie tous les hôtels en commun de ces deux sites et lui attribue un identifiant unique partagé. Les résultats de cette première étape renvoient deux listes d'URL d'hôtels.

2. Ensuite, en accédant via les URL, le crawler extrait toutes les commentaires de l'hôtel et les informations associées respectivement dans les deux sites. Étant donné qu'un hôtel peut avoir un ou plusieurs commentaires, les caractéristiques de l'hôtel peuvent être répétées, par exemple, le nom, l'étoile, le score et l'adresse.

Quant à la structure du site, l'une des différences entre Booking et Mafengwo est que les commentaires et informations sur le second sont directement accessibles via les codes sources de la page HTML. Une simple application du module BeautifulSoup de Python réussit à extraire toutes les informations nécessaires. Mais les commentaires de Booking ne sont pas affichés dans les codes sources de la page HTML. Ce sont des résultats de requêtes JavaScript. Dans notre script, nous avons simulé le fonctionnement de la requête afin d'obtenir des commentaires de chaque page.

Au total, le corpus contient 4 457 hôtels en Booking avec 12 131 commentaires d'une longueur moyenne de 36,224 1 caractères par commentaire et 2 018 hôtels sur Mafengwo avec 10 910 commentaires d'une longueur moyenne est 82,702 5 caractères par commentaire (Taleaux 3.1).

3.5 Prétraitement du corpus

En raison de la particularité de la langue chinoise et de la problématique spécifique de la recherche, les prétraitements du corpus sont nécessaires avant de passer à la partie suivante. Ces prétraitements consistent en une segmentation du chinois réalisé à l'aide d'un dictionnaire spécifique, une normalisation des entités nommées et un étiquetage automatique des parties du discours.

3.5.1 *Segmentation*

Puisqu'il n'y a pas de caractère d'espace entre chaque idéogramme, le chinois est une langue complexe à segmenter en mots. cependant, il existe plusieurs outils pour segmenter les textes. Nous présentons les outils les plus utilisés.

3.5.1.1 Outil : Jieba

Jieba est un outil de traitement automatique de la langue chinoise. Il est disponible gratuitement sur GitHub[3]. Ce logiciel propose trois modes de segmentation :

– Le **mode précis** tente de couper la phrase aussi précisément que possible, ce qui convient à l'analyse de texte. Ce mode offre une meilleure solution après avoir analysé les cas d'ambiguïtés sémantiques.

3. `https://github.com/fxsjy/jieba`

ZH : 酒店/ 离/ 埃菲尔铁塔/ 步行/ 10/ 分钟/ 。
FR : L'hôtel/ à / tour Eiffel/ à pied/ 10/ minutes/ . (traduction mot à mot)

– Le **mode complet** récupère tous les mots possibles de la phrase. C'est est rapide, mais cela ne tient pas compte de l'ambiguïté.
ZH : 酒店/ 离/ 埃菲尔/ 埃菲尔铁塔/ 菲尔[4]/ 铁塔/ 塔步/ 步行/ 10/ 分钟/ 。
FR : L'hôtel/ à/ Eiffel/ tour Eiffel/ ffel[5]/ tour/ inconnu*[6]/ à pied/ 10/ minutes/ . (traduction mot à mot)
– Le **mode moteur de recherche**, basé sur le mode précis, tente de diviser les mots longs en plusieurs mots courts, ce qui peut augmenter le taux de rappel. Il convient aux moteurs de recherche. Dans l'exemple, les résultats suggèrent 4 combinaisons provenant de "埃菲尔铁塔". Distingué du mode complet qui récupère tous les mots possibles au niveau de la phrase, 塔步 qui combine le dernier caractère de "埃菲尔铁塔" et le premier caractère de "步行" n'existe pas dans le mode de moteur de recherche, car ce dernier mode coupe un mot long dans les résultats du mode précis en plusieurs mots courts, mais pas entre les mots découpés.
ZH : 酒店/ 离/ 菲尔/ 铁塔/ 埃菲尔/ 埃菲尔铁塔/ 步行/ 10/ 分钟/ 。
FR : L'hôtel/ à/ ffel/ tour/ Eiffel/ Tour Eiffel/ à pied/ 10/ minuites/ . (traduction mot à mot)

Son algorithme est basé sur une structure de dictionnaire de préfixes pour réaliser une analyse efficace du graphe de mots. Il construit un graphe acyclique dirigé (DAG) pour toutes les combinaisons de mots possibles et utilise la programmation dynamique pour trouver la combinaison la plus probable en fonction de la fréquence des mots.

Ce logiciel n'est pas seulement un outil de segmentation, il propose également d'autres tâches de TAL comme l'étiquetage des parties du discours, l'extraction d'information, etc.

3.5.1.2 Ajout du dictionnaire personnalisé

Jieba permet également de personnaliser un dictionnaire. Pour détecter les mots inconnus, un modèle basé sur HMM est utilisé avec l'algorithme de Viterbi. De plus, les développeurs peuvent spécifier leur propre dictionnaire et l'intégrer dans le dictionnaire par défaut de jieba. Même si Jieba est capable d'identifier de nouveaux mots, l'ajout du dictionnaire personnel permet d'améliorer la précision et empêcher également les expressions figées d'être segmentées.

Pour les sites touristiques, au lieu de les segmenter en unités les plus petites, il vaux mieux garder le mot entier. Par exemple, '巴黎圣母院"(Notre Dame de Paris) peut être segmenté en deux mots '巴黎" (Paris) et "圣母院"(notre dame). Lors de l'évaluation de localisation d'un hôtel, les deux mots séparés ne représentent plus un lieu emblématique à Paris, de même pour les vocabulaires des transports et du tourisme. Si l'on ne précise

4. La couleur rouge indique l'endroit où il y a des erreurs de découpage.
5. 菲尔 n'est pas un mot, mais est détecté comme un mot par Jieba
6. 塔步 idem

pas,"穿梭巴士" (navette) est divisé en "穿梭"(faire la navette) et "巴士"(bus), ce qui n'a plus le sens de "navette". En conséquence, la personnalisation d'un dictionnaire dans Jieba améliore la qualité de la segmentation, au moins pour notre corpus spécifique.

Le corpus étant composé de commentaires d'hôtels à Paris, nous avons défini notre propre dictionnaire en conservant les mots spécifiques dans le domaine de l'hôtellerie de tourisme et des sites touristiques à Paris. Ce dictionnaire contient 7 listes de lexiques et 1 926 mots au total (voir Tableau 3.2).

Liste de lexique	Nombre de mots	Nombre de caractères	Longueur moyenne par mot
Tourisme	544	2550	4,6875
Transport	101	387	3,8316
Carrière	141	755	5,3546
Restauration	181	640	3,5359
Hotêllerie	798	3184	3,9899
Sites touristiques à Paris	46	220	4,7826
Sites touristiques mondiales	114	768	6,7368
Total	1925	8504	4,4176

Tab. 3.2 : Statistique du dictionnaire intégré dans la segmentation Jieba

3.5.2 *Transformation du chinois traditionnel vers chinois simplifié*

Le chinois simplifié (简体中文) est une écriture standardisée du chinois moderne, par opposition au chinois traditionnel(繁体中文). Le chinois simplifié est essentiellement composé de caractères hérités et de caractères simplifiés promus par le gouvernement de la République populaire de Chine à partir des années 1950. Le chinois simplifié est principalement utilisé dans les communautés chinoises de Chine continentale, de Malaisie, de Singapour et de certains pays d'Asie du Sud-Est. Par le passé, le chinois simplifié et le chinois traditionnel coexistaient dans divers documents des Nations Unies. Depuis que le gouvernement de la République populaire de Chine a remplacé les autorités taïwanaises et est revenu à l'ONU en 1971, le chinois simplifié est devenu l'une des six langues officielles de l'ONU.

Outre les différences d'écriture des caractères chinois, le chinois simplifié et traditionnel ont également des différences de vocabulaire. Par exemple, pour le "stylo à bille", "原子笔"(atome stylo) est souvent utilisé en chinois traditionnel et "圆珠笔" (perle ronde stylo) est utilisé à la place en chinois simplifié. En particulier au niveau des noms propres, l'utilisation du vocabulaire est très différente. En chinois simplifié, "Corée du Nord" s'écrit "朝鲜" (vers frais), tandis que le chinois traditionnel l'appelle "北韩"(nord Corée).

Les différences d'emploi du chinois traditionnel et simplifié varient selon les populations. Au milieu du XXe siècle, la Chine continentale et Taïwan avaient moins de communication, ce qui creusait la différence dans l'usage des mots. Ce phénomène est devenu plus évident depuis les années 1980 avec l'accélération des progrès technologiques et

scientifiques, ce qui a mis en évidence la différence de terminologie. Par exemple, pour la “souris” de l’ordinateur, ‘滑鼠”(glisser souris) est utilisé en chinois traditionnel et “鼠标” (souris marque) en chinois simplifié. En raison de cette différence, le terme “langage simplifié” a vu le jour.

En conséquence, le passage du chinois traditionnel au chinois simplifié est obligatoire avant l’analyse de texte afin d’unifier les vocabulaires. Cette tâche n’est pas seulement un changement de forme du caractère. Nous avons utilisé l’outil OpenCC[7] qui permet de transformer un texte mixte du chinois traditionnel et simplifié en une langue choisie. Nous avons finalement transformé tout le chinois traditionnel en chinois simplifié. Les encodages de caractères pris en charge sont tous en UTF-8.

3.5.3 *Standardisation des entités nommées*

Le corpus étant constitué de commentaires postés sur Internet, il ne respecte pas strictement les normes de vocabulaire, notamment les translittérations. Pour un nom propre, il existe plusieurs formes d’écriture dans notre corpus. Nous notons 11 translittérations différentes de “Champs Élysées” dans notre corpus (“香榭里舍大街”,“香榭里舍大道”,“香榭利舍大街”, etc).

Il est nécessaire d’unifier les variantes en une seule forme, car

1. Premièrement, Jieba ne peut pas segmenter correctement les mots qui ne sont pas enregistrés dans le dictionnaire interne de Jieba. Un nom propre translittéré correspond à une seule forme régulière. L’outil considérera les autres variants comme de nouveaux mots. Ce n’est pas le résultat que nous attendons de la segmentation.
2. Ensuite, les formes non unifiées apportent des erreurs lors de l’analyse statistique du texte. Même si les deux variantes se réfèrent à un seul emplacement, le système considère que l’un des deux est un mot non enregistré.

Au lieu de choisir une forme standard, nous prenons l’approche adoptée par L. Jiang et al. (2007) en utilisant la translittération en pinyin de mots étrangers et en supprimant les marqueurs de ton. Autrement dit, nous ignorons si les caractères utilisés sont corrects ou non et les réécrivons tous en caractères latins. Par exemple, pour “Champs Élysées”, il peut s’écrire “香榭里舍大街”,“香榭里舍大道”,“香榭利舍大街”,“香榭利舍大道”,“香街”,“香榭丽大街”,“香榭丽”,“香榭丽大道”. Après normalisation, il a une seule forme “xiang xie li she”.

3.5.4 *Partie du discours*

Les informations de la partie du discours aideront non seulement l’analyse textométrique du corpus, mais seront également utilisées comme caractéristiques du modèle d’apprentissage automatique dans l’expérimentation.

Notre corpus est sur les commentaires d’hôtels à Paris.Comparé à un corpus générique, il est plus sensible à l’ambiguïté lexicale. Mais il n’y a pas d’outil spécifique pour

7. `https://pypi.org/project/OpenCC/`

annoter un corpus hôtelier. Avant d'utiliser l'outil Jieba directement pour l'étiquetage automatique, nous voulons d'abord comparer la performance de l'étiquetage automatique et manuel afin de vérifier si un outil générique suffira pour un corpus spécifique prenant compte à la fois l'efficience et la performance.

Pour l'annotation manuelle, les catégories de la partie du discours sont conformes à la version simplifiée de "Chinese symbol part-of-speech tagset" [8]. Nous n'avons utilisé que les étiquettes simplifiées, soulignées en rouge dans le tableau 3.3, car cela représente déjà un jeu de 18 étiquettes et ces catégories sont suffisantes pour nos travaux sur les inférences et la fouille d'opinion démontrée par Sun et Wan (2016).

POS simplifiés	Abbréviation	Interprétation
Adjectif	A	Adjectif non prédicatif
Conjonction	Caa	Conjonction conjonctive，e.g. 和、跟
	Cab	Conjonction, e.g. 等等
	Cba	Conjonction, e.g. 的話
	Cbb	Conjonction corrélative
Adverbe	D	Adverbe
	Da	Adverbe quantitatif
的-Construction	DE	的, 之, 得, 地
Adverbe	Dfa	Adverbe pré-verbal de degré
	Dfb	Adverbe post-verbal de degré
	Di	Adverbe Aspectuel
	Dk	Adverbe sentimental
Mot étranger	FW	Mot étranger
Interjection	I	Interjection
Nom	Na	Nom commun
	Nb	Nom propre
	Nc	Nom de lieu
	Ncd	Localisateur
	Nd	Nom de temps
Déterminant	Nep	Déterminant démonstratif
	Neqa	Déterminant quantitatif
	Neqb	Déterminant post-quantitatif
	Nes	Déterminant spécifique
	Neu	Déterminant numérique
Mesure	Nf	Mesure
Postposition	Ng	Postposition
Pronom	Nh	Pronom
Préposition	P	Préposition
verbe 是	SHI	être 是

8. `https://www.sketchengine.co.uk/chinese-symbol-part-of-speech-tagset/`

Table 3.3 continued from previous page

POS simplifiés	Abbréviation	Interprétation
Particule	T	Particule
Verbe	VA	Verbe intransitif actif
	VAC	Verbe causal actif
	VB	Verbe pseudo-transitif actif
	VC	Verbe transitif actif
	VCL	Verbe actif avec un objet locatif
	VD	Verbe transitif
	VE	Verbe actif avec un objet sentimental
	VF	Verbe actif avec un objet verbal
	VG	Verbe classificatoire
	VH	Verbe intransitif statif
	VHC	Verbe causatif statif
	VI	Verbe pseudo-transitif statif
	VJ	Verbe transitif statif
	VK	Verbe statif avec un objet sentimental
	VL	Verbe statif avec un objet verbal
	V_2	avoir 有
Ponctuation	COMMACATEGORY	, virgule
	PERIODCATEGORY	. point
	PARENTHESISCATEGORY	“, 》, ’ , (parenthèse
	PAUSECATEGORY	、 pause
	COLONCATEGORY	: deux points
	QUESTIONCATEGORY	? point d’intérogation
	SEMICOLONCATEGORY	; point virgule
	ETCCATEGORY	…points de suspension
	EXCLANATIONCATEGORY	! point d’exclamation

Tab. 3.3 : Catégories et les interprétations de la partie du discours pour l’annotation manuelle

Afin de vérifier la pertinence de l’étiquetage dans les parties du discours proposé par Jieba sur un corpus provenant d’Internet, nous avons réalisé une annotation manuelle pour comparer les performances obtenues.

Les démarches se composent de quatre étapes :

- Étiquetage manuel d’un corpus par deux chinois natifs
- Étiquetage automatique du même corpus avec Jieba
- Évaluation des résultats en comparant les deux versions d’annotation sur la partie du discours

Dans un premier temps, nous avons sélectionné 70 commentaires et avons fait annoter ce corpus par deux natifs chinois selon les catégories présentées dans le tableau 3.3.

Ensuite, les deux annotateurs ont comparé et corrigé leurs propres résultats pour produire une version de consensus. Dans cette étape, nous avons aussi annoté les éléments d'une opinion : cible, opinion, lieu, polarité. Même si les étiquettes d'une opinion n'ont pas été utilisées dans l'évaluation, cela nous a donné une première observation linguistique de la fouille d'opinion. Les résultats statistiques sont présentés dans le tableau 3.4.

	Étiquettes morpho-syntaxiques	Temps moyen d'annotation
Natif 1	2226	4,76 heures/doc
Natif 2	1815	5,99 heures/doc
Adjudication	2427	1,66 heures/doc

Tab. 3.4 : Nombre d'étiquettes et temps d'annotation

Par la suite, nous avons effectué une annotation automatique des parties du discours avec l'outil Jieba. Enfin, nous avons comparé les résultats de Jieba (55 catégories) avec la version consensus des humaines (18 catégories) en calculant les scores de rappel, de précision et de F-mesure qui sont respectivement de 0,734, 0,815 et 0,746. Ces résultats n'étaient pas très satisfaisants. Toutefois, nous avons choisi la méthode automatique pour deux raisons : 1) L'étiquetage morpho-syntaxique est basé sur la segmentation. Il n'existe pas une segmentation standard, mais plutôt plusieurs solutions en fonction de la granularité, du contexte ou de l'interprétation. Même si certaines des solutions de Jieba sont différentes de celles des êtres humains, les jugements comme faux sont arbitraires. 2) La catégorisation des parties du discours de Jieba est plus complète que celle de l'annotation manuelle. Afin de permettre la comparaison entre les annotations automatiques et manuelles, nous avons homogénéisé les annotations autour de 15 catégories. Par exemple pour la catégorie du"nom", cet outil l'affine en 7 sous-catégories telles que les noms d'établissements, les noms de lieux, les noms propres, les morphèmes nominaux, les anciens noms personnels, les noms personnels modernes et les noms personnels translitérés. Si nous demandions aux annotateurs de distinguer ces 55 catégories, le travail serait très compliqué et le temps d'annotation serait beaucoup plus long.

De plus, l'utilisation d'un dictionnaire utilisateur dans Jieba permet également de s'assurer que les POS sur les mots étrangers sont corrects. Par exemple, la séquence en pinyin de "Champs Elysées" est correctement étiquetée.

3.6 Conclusion

Dans ce chapitre, les étapes que nous avons suivies sont :
- identification des sites
- définition des critères de sélection des messages
- récupération des données
- pré-traitements du corpus (segmentation, conversion du chinois traditionnel vers

simplifié, translittération des mots étrangers en pinyin et étiquetage des parties du discours)

En résumé, notre corpus est composé de deux parties principales : les textes et les métadonnées. Une comparaison statistique des données selon des étoiles est présentée dans le tableau 3.5. À partir de ce corpus normalisé, nous continuons une annotation manuelle dans le chapitre suivant.

Sites	Catégorie statistique	Étoile					
		0	1	2	3	4	5
Booking.com	Nombre d'hôtels	437	22	190	630	337	64
	Nombre de commentaires	5783	327	3435	11082	5686	813
	Nombre de tokens (moyen des commentaires)	135552 (23,44)	6774 (20,72)	70463 (20,51)	238840 (21,55)	126050 (22,17)	18615 (22,90)
	Nombre de mots (moyen des commentaire)	357593 (61,83)	18040 (55,16)	187066 (54,54)	634326 (57,2)	335134 (59,23)	49636 (61,05)
Mafengwo.cn	Nombre d'hôtels	88	452	254	737	417	69
	Nombre de commentaires	662	2314	3760	10548	5390	879
	Nombre de tokens (moyen des commentaires)	25263 (38,16)	93283 (40,31)	147447 (39,21)	423441 (40,14)	213252 (39,54)	37261 (42,39)
	Nombre de mots (moyen des commentaires)	68493 (103,46)	252716 (109,21)	399806 (106,33)	1146423 (108,69)	577280 (107,10)	101421 (115,38)

Tab. 3.5 : Données statistiques des hôtels dans chaque catégorie

Chapitre 4

Identification manuelle des inférences dans un corpus de commentaires

Contents

4.1 Introduction 37

4.2 Guide d'annotation 38

4.2.1 Objectifs 38

4.2.2 Annotation des inférences 38

4.2.3 Annotation des types d'inférences 40

4.2.4 Niveaux d'analyse 41

4.3 Processus d'annotation 42

4.3.1 Sélection des commentaires à annoter 42

4.3.2 Outil et configuration d'annotation 43

4.3.3 Accords inter-annotateurs 45

4.3.4 Difficultés rencontrées 47

4.4 Analyse des résultats d'annotation 48

4.4.1 Corrélation et distribution des catégories selon différents niveaux 48

4.4.2 Analyse textométrique du corpus 52

4.5 Conclusion 55

4.1 Introduction

Au chapitre 2.4, nous avons proposé une classification des inférences en cinq types : inférence logique, pragmatique, lexicale, énonciative et discursive. Sur la base des types d'inférence définis, dans ce chapitre, nous présentons d'abord une annotation manuelle sur un sous-corpus représentatif du corpus complet. Cette annotation consiste à étiqueter d'un côté la présence des inférences et leurs types, de l'autre côté les informations de la fouille d'opinion comme la polarité et le thème. Ensuite, nous faisons des analyses sur ce corpus d'annotation.

4.2 Guide d'annotation

Pour effectuer l'annotation manuelle des inférences de corpus, nous avons produit un guide d'annotation. Cette section présente les grands principes de ce guide. Les guides en française et en chinois sont annexés.

4.2.1 *Objectifs*

Cette annotation a plusieurs objectifs :

1. Nous voulons identifier les frontières des inférences. Même si un commentaire contient des inférences, tous les mots de ce commentaire ne font pas partie des inférences. Les frontières d'une inférence ne sont pas toujours évidentes. Le premier but de l'annotation manuelle est de déterminer l'existence et les frontières des inférences dans le corpus.
2. Nous vérifions en corpus la pertinence de notre classification. Les cinq types d'inférence sont tirés des recherches existantes. Avant d'incorporer les informations inférentielles pour la fouille d'opinion, nous souhaitons vérifier la couverture des cinq types que nous avons définis.
3. Nous voulons vérifier à quelle thématique chaque inférence précédemment annotée est liée. En effet, les longs commentaires peuvent contenir plusieurs inférences et traiter plusieurs thèmes. Les inférences se situent au niveau du lexique, du syntagme, de la phrase ou de l'enchaînement de plusieurs phrases. Par la suite, nous pouvons continuer à découvrir la relation entre le type d'inférence et ces quatre niveaux.

4.2.2 *Annotation des inférences*

Afin d'aider les annotateurs manuels à repérer et annoter les inférences de corpus, nous avons proposé les règles suivantes.

Inspirée par la définition de Doussau et Rigal (2011), nous interprétons une inférence comme un "ajout d'information" qui n'est pas explicitement exprimée dans le texte. Par ailleurs, les inférences constituant un processus d'interprétation ou de raisonnement mental, dans la mesure où elles relient différentes parties d'un message (Fayol 2003). Si le lecteur doit mobiliser des connaissances personnelles absentes du texte, c'est qu'une inférence est probablement présente. De même, si un lecteur relie plusieurs parties liées ou séparées dans le texte pour aboutir à une conclusion, le processus mental contient également une ou des inférences.

Plus concrètement, une portion de texte ne contient des inférences que si l'on a besoin d'une interprétation pour arriver à une relation binaire de <thème, polarité>. Par exemple, un commentaire de"bonne localisation" peut être extrait comme <localisation, positive>. Les deux éléments de cette relation binaire s'expliquent d'eux-mêmes. En revanche, si un commentaire est que "Le personnel parlait couramment l'anglais", la relation binaire est

déduite comme <personnel, positive>. Mais la polarité positive n'est pas explicitement exprimée dans le texte. L'interprétation mentale emprunte des connaissances externes du texte. Le processus mental est que :

- L'auteur était un Chinois qui voyageait en France.
- Il ne parlait pas français.
- Si le personnel ne parlait pas anglais ni chinois, ils ne pouvaient pas communiquer entre eux.
- Ce commentaire est positif.

En plus des informations présentes dans le texte, le lecteur ajoute les connaissances complémentaires afin de comprendre ce commentaire. Ainsi, ce commentaire contient l'inférence. Pareillement pour "proche de la tour Eiffel" (靠近埃菲尔铁塔), il peut être transformé en une relation binaire <localisation, positive>. L'interprétation de ces deux éléments utilise l'inférence.

Dans cette annotation, nous avons distingué la polarité positive, négative et neutre. Quant au thème, nous avons d'abord défini 9 thèmes : localisation, équipement, personnel, propreté, qualité de service, prix, sécurité, clientèle et évaluation générale. Afin de faciliter l'accès à la relation binaire, nous avons ensuite affiné les 9 thèmes en plusieurs sous-thèmes. Par exemple, le thème <localisation> contient des sous-thèmes tels que <transport>, <site touristique>, <restauration>,<environnement> et <transport>. "Près de la tour Eiffel" (靠近埃菲尔铁塔) peut être transformé en une relation binaire <localisation(site touristique), positif>. Les sous-thèmes ne sont pas utilisé directement dans l'annotation manuelle, mais aident à l'annotateur à prendre la décision. La catégorisation des thèmes et des sous-thèmes est présentée dans le tableau 4.1. Si certains énoncés ressemblent à une inférence, mais ne se rapportent pas à un thème ou un sous-thème parmi ceux énumérés dans le tableau, l'annotateur peut proposer un nouveau thème qu'il juge plus approprié.

Thème	Sous-thème						
localisation	géographie	site touristique	shopping et course	restauration	environnement	transport	
équipement	chambre	salle de bain	restauration	région commune	wifi	gratuitement	isolation
personnel	qualité service	attitude	compétence linguistique				
propreté	chambre	salle de bain	région commune				
qualité de service	hôtel	restauration	navette (site, aéroport)	autre service			
prix	hôtel	restauration	autre service				
sécurité	inférieur de l'hôtel	extérieur de l'hôtel					
clientèle							
général	évaluation globale	intégrité					

Tab. 4.1 : Définition des thèmes et des sous-thème

4.2.3 *Annotation des types d'inférences*

Dès qu'une phrase ou une portion de phrase contient une inférence, nous annotons son type en distinguant l'inférence logique, pragmatique, lexicale, énonciative et discursive.

1. Une inférence logique se réalise en effectuant un raisonnement fondé sur une interprétation littérale du texte et ne nécessitant pas de faire appel à des connaissances extérieures. Par exemple, l'interprétation littérale de la phrase "l'hôtel est à côté du métro (酒店在地铁旁边)" signifie que la localisation de l'hôtel est pratique (positive). Cette interprétation est provoquée par un raisonnement de "à côté de" à "proche".
2. Une inférence pragmatique s'appuie sur l'ensemble des connaissances acquises par un individu au cours de ses expériences passées. Par rapport à l'inférence logique, un schéma des connaissances du lecteur est ajouté au processus de raisonnement. Si nous changeons l'exemple précédent en "l'hôtel est à côté du champ de Mars" (酒店在战神广场旁边)", l'interprétation demande d'abord des connaissances sur "champs de Mars". C'est parce que le lecteur sait que le champs de Mars est à côté de la tour Eiffel qu'il peut en déduire que "l'hôtel est proche du site touristique", avec une polarité positive.
3. Une inférence lexicale est le seul type qui se trouve au niveau du groupe lexical. Si un lexique n'est pas une sémantique positive ou négative et porte un sentiment dans le domaine spécifique de l'hôtellerie, nous le classons comme inférence lexicale. Par exemple, "Carrefour", "anglais", "ventilation", "bouilloire", etc.
4. Une inférence énonciative est actualisée en contexte. Concrètement, nous le considérons comme une inférence énonciative, si une ou plusieurs déclarations :
   a) contiennent des adverbes (无济于事, en vain ; 足矣, suffisamment) ou des verbes (号称, soi-disant) qui déterminent la polarité ;
   b) sont interrogatives (房间真的有 20 平米吗? La pièce fait-elle vraiment 20 mètres carrés ?) ;
   c) impliquent des reproches (电梯迷你到不刻意找你都发现不了。L'ascenseur est si petit que vous ne pouvez pas le trouver à moins de le chercher.) ou des appréciations (下次还会入住。Nous allons toujours choisir cet hôtel la prochaine fois.).
5. Une inférence discursive est basée sur des connaissances extralinguistiques au niveau du discours. Ces connaissances permettent notamment de moduler les premières inférences linguistiques et de leur ajouter de nouvelles significations. Il est courant qu'une inférence discursive soit construite par plusieurs énoncés. La principale caractéristique est qu'il est le plus longue par rapport aux quatre types précédents. Par exemple, 入住后洗完澡，穿了浴袍发现浴袍的袖子和口袋里全部是头发，而且口袋里面的头发非常多。(Après l'enregistrement, j'ai pris une douche et mis un peignoir. J'ai trouvé que les manches et les poches de la robe étaient plein de cheveux, et qu'il y avait même beaucoup de cheveux dans la poche.)

Parmi les cinq types, des inférences logiques sont sémantiques, tandis que les inférences pragmatiques et énonciatives se situent davantage au niveau de la pragmatique. Les inférences lexicales et discursives appartiennent à ces deux catégories. Du point de vue des connaissances externes apportées par les lecteurs, cet indice est indispensable pour les inférences pragmatiques et lexicales. Mais il n'y a pas besoin d'inférences logiques et énonciatives. La production de ces deux derniers types dépend plutôt de l'enchaînement du texte. Autrement dit, le lien entre les différentes parties dans un commentaire. Il n'y a que des inférences discursives qui nécessitent à la fois l'introduction des connaissances personnelles et la prise en compte de l'enchaînement du texte. Nous ne pouvons pas prédire les inférences logiques et pragmatiques par leurs longueurs. Par contre, les trois types restants se distinguent selon leur taille textuelle (les inférences discursives étant les plus longues).

Quant au degré de subjectivité, l'inférence logique est le seul type objectif, car elle est le résultat logique du texte. Les quatre autres types nécessitent une interprétation du texte ; cette interprétation conduit à la subjectivité. En d'autres termes, la compréhension du texte par les quatre inférences est influencée par les expériences personnelles de chaque lecteur. L'interprétation personnelle peut être différente sur le même commentaire.

Les cinq types d'inférences ne présentent pas les mêmes niveaux de difficulté du point de vue de l'annotation : les plus faciles à annoter concernent les inférences locales ne demandant pas d'interprétation (inférences logiques), les plus complexes étant celles qui impliquent de prendre en compte de grandes portions textuelles (inférences discursives). Les inférences qui font appel aux connaissances du locuteur (inférences pragmatiques) seront annotées en fonction du niveau de culture générale de l'annotateur. Enfin, il existe un cas possible d'imbrication d'inférences dans la mesure où les inférences discursives couvrent une portion de texte qui peut être décomposées en sous-portions, chaque sous-portion faisant référence à d'autres types d'inférences. Ces différents conseils pour annoter manuellement les inférences révèlent la difficulté de la tâche pour un humain.

4.2.4 *Niveaux d'analyse*

Le troisième objectif est de localiser les inférences dans un commentaire. Nous nous contentons de quatre niveaux.

1. Commentaire : l'ensemble des phrases écrites par un utilisateur. Il peut contenir des opinions sur différents thèmes. Par exemple, un commentaire jugera à la fois de la propreté de la chambre, de la localisation de cet hôtel et de la qualité des services. Un commentaire contient au moins deux signes de ponctuation de fin de phrase.
2. Phrase : une phrase dans un commentaire, dont la fin est indiquée par une ponctuations de fin de phrase telle que 。?! ….
3. Syntagme : une proposition d'une phrase complexe, segmentée par ponctuations comme , ou ;. Il ne contient aucune ponctuation de fin de phrase, mais une ponc-

tuation dans la phrase . Un syntagme peut exprimer une opinion complète ou incomplète. Il ne s'agit donc pas de la définition en linguistique (syntagme nominal, syntagme verbal, etc.)

4. Lexical : un mot ou une expression figé. Il ne contient aucune ponctuation. Ce niveau est spécifiquement désigné pour l'inférence lexicale. Les quatre autres types ne sont jamais localisés au niveau lexical.

4.3 Processus d'annotation

4.3.1 *Sélection des commentaires à annoter*

Nous avons recruté deux stagiaires de niveau M1 pour réaliser ce travail d'annotation manuelle de corpus. En raison de la durée limitée des stages et du coût de ces stages, nous avons sélectionné un sous-ensemble du corpus que nous avons jugé représentatif de l'ensemble du corpus et composé de 1 391 commentaires sur un total de 27 145, soit 5% du corpus global. Pour cela, nous avons appliqué les trois critères suivants.

- polarité : Le nombre de commentaires positifs, négatifs et neutres soit égal.
- taille des phrases : la longueur moyenne des phrases du corpus à annoter doit être strictement supérieure à la longueur moyenne des phrases du corpus d'origine, avec une taille minimum de 10 mots par commentaire. Ce critère permet d'éviter de sélectionner des commentaires très simples sans inférence.
- distribution : la distribution des commentaires par arrondissement, étoile et score suit la même tendance dans les corpus d'origine et d'annotation.

Dans le tableau 4.2, nous présentons la distribution des commentaires par arrondissement dans le corpus d'origine et dans le corpus annoté. Nous observons que les arrondissements les plus touristiques sont aussi bien représentés dans le corpus global que dans le sous-ensemble annoté. La répartition des commentaires par arrondissement dans les deux corpus est représentée sur la figure 4.1. Nous avons contrôlé la distribution d'autres informations de la même manière.

Arrondissement	Corpus d'annotation	Corpus d'origine
1	7,95%	8,48%
2	3,57%	5,03%
3	2,04%	1,58%
4	0,82%	2,80%
5	8,44%	3,59%
6	3,41%	2,16%
7	3,84%	2,73%
8	11,28%	9,42%
9	14,01%	8,99%
10	7,16%	5,46%

Arrondissement	Corpus d'annotation	Corpus d'origine
11	2,23%	3,09%
12	4,71%	6,11%
13	4,57%	6,11%
14	5,72%	8,05%
15	6,76%	5,82%
16	2,67%	4,60%
17	3,33%	3,67%
18	1,92%	3,81%
19	1,04%	1,80%
20	0,71%	2,59%
Autre	8,84%	4,10%

Tab. 4.2 : Comparaison de la distribution des arrondissements

4.3.2 *Outil et configuration d'annotation*

Brat[1] est un outil d'annotation de données textuelles en ligne. En particulier, il est destiné aux annotations structurées. Cela nous permet de créer notre propre schéma d'annotation à partir d'un fichier de configuration. Une version est disponible pour une utilisation locale sur les machines de chaque utilisateur. Quant à la configuration, Brat donne la possibilité de définir quatre sections : [entities], [relations], [events] et [attributes]. Notre travail a utilisé deux des quatre sections. Le schéma d'annotation est désigné comme suivant 4.2 :

Comme présenté dans la section 4.2.4, les annotateurs vérifient pour chaque niveau d'analyse dans le corpus, s'il contient une ou plusieurs inférences. "Inférence" représente la présence d'inférence. "Absence" est l'absence d'inférence. "Incertitude" représente "nous ne savons pas s'il contient une inférence", ce qui inclut les situations suivantes :

A) Erreurs orthographiques incompréhensibles.
B) Un syntagme qui n'a pas de sens.
C) Un syntagme qui ne contient pas une signification complète.

Pour chaque entité annotée comme "Inférence", il faut alors décider de son type. Comme les 5 types ne sont pas exclusifs, il est possible qu'une partie du texte contienne plusieurs types. L'étiquette "sans-type" est réservée aux cas où les annotateurs sont sûrs qu'il y a une inférence, mais ils ne savent pas de quel type il s'agit parmi les types disponibles (possiblement, c'est une inférence d'un type nouveau, non prévu dans le guide).

1. `https://brat.nlplab.org/`

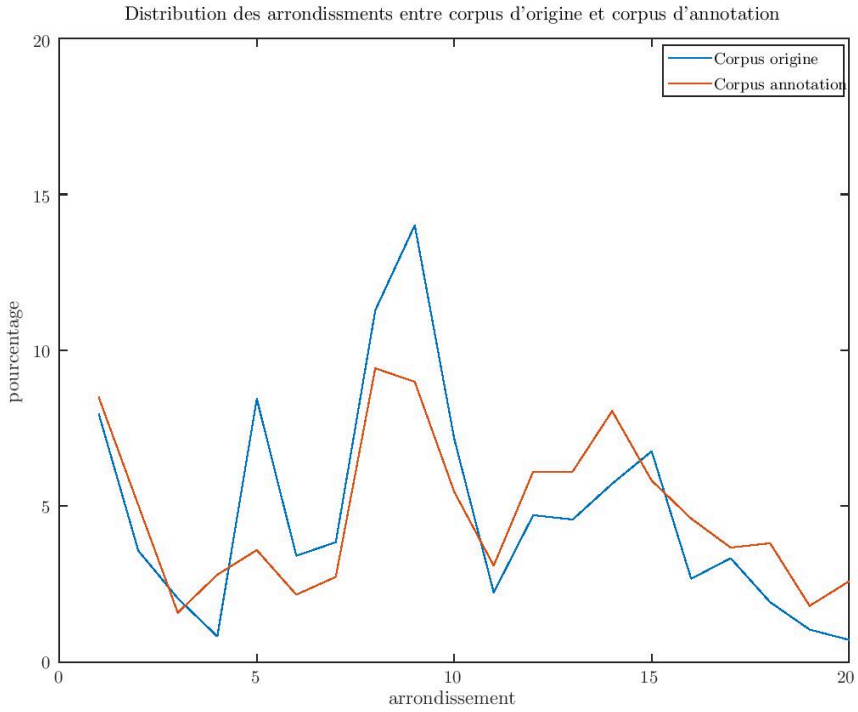


Fig. 4.1 : Comparaison de la distribution des arrondissements

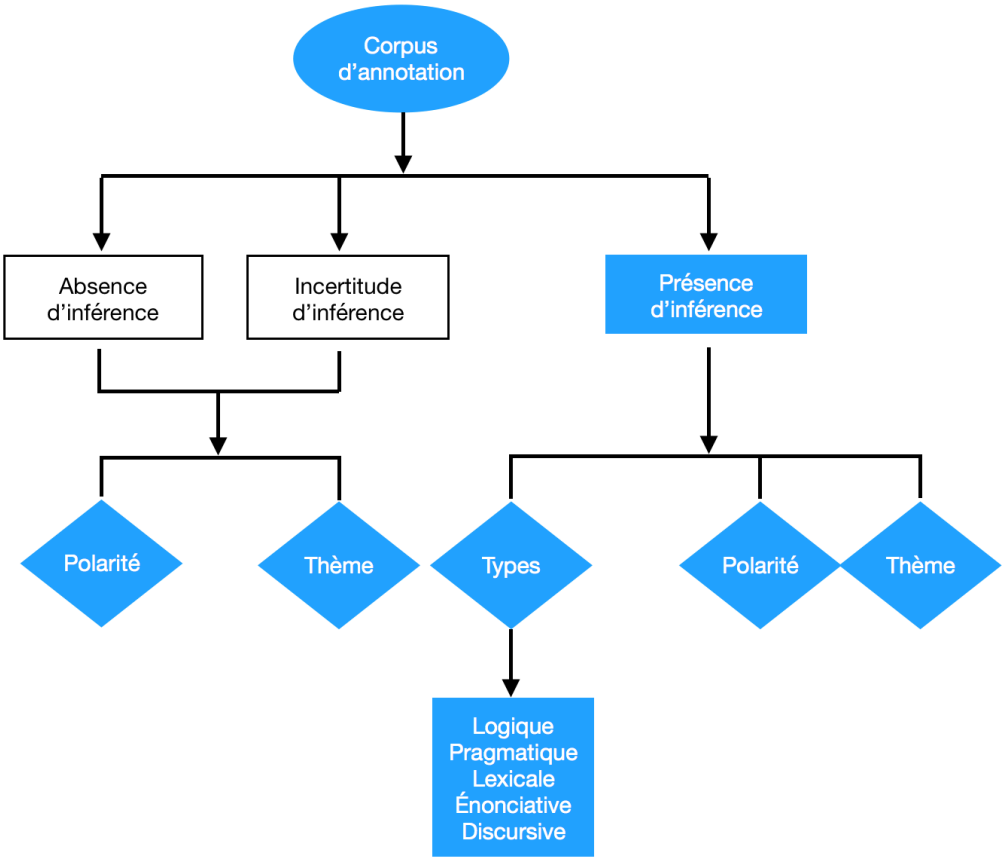


Fig. 4.2 : Schéma d'annotation

La dernière étape consiste à associer un thème et une polarité (parmi positive, négative, neutre et inconnue) à chaque entité. Pour rappel, les thèmes et leurs sous-thèmes sont présentés dans le tableau 4.1.

4.3.3 *Accords inter-annotateurs*

Ce travail d'annotation a été réalisé par trois annotatrices chinoises quatre heures par jour pendant un mois. Au cours de la première semaine, elles ont commencé par lire le guide d'annotation et les exercices. Ensuite, les trois annotatrices ont annoté 50 commentaires. Elles ont comparé et validé une version qu'elles ont réalisée ensemble, ce qui les a aidées à mieux comprendre le guide. Le reste du travail a été réalisée en double annotation, chaque commentaire ayant été annoté par deux annotatrices à chaque fois.

Comme l'auteur du guide d'annotation n'a pas participé au travail d'annotation et

que ce travail n'a été réalisé que par les trois annotatrices, pour vérifier la qualité de l'annotation, nous avons comparé des scores des accords entre la version avant et après la correction.

Ayant un score de Kappa moyen de 0,572, les scores avant la phrase consensuelle n'étaient pas satisfaisants. Cela montrait la difficulté de reconnaître les inférences. Même pour les natifs chinois, un certain nombre de divergences subsistaient. Il est donc nécessaire de poursuivre une correction entre deux versions par une phase d'adjudication des annotations jusqu'à ce qu'un consensus entre les annotatrices soit atteint.

4.3.3.1 Deux versions corrigées en conservant les écarts

Dans la première étape de la correction, si les deux annotatrices ne sont pas d'accord l'une avec l'autre, elles ont le droit de conserver leur propre réponse pour les cas douteux. Après cette correction, la moyenne de Kappa était de 0,939. La plupart des opinions pourraient être unifiées.

La présence d'étiquettes d'inférence trie le corpus en distinguant la présence, l'absence et l'incertitude de l'inférence. Ils définissent principalement la portion d'une inférence. Selon le tableau 4.3, la valeur globale de F-mesure s'élève à 0,998, indiquant que chaque couple d'annotatrices est presque toujours d'accord, en particulier pour l'étiquetage des parties avec des inférences.

Selon le tableau des scores des attributs 4.4, la valeur de la F-mesure globale descend à 0,829. Parmi les cinq types d'inférences, les annotatrices ont eu le plus de difficulté à identifier les inférences discursives (F=0,502). Le type le plus simple à identifier est celui des inférences lexicales. Cela peut être dû au fait que nous avons déjà défini une liste de lexiques contenant des inférences lexicales dans le guide d'annotation. En s'appuyant sur cette liste de base, les annotatrices ont localisé plus facilement les inférences lexicales. Les erreurs se sont concentrées sur les nouvelles inférences lexicales détectées. Les scores des inférences logiques et pragmatiques sont tout autour de la moyenne. Quant au choix du thème et de la polarité, la tâche de la polarité est légèrement mieux accomplie par rapport au thème.

Présence d'inférence	Vrai positif	Faux positif	Faux négatif	Précision	Rappel	F-mesure
Absence	1379	8	4	0,9942	0,9971	0,9957
Présence	4915	3	0	0,9994	1,0000	0,9997
Incertitude	1368	8	6	0,9942	0,9956	0,9949
Global	7662	19	10	0,9975	0,9987	0,9981

Tab. 4.3 : Scores des entités après l'accord

4.3.3.2 Une version de consensus

Dans la deuxième étape de correction, au lieu de conserver les écarts, les annotatrices ont discuté des différents cas afin de produire une version consensuelle. Cette version

Attributs	Vrai positif	Faux positif	Faux négatif	Précision	Rappel	F-mesure
Discursive	72	72	71	0,5000	0,5035	0,5017
Énonciative	305	146	156	0,6763	0,6616	0,6689
Lexicale	2739	138	146	0,9520	0,9494	0,9507
Logique	2736	511	465	0,8426	0,8547	0,8486
Pragmatique	2095	549	479	0,7924	0,8139	0,8030
Polarité	5237	877	849	0,8566	0,8605	0,8585
Thème	4842	1529	1414	0,7600	0,7740	0,7669
Global	18026	3822	3580	0,8251	0,8343	0,8297

Tab. 4.4 : Scores des attributs après l'accord

validée sera utilisée dans les expériences automatiques.

	Version 1	Version 2	V1 vs. V2
Présence d'inférence	0,9627	0,9536	0,9382
Types d'inférence	0,7050	0,6854	0,6505
Polarité	0,8449	0,8170	0,8255
Thème	0,7454	0,7071	0,7056
Global	0,8145	0,7908	0,8612

Tab. 4.5 : F-mesure entre deux versions et version consensuelle

D'après le tableau 4.5, le score de la F-mesure des étiquettes pour la présence d'inférence est supérieur à 0,95, ce qui montre que chacune des deux annotatrices est en accord quasi-préfet concernant la présence et la portion d'inférence (*présence*, *absence* et *incertitude*). Cependant, les scores de la F-mesure pour les types d'inférence chutent à 0,6 - 0,7. Les résultats calculés montrent que les inférences lexicales étaient les plus simples à traiter tandis que les inférences discursives étaient les plus complexes. Cela peut être dû au fait que les frontières de l'inférence discursive sont plus difficiles à définir, car elles contiennent souvent plusieurs phrases dans une seule inférence. Concernant le choix du thème et de la polarité, l'accord inter-annotateurs de la polarité est légèrement supérieur à celui du thème, car le thème a 7 candidats de plus par rapport à la polarité.

En général, le corpus annoté répondait à notre besoin de qualité et de quantité. En effet, les résultats obtenus par les humaines sont représentatifs des difficultés de la tâche et éventuellement nous obligent à travailler sur différentes méthodes pour identifier les inférences.

4.3.4 *Difficultés rencontrées*

Dans cette section, nous présentons une revue du travail d'annotation manuelle.

En premier lieu, le résultat d'annotation obtenu est assez coûteux en temps et en prix. Comme le score de Kappa n'était pas convaincant avant la correction, tout le travail a été validé par au moins deux annotatrices. Trois locutrices natives chinoises ont annoté 1 391 commentaires pendant un mois. Mais par rapport au corpus total, ce corpus annoté n'occupe que 5,12%. Par ailleurs, les inférences sont un objet d'étude complexe pour l'homme, d'autant plus que la classification en cinq catégories est complexe, à la fois pour se familiariser avec les définitions et pour prendre les bonnes décisions d'annotation, notamment au niveau des frontières.

Entrant dans les détails de l'annotation, nous avons clarifié plusieurs points difficiles à annoter :

1. Frontières : il est difficile de reconnaître les frontières du type discursives car un paragraphe raconte plus d'un sujet ;
2. Une grande partie du désaccord est due à l'absence d'un thème clairement exprimé ;
3. Chaque annotatrice a alors sa propre interprétation du contenu du message. Par exemple, si un commentaire mentionne le 15e arrondissement, une annotatrice aura indiqué une incertitude sur la présence d'une inférence dans ce commentaire, tandis que la deuxième annotatrice aura considéré qu'il y a une inférence de type pragmatique car c'est possible d'inférer de cette localisation que l'hôtel est situé au sud-ouest de Paris. Dans cet exemple, nous considérons qu'il n'y a pas d'inférence dans la mesure où le 15e arrondissement ne représente pas un endroit connu ou recherché des touristes chinois.

4.4 Analyse des résultats d'annotation

Notre analyse du corpus annoté se concentre sur trois aspects : la corrélation et la distribution des étiquettes à différents niveaux, une analyse textométrique du corpus annoté et les difficultés rencontrées.

4.4.1 *Corrélation et distribution des catégories selon différents niveaux*

Nous avons distingué trois niveaux d'analyse : un syntagme, une phrase ou un commentaire. Un commentaire contient au moins deux signes de ponctuation finaux, alors qu'une phrase contient un seul. Notre corpus est composé des commentaires des forums. C'est informel. L'utilisation des ponctuations n'est pas régulière. Cependant, une "phrase" est souvent une sorte de "commentaire". Au lieu d'être séparés par la ponctuation finale, des énoncés dans ce type de "commentaire" sont séparés par des espaces. Ces espaces peuvent séparer des phrases ou des syntagmes. Nous avons essayé de produire un modèle qui réintroduit automatiquement des points (frontière de phrase) et des virgules (frontières de syntagmes) à la place d'espaces. La performance de ce modèle s'est avérée assez médiocre, avec une F-mesure de 0,409. En conséquence, nous n'avons pas

utilisé ce modèle ni cherché une autre façon de réintroduire des frontières de phrases et de syntagmes pour ne pas apporter de bruit dans le corpus.

Compte tenu des caractéristiques intrinsèques de chaque type d'inférence, notre hypothèse est que les inférences discursives n'apparaissent pas beaucoup au niveau du syntagme, car une inférence discursive représente un enchaînement de phrases ou une combinaison de phrases. Ce n'est pas au niveau du syntagme, mais à un niveau plus large que le syntagme. Les inférences énonciatives sont basées sur des énoncés au moins avec une signification complète. De même, elles sont moins fréquentes au niveau du syntagme. Quant aux inférences logiques et pragmatiques, nous ne pouvons pas pré-classer leurs présences en niveaux, car ces deux types ne se limitent pas à des caractéristiques morphologiques. Leur présence dépend du déroulement et du contexte du texte. Autrement dit, nous ne pouvons pas spécifier la longueur des inférences logiques et pragmatiques, mais les inférences discursives sont les plus longues et les inférences lexicales sont les plus courtes.

4.4.1.1 Les inférences

	Syntagme	Phrase	Commentaire	Total
Présence	2740	1614	215	4569
Absence	1637	66	0	1730
Incertain	1624	76	0	1700
Total	6001	1756	215	7972

Tab. 4.6 : Distribution des étiquettes de la présence d'inférence selon les niveaux syntagme, phrase et commentaire

Types d'inférence	Syntagme	Phrase	Commentaire
Logique	2297 (83,83%)	1504 (93,18%)	210 (97,67%)
Pragmatique	1839 (67,12%)	1319 (81,72%)	193 (89,77%)
Lexicale	1007 (36,75%)	902 (55,89%)	105 (48,84%)
Énonciative	291 (10,62%)	285 (17,66%)	90 (41,86%)
Discursive	3 (0,11%)	96 (5,95%)	87 (40,47%)
Total (Présence d'inférence)	2740	1614	215

Tab. 4.7 : Distribution des étiquettes des types d'inférence selon les niveaux syntagme, phrase et commentaire

Le tableau 4.6 présente la distribution des étiquettes de présence, d'absence et d'incertitude d'inférence dans chaque niveau. Nous avons un total de 6 001 syntagmes, 1 756 phrases et 215 commentaires dans ce corpus annoté. Les inférences sont présentes dans

la moitié du corpus. Nous observons que la présence des inférences dans le corpus varie en fonction du niveau d'analyse. Au niveau du commentaire, les inférences sont toujours présentes, tandis qu'au niveau des phrases et des syntagmes, des inférences ne sont pas présentes dans chacune des phrases et chacun des syntagmes. En particulier au niveau des syntagmes, la présence d'inférence n'occupe que 45,66%. Plus l'opinion est longue, plus la possibilité d'avoir des inférences est grande.

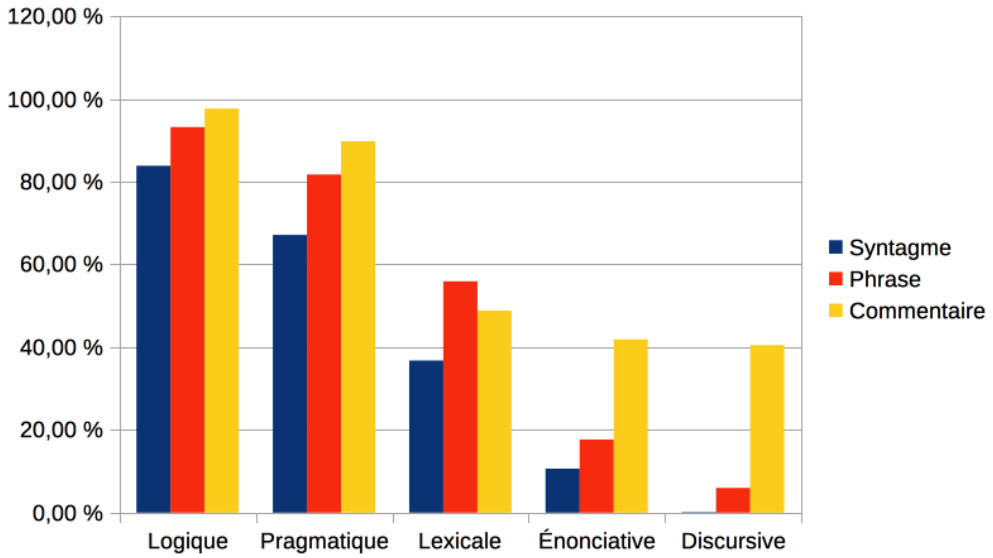


Fig. 4.3 : Distribution des 5 types selon les niveaux syntagme, phrase et commentaire

Au niveau du commentaire, presque tous les candidats contiennent des inférences logiques (97,67%). Ce qui est raisonnable, car les inférences logiques s'effectuent au fil des textes. Pareillement, les inférences pragmatiques occupent une place importante dans des commentaires. Les longs commentaires obligent le lecteur à s'appuyer davantage sur des connaissances extérieures, aussi bien pour comprendre les informations énoncées que pour relier les propositions. Les inférences énonciatives et discursives sont majoritairement présentes dans les commentaires, le taux d'occupation étant plus élevé par rapport aux deux autres niveaux (figue 4.3).

Au niveau de la phrase, ce qui est inattendu, c'est qu'il y a plus d'inférences discursives dans les phrases que dans les commentaires. Ceci est dû aux espaces utilisés comme les points finaux dans un corpus informel. Ce genre de texte, même s'il est caractérisé comme une phrase, c'est plutôt un commentaire. Cette tendance a également lieu des inférences lexicales. Les inférences lexicales font référence à des termes neutres (non polarisés), à l'exception des termes du domaine de l'hôtellerie. Ce genre de lexique est présenté dans des énoncés de coordination. Par exemple, "L'hôtel est proche du centre

commercial, une grande chambre donnant sur la tour Eiffel, une salle de bain propre avec des produits de l'Occitane" dont "tour Eiffel", "centre commercial" et "l'Occitane" sont toutes des inférences lexicales et toutes ont une valence positive.

Au niveau du syntagme, il y a très peu d'inférences discursives. Si un passage est court ou même incompréhensible, il n'est pas possible d'identifier une chaîne d'idées claires.

4.4.1.2 La polarité

	Positif	Négatif	Neutre	Inconnu
Lexique	1315	53	40	0
Syntagme	2096	1618	177	8
Phrase	688	658	92	3
Commentaire	37	130	6	0
Total	2821	2406	275	11

Tab. 4.8 : Distribution des polarités selon les niveaux syntagme, phrase et commentaire

Dans l'annotation, nous avons défini 4 étiquettes de polarité : positive, négative, neutre et inconnue. Le dernier est ajouté pour les opinions incompréhensibles et pour lesquelles nous ne pouvons pas prédire la polarité. En calculant les nombres d'étiquettes (tableau 4.8), le nombre d'inférences positives (2 821) et négatives (2 406) maintient un équilibre.

Il y a 30 phrases positives de plus par rapport aux phrases négatives et 487 plus de syntagmes positifs que négatifs, voire 25 fois plus de lexiques positifs que négatifs. Mais au niveau des commentaires, le nombre négatif augmente fortement à 130 contre 37 de positif. Nous concluons que lorsque les utilisateurs donnent leurs opinions sur plusieurs phrases, il est très probable qu'ils expriment une polarité négative lors de la description de ces expériences.

Les catégories neutres et inconnues sont principalement présentes au niveau des phrases et des syntagmes. En effet, au niveau lexical, les termes portent toujours une valeur de polarité, alors qu'au niveau du commentaire, on observe une combinaison de plusieurs termes positifs, négatifs, ou neutres, à partir desquels il est toujours possible de déterminer une opinion globale (aucune polarité inconnue).

4.4.1.3 Les thèmes

Nous avons déterminé 9 thèmes avec leurs sous-thèmes dans la section 5.2.2 (tableau 4.1). Nous avons trié le tableau 4.9 de la distribution des thèmes par le nombre total. L'équipement et la localisation sont les deux aspects les plus mentionnés par les utilisateurs. Nous pouvons l'interpréter comme les principaux critères d'évaluation. Le thème mixte qui désigne plusieurs thèmes évoqués en une seule partie occupe la troisième place. Le niveau du commentaire a le plus de thèmes mixtes, car les clients mentionnent habituellement plus qu'un thème dans une longue description.

	Lexique	Syntagme	Phrase	Commentaire	Total
équipement	507	1497	488	50	2542
localisation	829	1086	300	17	2232
mixte	0	160	427	45	632
personnel	39	363	112	24	538
général	47	367	34	4	452
qualité Service	4	253	79	30	366
sécurité	10	107	29	11	157
propreté	0	127	20	1	148
prix	0	86	15	1	102
clientèle	9	9	1	1	20

Tab. 4.9 : Distribution des thèmes selon les niveaux syntagme, phrase et commentaire

Quant au lexique, il n'y a pas du tout de thèmes mixtes, propreté et prix. Autrement dit, il n'y a pas d'inférence lexicale qui commente la propreté ou le prix dans un seul lexique. Au moins, ce cas particulier n'existe pas dans notre corpus annoté.

Parmi les 9 thèmes, les opinions portent moins sur la clientèle. Premièrement, les hôtels que nous avons extraits sont sur Booking et Mafengwo, contrairement à Airbnb, la plupart sont des hôtels étoiles. Traditionnellement, les clients ne partagent pas la chambre ou le salon avec d'autres clients. Ils ont moins d'occasions de rencontrer d'autres clients dans un même hôtel. Le thème clientèle ne les concerne pas. Deuxièmement, pour les touristes chinois en France, ils n'ont pas l'habitude de distinguer les Européennes. Ils commentent très peu les autres clients.

4.4.2 *Analyse textométrique du corpus*

Le corpus composé de commentaires d'hôtels contient des caractéristiques telles que des caractéristiques émotionnelles (positive, négative, neutre), des caractéristiques thématiques (service, équipement, localisation) et des caractéristiques inférentielles du corpus. Un corpus peut se diviser en plusieurs sous-parties selon ses caractéristiques. L'outil TXM permet de comparer les vocabulaires entre les différentes sous-parties d'un corpus.

4.4.2.1 Polarité et thème

Le corpus annoté est divisé en quatre sous-corpus par les critères de polarité : positive, négative, neutre et inconnue. Chaque partie contient une seule polarité. En calculant un indice de spécificité des vocabulaires de chaque sous partie, on peut distinguer un mot le plus représentatif d'une sous-partie par rapport aux autres. La figure 4.4 est un extrait de la liste de spécificité en triant le score de la partie négative.

Unités	Fréquence T 70474		negatif t=29472	score	neutre t=3101	score	positive t=28589	score	inconnu t=9312	score
没有	545		447	82,7	18	-0,9	35	-74,1	45	-3,7
不	517		362	37,9	23	0,3	59	-48,4	73	0,5
房间	876		524	26,5	58	2,8	240	-15,7	54	-1...
很小	75		71	21,7	1	-0,8	1	-15,2	2	-2,7
太	171		131	19,7	9	0,5	10	-24,7	21	-0,4
隔音	100		83	16,6	4	-0,3	13	-8,8	0	-6,2
差	68		61	15,7	3	0,2	3	-11,1	1	-3,1
了	1096		588	14,9	50	0,4	230	-43,2	228	11,8
有点	163		118	14,6	5	-0,6	20	-14,8	20	-0,4
旧	43		42	14,5	1	-0,4	0	-9,7	0	-2,6
不好	76		65	14,4	3	-0,2	5	-10,7	3	-2,2
楼梯	45		43	13,7	1	-0,4	0	-10,2	1	-1,9
没	155		110	12,7	2	-1,5	12	-19,6	31	1,9
太小	73		61	12,6	6	1,0	4	-11,1	2	-2,6
厕所	50		45	11,9	1	-0,5	1	-9,8	3	-1,1
要	236		149	10,5	16	1,2	32	-19,0	39	1,1
不是	140		97	10,3	9	0,8	7	-21,5	27	1,6
加	54		46	10,2	2	-0,2	2	-9,4	4	-0,8
小	323		193	10,2	19	0,9	92	-5,4	19	-4,9
的	2574		1234	10,0	110	-0,4	872	-12,2	358	0,8
给	171		113	9,9	10	0,7	24	-13,5	24	0,4
有些	41		36	8,9	2	0,3	1	-7,8	2	-1,1
只有	86		63	8,5	3	-0,3	12	-7,1	8	-0,7
我	449		249	8,4	17	-0,5	68	-31,1	115	11,9
声音	40		34	7,6	2	0,3	0	-9,0	4	-0,4
坏	48		39	7,6	0	-0,9	5	-5,4	4	-0,7
钱	55		43	7,4	0	-1,1	4	-7,5	8	0,4
收费	30		27	7,3	1	-0,2	0	-6,8	2	-0,7
最差	19		19	7,2	0	-0,4	0	-4,3	0	-1,2
上	140		90	7,2	9	0,8	22	-9,8	19	0,3
才	78		56	7,1	1	-0,9	3	-13,2	18	1,9
居然	38		32	7,0	0	-0,7	0	-8,6	6	0,4

Fig. 4.4 : Liste de spécificités des commentaires positifs, négatifs, neutres et inconnus

La négation est la plus représentative de la partie négative avec des variantes des expressions comme “没有”(ne pas avoir), “不”(ne), “不好”(n’être pas bon), “不是”(n’être pas). La négation est beaucoup moins spécifique dans les trois autres parties. Par exemple, l’indice spécifique de “没有” est respectivement 82,7, −0,9, −74,1 et −3,7 dans la partie négative, neutre, positive et inconnue.

Ensuite, des adjectifs qualificatifs comme “小” (petit),“旧” (vieux), “差” (mauvais, faible), “坏” (mauvais, méchant, abîmé) sont plus présents dans les commentaires négatifs. Nous observons également que “小” (petit) est l’adjectif qui est le plus souvent modulé par des adverbes de degré comme “很” (très), “太”(trop). Cependant, certain adverbes “太” (trop), “有点” (un peu), “有些” (quelque peu), “只有” (seulement), “居然” (imprévisiblement) apparaissent seuls également spécifiquement dans cette partie. Le degré de la plupart des adverbes ou plutôt des qualificateurs est léger, contrairement à ceux de la partie positive, dans laquelle les qualificateurs sont le plus souvent intenses comme “非常” (extrêmement), “很” (très) et positivement doux comme “挺” (assez), “蛮” (assez). Si nous distinguons le degré des qualificateurs entre la partie positive et négative avec un score de 1 à 10 (faible à fort), la partie négative se localise vers 2 ou 3, tandis que la partie positive peut atteindre 9 et 10. Cette observation linguistique s’explique également par la culture sociale de la communauté chinoise. Afin d’atténuer les termes d’un commentaire négatif, les utilisateurs chinois ont tendance à utiliser des qualificateurs plus failes. Au contraire, s’ils étaient satisfaits de certaines services, ils ne seraient pas avares des compliments.

Quant aux noms, “隔音” (isolation du bruit), “楼梯”(escalier), “厕所” (toilette), “收费” (frais), “床单” (drap), “洗澡” (douche), “水” (eau), “箱子” (valise) sont plus spécifiques dans les commentaires négatifs par comparaison à la partie positive, neutre et inconnue. Autrement dit, quand un utilisateur mentionne ces thèmes, c’est plutôt pour émettre une critique qu’un avis positif. Cependant, autour des commentaires positifs, nous repérons souvent ” 地理位置” (localisation), “地铁” (métro), ” 超市” (supermarché), ” 景点” (monuments historiques). Étant donné des touriste à Paris, ils sont plus préoccupés par ces points. Lors de l’analyse des indices spécifiques, nous constatons que les commentaires négatifs concernent davantage les services et les équipements à l’intérieur des hôtels, alors que les opinions positives se focalisent plutôt sur l’environnement des hôtels.

Nous observons que dans les passages neutres, les termes porteurs d’opinion employés relèvent des différentes catégories morpho-syntaxiques :

- Les adjectifs et les adverbes comme “普通” (ordinaire), “一般” (moyen), “还好” (ça va), “差不多” (passable).
- Les noms se rapportent à des thèmes mineurs. Par exemple, “火车站” (gare) près de l’hôtel n’est pas aussi important qu’un supermarché à côté. La chambre est un“阁楼” (grenier).
- Les conjonctions sont plus représentatives comme “但是” (cependant), “不过” (pourtant), “但”(mais).

Cette observation nous amène à mettre en évidence trois types de commentaires neutres :

(i) ceux qui utilisent des adjectifs et des adverbes modérés, (ii) ceux qui énoncent un fait sans révéler une opinion personnelle, ou (iii) ceux qui expriment une opinion pleinement positive ou négative avec usage de conjonctions.

La partie “inconnu” regroupe les commentaires qui sont incomplets et incompréhensibles à cause du mauvais découpage du crawler ou des fautes orthographiques. Les termes spécifiques de cette partie sont souvent des pronoms personnels (“我” (je),“我们” (nous), “他” (il)), des ponctuations (“.。””), et des particules (“了”, “着”). Il n’est pas possible de mettre en évidence des propriétés morpho-syntaxiques propres aux commentaires classés “inconnus” dans la mesure où cette catégorie regroupe des commentaires positifs, négatifs, ou neutres.

4.4.2.2 Présence des inférences

Nous avons également étudié les propriétés du corpus selon que nous effectuons une partition entre corpus avec inférence et corpus sans inférence. Au niveau des parties du discours, les verbes, les noms, les mots de localité, les noms de lieux, les noms propres, les noms de personnes, les expressions adverbiales et prépositions.

Au niveau du vocabulaire, les commentaires sans inférences indiquent directement le thème tels que la localisation, la chambre, le personnel, le transport, le service. Nous n’avons pas besoin d’inférer le thème à partir des description. Ensuite, les adjectifs sont à la fois explicites et sentimentaux, par exemple “好” (bon), “不错” (pas mal), “干净” (propre), “方便” (pratique), “热情” (chaleureux), “舒服” (confortable), “友好” (amical) qui peuvent être qualifiés par des adverbes comme “很” (très), “非常” (extrêmement).

Quant au corpus avec inférence, les mots-clés du thème ne sont plus généraux. Les sous-thèmes deviennent plus spécifiques. Par exemple, au lieu du mot localisation, les commentaires mentionnent “地铁站” (métro), “超市” (supermarché), “铁塔” (tour Eiffel).

4.5 Conclusion

Dans ce chapitre, nous avons présenté le travail d’annotation manuelle effectué par trois annotatrices à la fois d’inférences et de fouille d’opinion. Ce travail consiste à annoter les inférences avec cinq types (logique, pragmatique, lexicale, énonciative et discursive), polarité (positive, négative, neutre, inconnue) et 9 thèmes (localisation, service, équipement, etc). Ce travail a été réalisé sur quatre niveaux textuels : lexical, syntagme, phrase et commentaire. Les accords inter-annotateurs avant la mise en enchères ont été mesurés au moyen du coefficient Kappa, avec un score plutôt faible de 0,572, ce qui témoigne de la difficulté de la tâche. La difficulté s’explique à la fois par la complexité de l’objet d’étude (aussi bien pour déterminer la présence d’inférence que pour déterminer le type d’inférence) et par les propriétés du corpus (des commentaires postés sur internet avec des fautes d’orthographe et une absence de ponctuation marquant les frontières de phrases, ce qui conduit à des problèmes de segmentation).

Nous avons aussi fait une analyse textométrique du corpus annoté qui permet de mettre en évidence des propriétés différentes morpho-syntaxiques en fonction du type d'inférence, de la polarité, ou du thème annoté.

Troisième partie

Identification et classification des inférences pour la fouille d'opinion en chinois

Chapitre 5

Identification et classification automatiques des inférences

Contents

5.1 Introduction 59

5.2 Présentation des caractéristiques et des méthodes 60

5.2.1 Corpus d'entraînement 60

5.2.2 Caractéristiques généralisées 60

5.2.3 Plongement des mots 61

5.3 Expérimentation 62

5.3.1 Identification des inférences 62

5.3.2 Classification des types d'inférence 63

5.3.3 Amélioration des performances 65

5.4 Analyse des résultats 65

5.4.1 Identification des inférences 66

5.4.2 Classification des types d'inférence 67

5.4.3 Amélioration des performances du modèle 70

5.5 Conclusion 72

5.1 Introduction

Dans ce chapitre, nous nous concentrons sur l'identification et la classification automatique des inférences. Tout d'abord, nous définissons et normalisons les caractéristiques du modèle SVM. Ensuite, l'expérimentation porte sur plusieurs modèles. À la fin, nous comparons les résultats des modèles afin de trouver une meilleure solution.

5.2 Présentation des caractéristiques et des méthodes

Les machines à vecteurs de support (SVM) peuvent être utilisées pour résoudre des problèmes de discrimination à deux classes ou multi classes (Guermeur 2007). Elles permettent de traiter des problèmes de discrimination non linéaire, et de reformuler le problème de classement comme un problème d'optimisation quadratique. Les performances sont assez élevées lorsque la taille de l'échantillon est petite (<3000 mots) ou moyenne (<50k mots). Comme notre corpus annoté manuellement ne contient que environ38 000 tokens, nous avons choisi SVM comme algorithme pour nos expériences, implémenté dans le module *scikit-learn* (*sklearn*) en Python (Pedregosa et al. 2011).

5.2.1 *Corpus d'entraînement*

Le corpus d'entraînement contient 8 713 commentaires, dont 178 commentaires annotés, 1 520 entités, 1 603 phrases annotées et 5 412 syntagmes annotés. Nous avons défini 2 500 textes aléatoires comme le corpus de test. Les 4 000 textes constituent le corpus d'entraînement et les 2 213 textes restants constituent le corpus de développement.

5.2.2 *Caractéristiques généralisées*

Nous avons d'abord généralisé les caractéristiques du modèle. Ce travail consiste à transformer les métadonnées textuelles en numérique. Les caractéristiques se composent de deux parties : métadonnées morpho-syntaxiques et les métadonnées relatives à l'hôtel.

Les métadonnées morpho-syntaxiques décrivent les textes au niveau linguistique. Avant d'accéder directement aux informations textuelles, nous avons essayé de représenter les textes par ces métadonnées morpho-syntaxiques.

5.2.2.1 Métadonnées morpho-syntaxiques

Négation Nous avons créé une liste de mots de négation. Cette liste a deux types de négation. D'abord elle contient les indices de la négation présentés par un seul idéogramme ("非", "别", "勿", "没", "不", "无", "未", "否", "差"). Ils peuvent être traduits par "ne ... pas", "sans"', ou "non". Ces mots sont souvent combinés avec des verbes, des noms, des adjectifs ou des adverbes pour exprimer la négation. Par exemple, "不好" (n'être pas bon) est une combinaison de "不" et l'adjectif "bon". "没有" (n'avoir pas) est une combinaison de "没" et le verbe "avoir". "无关" (sans aucune relation) est une combinaison de "无" et le nom "relation". Ensuite, la liste de négation contient aussi des mots et expressions figées qui sont composés de plusieurs idéogrammes et qui ne contiennent pas d'indice de négation. Par exemple, "缺少" (pas de), "匮乏" (peu de), "难以" (difficile à), "乏善可陈" (sans valeurs particulières), etc.

Pour chaque commentaire, nous avons détecté s'il contenait des mots de négation dans cette liste. La fréquence de négation a été enregistrée comme une caractéristique.

Longueur de texte Comme présenté dans les analyses des résultats d'annotation manuelle, la longueur du texte a une corrélation positive avec certains types d'inférence. L'inférence discursive est généralement la plus longue parmi les cinq types, alors que l'inférence lexicale est la plus courte. Par conséquent, nous avons également utilisé la longueur de chaque texte comme une caractéristique.

Partie du discours Dans la partie précédente, nous avons déjà prouvé que compte tenu de la performance et du prix, l'annotation automatique dans certaines parties du discours était la meilleure solution. Nous avons donc étiqueté automatiquement le corpus avec Jieba[1].Cet outil propose 54 catégories de parties du discours pour le chinois. Pour chaque texte, la fréquence des catégories grammaticales apparues a été calculée. Nous ajoutons ces 54 catégories comme caractéristiques du modèle.

Derniers caractères Afin d'identifier plus facilement les inférences dans les textes ainsi que leurs frontières, nous nous intéressons aux derniers caractères des portions annotées. Le dernier caractère de chaque portion annotée peut être différent selon les types d'inférence. Par exemple, l'inférence lexicale ne contient aucune ponctuation. Le dernier caractère est probablement un idéogramme. Les inférences discursives et énonciatives expliquent un point de vue complet. Normalement elles ne se terminent pas par une virgule. Pour cette raison, il est intéressant de conserver le dernier caractère, même deux derniers caractères de chaque portion annotée. Nous avons distingué l'idéogramme, l'espace, la ponctuation de fin de phrase (.?!) et la ponctuation de fin de syntagme (,;).

5.2.2.2 Métadonnées de l'hôtel

Les métadonnées d'un hôtel sont extraites du corpus d'origine. Dans l'expérimentation, nous avons sélectionné certaines données comme les métadonnées des hôtels :

- Score global d'un hôtel de 0 à 10
- Classement d'un hôtel en nombre d'étoiles de 0 à 5
- Localisation : standardisé par arrondissement à Paris du 1er au 20e (1 à 20)
- Score de l'hôtel évalué par l'utilisateur de 0 à 10

Avant de passer ces métadonnées numériques directement au modèle, il nous faut standardiser les données avec la fonction *StandardScaler* du sous-module *preprocessing*. Cette fonction garantit que toutes les données sont centrées autour de 0 et ont une variance dans le même ordre.

5.2.3 *Plongement des mots*

En ce qui concerne les informations textuelles, nous avons transformé les textes en vecteur avec différentes méthodes : doc2vec, sac de mots, TF-IDF, N-gram et word2vec

1. `https://github.com/fxsjy/jieba`

avec les mots-clés. Les quatre premiers vecteurs ont été créés par la fonction *CountVectorizer* de sklearn.

Pour la méthode Word2Vec (W2V), nous avons appliqué la méthode suivante. À partir d'un corpus complet de 817 758 tokens, segmenté par l'outil jieba, nous avons ensuite entraîné un modèle Word2Vec en utilisant le module gensim[2] de Python.

Sur le même corpus, nous avons effectué une classification non-supervisée en 10 clusters avec le module *LDA*, et sélectionné les 25 mots-clés les plus représentatifs de chaque groupe. En supprimant les doublons, nous obtenons 216 mots clés.

Puis, nous avons calculé le score de similarité entre chaque mot du texte et tous les mots-clés. Selon le score de similarité, nous avons défini quatre intervalles : 0,8 - 1, 0,6 - 0,8, 0,4 - 0,6, inférieur à 0. Pour un texte, nous avons additionné les scores qui tombaient dans le même intervalle. De cette manière, chaque texte a été transformé en vecteur de 864 (216 mots clés * 4 intervalles) dimensions.

5.3 Expérimentation

Dans les expériences suivantes, nous voulons vérifier s'il est possible d'identifier des inférences dans un commentaire, et de typer automatiquement les inférences précédemment identifiées. Dans cette étape, nous avons utilisé simplement les paramètres par défaut du modèle SVM, car le but n'est pas de prouver la puissance du modèle, mais de valider les caractéristiques choisies et le mode de l'expérimentation.

5.3.1 *Identification des inférences*

Nous commençons par identifier la présence des inférences. Pour chaque commentaire, le modèle prédira s'il contient des inférences sans que le modèle n'ait à en préciser le nombre. Nous avons défini trois catégories pour le repérage des inférences : absence, présence, et incertitude (pour le cas où il est difficile de déterminer s'il y a inférence ou pas). Afin d'identifier automatiquement les inférences, nous avons testé les six configurations suivantes :

1. Configuration 1 : les métadonnées de l'hôtel et des informations morpho-syntaxiques (ci-après dénommé "métadonnées").
2. Configuration 2 : métadonnées et vecteurs de Doc2Vec.
3. Configuration 3 : métadonnées et vecteurs de TF-IDF.
4. Configuration 4 : métadonnées et vecteurs de N-gram.
5. Configuration 5 : métadonnées et vecteurs de sac de mots.
6. Configuration 6 : métadonnées et vecteurs de Word2Vec.

À la fin des 6 configurations, nous envisageons de sélectionner la meilleure combinaison des caractéristiques pour l'identification des inférences. Plus concrètement, nous

2. `https://pypi.python.org/pypi/gensim`

choisissons un vecteur parmi les 5 et décidons si une combinaison de métadonnées et de textes vectoriels est une meilleure solution pour identifier les inférences.

Métadonnées Dans la première configuration, nous n'avons utilisé que les caractéristiques concernant les traits morpho-syntaxiques et hôteliers. Dans cette première expérience, aucune information du texte n'est utilisée. Après avoir été converties et standardisées en vecteur, les métadonnées sont utilisées comme caractéristiques du modèle SVM. Au total, ce modèle contient 61 caractéristiques.

Métadonnées et des vecteurs de Doc2Vec Dans la deuxième configuration, les caractéristiques a été ajoutées par les vecteurs de Doc2Vec. Nous avons considéré une ligne de texte comme un mini document et avons utilisé le module *Gensim.Doc2Vec* pour convertir chaque ligne en vecteur de 250 dimensions. Ce modèle contient 311 caractéristiques.

Métadonnées et des vecteurs de TF-IDF, N-gram et sac de mots Dans les trois étapes suivantes, nous avons utilisé *sklearn.CountVectorizer* pour convertir respectivement les textes en vecteur de TF-IDF, N-gram et sac de mots dont N-gram était défini comme uni-gramme, bi-gramme et tri-gramme avec un seuil minimal de 1. Les autres paramètres étaient ceux définis par défaut.

Métadonnées et des vecteurs de Word2Vec À la fin, nous avons appliqué la méthode de Word2Vec. Puisque le SVM demande le même nombre de caractéristiques entre chaque texte d'entraînemnt et que la longueur de chaque texte n'est pas identique, nous ne pouvons pas utiliser les vecteurs de word2vec directement. Nous avons créé une méthode qui combine des mots-clés de clusters et word2vec.

D'abord, nous avons utilisé LDA non-supervisé afin de classifier les textes en 10 groupes. Ensuite, nous avons sélectionné 216 mots-clés les plus représentatifs parmi les 10 groupes. Sur cette base des vecteurs de Word2vec, dans chaque texte, nous avons calculé la similitude de tous les mots avec les 216 mots-clés et divisé le score de similarité en 4 intervalles. Pour un mot-clé, nous avons examiné tous les mots dans un texte si le score de similarité des deux mots est moins de 0,4, entre 0,4 et 0,6, entre 0,6 et 0,8, ou supérieure à 0,8. De cette manière, nous avons obtenu un vecteur avec une dimension de 864 (216*4). Par conséquent, les caractéristiques contenaient 61 plus 864 éléments.

Parmi les 6 configurations, le dernier a obtenu un meilleur résultat. Nous présentons les résultats de ces 6 configurations dans la section *Analyse des résultats*.

5.3.2 *Classification des types d'inférence*

Pour les commentaires qui ont été identifiés comme contenant une ou plusieurs inférences, nous avons ensuite tenté de classer automatiquement les inférences parmi les

cinq types présentés au chapitre 5 : logique, pragmatique, lexicale, énonciative et discursive. Étant donné que les cinq types ne sont pas exclusifs, une inférence candidate peut contenir à la fois plusieurs types.

Dans cette expérience, nous avons testé deux modèles :

- un modèle par type d'inférence (soit 5 modèles qui prédisent chacun un seul type d'inférence)
- un modèle multi-labels qui peut prédire les 5 types d'inférences

5.3.2.1 Modèles spécifiques à chaque type

Cette modalité transforme la classification multiple en binaire. Pour chaque type, nous avons construit un modèle pour prédire l'absence ou la présence de ce type. Les raisons pour lesquelles nous commençons par ce genre de modèle sont qu'il nous donne une idée globale de la difficulté de prédire les types d'inférences et de comparer la performance du modèle en appliquant les différentes caractéristiques.

Comme les 6 configurations de l'identification des inférences, nous avons appliqué les mêmes combinaisons des caractéristiques.

1. Configuration 1 : les métadonnées de l'hôtel et des informations morpho-syntaxiques (ci-après dénommé "métadonnées").
2. Configuration 2 : métadonnées et vecteurs de Doc2Vec. Les résultats de vecteurs sont générés par le module *Gensim.Doc2Vec*.
3. Configuration 3 : métadonnées et vecteurs de sac de mots. Les vecteurs sont des résultats de *sklearn.CountVectorizer* avec un seuil minimal de 1.
4. Configuration 4 : métadonnées et vecteurs de TD-IDF. Les vecteurs viennent des résultats de *sklearn.TfidfTransformer*.
5. Configuration 5 : métadonnées et vecteurs de N-gram. Les vecteurs viennent des résultats de *sklearn.CountVectorizer* avec uni-gramme, bi-gramme et tri-gramme.
6. Configuration 6 : métadonnées et vecteurs de Word2Vec.

Pour chaque type, nous avons obtenu 6 modèles. Les résultats sont présentés dans le tableau 5.2. Les meilleures caractéristiques sont toujours la combinaison des métadonnées et des vecteurs de Word2Vec. Les résultats seront discutés dans la section suivante.

5.3.2.2 Modèle multi-labels

Au lieu de construire un modèle pour chaque type, nous avons réalisé un modèle multi-labels qui nous permet de classer les cinq types en même temps. Pour chaque texte, ce modèle sélectionne tous les types qu'il considère approprié. Nous gardons des expériences précédentes la meilleure configuration (combinaison des méta-données et vecteurs W2V). Le modèle était réalisé par module *OneVsRestClassifier* de *sklearn.multiclass*. Nous avons toujours utilisé les paramètres par défaut.

5.3.3 *Amélioration des performances*

Comme nous pouvons le voir dans le tableau 5.3, la distribution des inférences par type n'est pas équilibrée : sur les 3 254 inférences, il n'y a que 68 inférences discursives et 185 énonciatives.

Pour améliorer les performances du modèle, nous envisageons deux solutions :

- fusionner certaines catégories
- balance les échantillons

5.3.3.1 Modèle des types fusionnés

Fusion des catégories "Absence" et "Incertitude" L'un des nos objectifs est de prendre en compte des inférences pour la fouille d'opinion. Pour ce faire, nous avons distingué une seule catégorie avec inférence et une autre catégorie pour les situations restantes. Autrement dit, les catégories "Absence" et "Incertitude" ont été fusionnées en une seule catégorie.

Fusion des inférences énonciatives et discursives Notre deuxième approche consiste à fusionner les inférences énonciatives et discursives dans une seule catégorie. Les deux types partagent quelques caractéristiques commun. Premièrement, les deux types d'inférence concernent des textes qui ont la particularité d'avoir une structure syntaxique complète et de ne véhiculer qu'un seul sens. Deuxièmement, ces deux types renvoient aux portions les plus longues parmi celles contenant des inférences. Par contre, le nombre des deux types d'inférence est le plus petit. Pour tous les textes contenant l'inférence énonciative ou discursive, nous l'avons ajouté dans une nouvelle catégorie. Ensuite, nous avons appliqué respectivement le modèle uni-label et multi-labels pour comparer les résultats avant et après la fusion.

5.3.3.2 Échantillonnage équilibré

Notre deuxième solution vise à équilibrer la distribution des inférences pour chaque type. On ne pouvait pas ajouter simplement les échantillons de l'inférence discursive, car une inférence discursive contient plusieurs autres types. Le pourcentage de chaque type d'inférence est stable. Nous avons donc réduit le nombre des quatre types au même nombre de l'inférence discursive.

Le défaut de cette méthode était que la taille des échantillons était considérablement réduite.

5.4 Analyse des résultats

Dans cette section, nous présentons et analysons les résultats correspondant à chacune des expériences précédemment décrites.

5.4.1 *Identification des inférences*

En identifiant des inférences, nous avons construit six modèles. Les résultats sont présentés dans le tableau 5.1.

Caractéristiques	Présence (1656)			Absence (347)			Incertitude (497)			Global (2500)
	P [7]	R [8]	F [9]	P	R	F	P	R	F	Macro F-mesure
M [1]	0,81	0,89	0,85	0,67	0,66	0,66	0,66	0,43	0,52	0.68
M+D2V [2]	0,66	0,96	0,80	0,61	0,33	0,43	0,50	0,02	0,04	0.42
M+T-I [3]	0,66	1,00	0,80	0,00	0,00	0,00	0,00	0,00	0,00	0.27
M+Ng [4]	0,66	1,00	0,80	0,00	0,00	0,00	0,00	0,00	0,00	0.27
M+BOW [5]	0,66	1,00	0,80	0,00	0,00	0,00	0,00	0,00	0,00	0.27
M+W2V [6]	0,84	0,90	0,87	0,75	0,85	0,80	0,68	0,47	0,56	0,74

Tab. 5.1 : Résultats de l'identification des inférences selon les différentes caractéristiques

[1] Métadonnées
[2] Vecteurs de Doc2Vec
[3] Vecteurs de TF-IDF
[4] Vecteurs de N-gram
[5] Vecteurs de sac de mots
[6] Vecteurs en Word2Vec
[7] Précision
[8] Rappel
[9] F-mesure

Métadonnées et étiquettes morpho-syntaxiques Même sans le contenu du texte, le modèle peut déjà faire une première identification des inférences avec les métadonnées. Le modèle a prédit à la fois la "présence", l'"absence" et l'"incertitude" des étiquettes des inférences dans les textes. Il a obtenu une macro F-mesure de 0,68. Par rapport à l'"absence" et l'"incertitude", le modèle a bien réussi à identifier la présence des inférences avec une F-mesure de 0,85. Son rappel de 0,89 montre que le modèle avec uniquement des métadonnées a la possibilité de retrouver la plupart des candidats avec des inférences. Mais il est moins capable de distinguer les textes sans inférences et les cas incertains. Les deux dernières catégories contiennent beaucoup moins de textes par rapport aux textes avec inférences. Ce déséquilibre des échantillons conduit à des performances différentes de chaque catégorie.

Métadonnées avec les vecteurs de Doc2Vec Comme la méthode Doc2Vec est plus sensible aux différences entre les documents, ce n'est pas une méthode adaptée à notre corpus spécifique. Le score de la F-mesure même descend par rapport à celui de la première configuration. La représentation du contenu des documents avec doc2vec ne suffit pas au classifieur pour identifier la présence d'inférences, et dégrade les résultats. On peut imaginer que c'est parce que les mêmes mots sont utilisés dans les commentaires avec et sans inférences que cette méthode ne fonctionne pas. Même si les commentaires

contiennent différents thèmes tels que “localisation”, “propreté”, “qualité de service”, etc, par rapport à la distinction entre les commentaires de l’hôtel et les critiques politiques, la différence des sous-thèmes n’est pas aussi distinctive. Les vecteurs générés par Doc2Vec ne représentaient pas clairement les textes.

Métadonnées et des vecteurs de TF-IDF, N-gram et sac de mots Les résultats de ces trois méthodes ne sont pas satisfaisants, en particulier pour la catégorie “absence” et “incertitude” des inférences. Les vecteurs n’ont pas été pas pris en compte lors de la prédiction de ces deux catégories. Avant de conclure que les trois vecteurs ne sont pas convenables pour les tâches des inférences, nous avons continué à appliquer les trois vecteurs dans la classification des types d’inférences ci-après.

Métadonnées et des vecteurs de Word2Vec Jusqu’à présent, ces vecteurs de Word2Vec ont obtenu le meilleur résultat avec une F-mesure de 0,74. Le modèle a réussi à récupérer les textes avec et sans inférences, avec une F-mesure respectivement de 0,87 et 0,80. La faiblesse réside dans les textes incertains. A priori, Word2Vec tient compte du contexte, il est donc nécessaire de s’appuyer sur les mots et leur contexte pour identifier les inférences.

5.4.2 *Classification des types d’inférence*

Avant de passer directement à la présentation des résultats, nous avons commencé par passer en revue les relations de ces cinq types d’inférence. Principalement, les cinq types ne sont pas exclusifs. Un commentaire peut contenir les cinq types, certains types ou un seul type. Ensuite, les relations entre eux ne sont pas identiques. L’inférence discursive qui est subsumée par les autres types représente l’interaction des phrases. Cependant, la longueur moyenne de ce type est normalement la plus longue. L’inférence lexicale qui est souvent produite par un mot ou un groupe nominal est la plus petite unité parmi les cinq types. Une inférence lexicale ne peut pas contenir une autre inférence. Par rapport aux inférences lexicales qui sont au niveau du lexique, les inférences énonciatives et discursives se produisent sur les phrases ou l’enchaînement de plusieurs phrases. Les deux dernières inférences ne se trouvent jamais dans un syntagme ni un mot. Quant aux inférences logiques et pragmatiques, la définition se fonde sur la manière dont une opinion est comprise. L’inférence logique repose sur une compréhension successive par déroulement de textes, tandis que l’inférence pragmatique s’appuie sur certaines informations au sein d’une opinion. De la compréhension et de l’interprétation de ces informations locales, les inférences se produisent. Les inférences logiques et pragmatiques ne sont donc pas limitées par la forme morphologique.

5.4.2.1 Modèles spécifiques à chaque type

Dans la première étape de la classification des types d'inférence, nous avons réalisé 5 modèles pour prédire 5 types en utilisant les 6 combinaisons des caractéristiques. Les résultats sont présentés dans le tableau 5.2.

Inférence logique, pragmatique et lexicale Pour les trois types, c'est toujours le modèle Word2Vec qui obtient les meilleures performances avec une F-mesure de 0,81, 0,78 et 0,85 respectivement. Cette observation a été confirmée par les résultats de l'identification des inférences dans la section précédente.

Nous pouvons en conclure que les meilleures caractéristiques sont une combinaison de toutes les métadonnées et vecteurs de Word2Vec. Nous n'utiliserons plus les autres vecteurs dans l'expérimentation suivante.

Inférences énonciatives et discursives Si un commentaire se compose de plusieurs phrases, il a la possibilité de contenir une inférence énonciative ou discursive. Ce commentaire contient en même temps plusieurs inférences lexicales, logiques ou pragmatiques. Le nombre des inférences logiques, pragmatiques est supérieur à celui des inférences énonciatives et surtout discursives. Cela provoque un déséquilibre des échantillons dans le corpus d'entraînement et de test. Le corpus contient un total de 3 587 inférences logiques, 2 947 inférences pragmatiques, 3 374 inférences lexicales, mais seulement 547 inférences énonciatives et 154 inférences discursives. Ce dernier n'occupe que 1,45% parmi l'ensemble des inférences. En raison de ce déséquilibre, la performance du modèle dans l'identification des inférences énonciatives et discursives n'est pas satisfaisante. Les scores de F-mesure pour tous les modèles sont inférieurs à 0,57. En particulier, les modèles n'ont pas réussi à trouver les textes avec des inférences énonciatives et discursives, car les scores du rappel sont tous inférieurs à 0,1. Aucun vecteur testé ne permet d'identifier les inférences de type énonciatif et discursif. Dans ce qui suit, nous ajusterons les modèles afin de résoudre ce problème de données déséquilibrées.

5.4.2.2 Modèle multi-labels

Par rapport au score de la macro F-mesure (0,712) des modèles unitaires, le score du modèle multi-labels est un peu plus faible (0,70). Un mélange de types d'inférence rend la tâche plus difficile. Les résultats (Tableau 5.3) montrent que les inférences logiques, pragmatiques et lexicales sont plus faciles à classer. La faiblesse réside toujours dans les inférences énonciatives et discursives avec les F-mesure inférieures à 0,54.

L'amélioration du modèle est davantage déterminée en trouvant les raisons pour lesquelles les inférences énonciatives et discursives sont moins bien identifiées. La première raison la plus évidente est le déséquilibre de l'échantillon. Les données d'entraînement sont manquantes afin que le modèle puisse mieux apprendre les propriétés de ces deux types. Cependant, un simple ajout des données ne peut pas résoudre le problème, car un

Types d'inférence	M [1]	D2V [2]	T-I [3]	Ng [4]	BOW [5]	W2V	Présence				Absence				Global	
							R [6] [7]	P [8]	F [9]	Nombre	R	P	F	Nombre	Macro F-mesure	Nombre
Logique	X						0,78	0,71	0,74	1081	0,79	0,85	0,82	1419	0,78	2500
	X	X					0,82	0,48	0,61		0,70	0,92	0,79		0,70	
	X		X				0,86	0,51	0,64		0,71	0,94	0,81		0,72	
	X			X			0,96	0,22	0,36		0,63	0,99	0,77		0,57	
	X				X		0,86	0,51	0,64		0,71	0,94	0,81		0,72	
	X					X	0,82	0,73	0,77		0,81	0,88	0,84		0,81	
Pragmatique	X						0,84	0,52	0,64	1006	0,74	0,94	0,83	1494	0,74	2500
	X	X					0,88	0,33	0,49		0,68	0,97	0,80		0,64	
	X		X				0,92	0,33	0,49		0,69	0,98	0,81		0,65	
	X			X			0,99	0,10	0,18		0,62	1.00	0,77		0,47	
	X				X		0,92	0,34	0,49		0,69	0,98	0,81		0,65	
	X					X	0,84	0,62	0,71		0,78	0,92	0,85		0,78	
Lexicale	X						0,86	0,61	0,71	914	0,81	0,94	0,87	1586	0,79	2500
	X	X					0,83	0,45	0,58		0,75	0,95	0,84		0,71	
	X		X				0,79	0,50	0,61		0,76	0,92	0,83		0,72	
	X			X			0,00	0,00	0,00		0,63	1.00	0,78		0,39	
	X				X		0,80	0,51	0,62		0,76	0,92	0,84		0,73	
	X					X	0,86	0,76	0,81		0,87	0,93	0,90		0,85	
Énonciative	X						1.00	0,05	0,09	185	0,93	1.00	0,96	2315	0,53	2500
	X	X					0,00	0,00	0,00		0,93	1.00	0,96		0,48	
	X		X				0,00	0,00	0,00		0,93	1.00	0,96		0,48	
	X			X			0,00	0,00	0,00		0,93	1.00	0,96		0,48	
	X				X		0,00	0,00	0,00		0,93	1.00	0,96		0,48	
	X					X	0,90	0,10	0,18		0,93	1.00	0,96		0,57	
Discursive	X						1.00	0,06	0,11	68	0,97	1.00	0,99	2432	0,55	2500
	X	X					0,00	0,00	0,00		0,97	1.00	0,99		0,49	
	X		X				1.00	0,06	0,11		0,97	1.00	0,99		0,55	
	X			X			0,00	0,00	0,00		0,97	1.00	0,99		0,49	
	X				X		1.00	0,04	0,08		0,97	1.00	0,99		0,54	
	X					X	1.00	0,06	0,11		0,97	1.00	0,99		0,55	

Tab. 5.2 : Comparaison des résultats de la classification des 5 types avec de différentes caractéristiques

[1] Métadonnées
[2] Vecteurs de Doc2Vec
[3] Vecteurs de TF-IDF
[4] Vecteurs de N-gram
[5] Vecteurs de sac de mots
[6] Vecteurs en Word2Vec
[7] Rappel
[8] Précision
[9] F-mesure

Types d'inférence	Précision	Rappel	F-mesure
Logique (1081)	0,86	0,79	0,82
Pragmatique (1006)	0,88	0,72	0,79
Lexicale (914)	0,88	0,79	0,84
Énonciative (185)	0,44	0,54	0,49
Discursive (68)	0,49	0,60	0,54
Micro-moyenne (3254)	0,83	0,75	0,79
Macro-moyenne (3254)	0,71	0,69	0,70

Tab. 5.3 : Résultats de la classification multi-labels des types d'inférence

nouveau texte contenant une inférence discursive contient souvent plusieurs autres types. La distribution des données ne sera pas modifiée.

La deuxième raison est que ces deux types, en particulier l'inférence discursive, sont difficiles à identifier, même pour les humains. Lors de l'annotation manuelle des inférences, l'accord inter-annotateurs sur l'inférence discursive est le plus faible (F-mesure de 0,501). La principale contradiction est centrée sur la frontière. Puisqu'elle peut contenir plusieurs phrases, il est nécessaire de décider la portion d'une inférence discursive. Les phrases écrites sous forme de commentaires ne sont pas toujours correctement séparées par des points-virgules et ne peuvent être séparées que par des espaces. Même les annotateurs chinois natifs ont du mal à standardiser la taille de la portion. Ce serait plus difficile pour une machine.

5.4.3 *Amélioration des performances du modèle*

5.4.3.1 Transformation les textes en pinyin

Pour le moment, le corpus d'entraînement est constitué d'idéogrammes. Nous nous sommes demandé si la transformation du texte intégral en pinyin améliorait les résultats. En gardant la même procédure et les mêmes caractéristiques, nous avons utilisé le corpus en pinyin pour entraîner le modèle. Les résultats ont été légèrement améliorés avec une augmentation de 0,02 pour le score de la F-mesure. Toutefois, les textes en pinyin ne sont pas pertinents pour l'analyse de texte. Par rapport à l'amélioration minime, nous perdrons beaucoup d'informations textuelles. Par conséquent, nous n'avons pas adopté cette méthode et avons continué à utiliser les textes des idéogrammes dans notre travail.

	Précision	Rappel	F-mesure
Logique (1081)	0,85	0,80	0,82
Pragmatique (1006)	0,88	0,73	0,80
Lexicale (914)	0,90	0,84	0,87
Énonciative (185)	0,49	0,53	0,51
Discursive (68)	0,56	0,59	0,57
Macro-moyenne (3254)	0,73	0,70	0,71

Tab. 5.4 : Résultats du modèle de multi-labels avec les textes en pinyin

5.4.3.2 Fusion des catégories

Absence et incertitude Pour reconnaître les inférences présentes dans les textes, la catégorie la plus essentielle est la "présence". Nous avons donc fusionné les catégories

“absence” et “incertitude” afin de simplifier la tâche.

Les cinq modèles ont tous été améliorés avec une augmentation du score de la F-mesure (Tableau 5.5). Le meilleur résultat reste le modèle de Word2Vec avec un score de 0,79. Nous pouvons conclure que la différence entre la catégorie “absence” et “incertitude” est inutile pour le repérage. L’essentiel est de distinguer les textes avec inférence des textes restants.

Identification des inférences	Macro F-mesure avant le fusionnement	Macro F-mesure après le fusionnement
Métadonnées	0,68	0,76
Métadonnées + Doc2Vec	0,42	0,65
Métadonnées + TF-IDF	0,27	0,59
Métadonnées + N-gram	0,27	0,40
Métadonnées + Sac de mots	0,27	0,59
Métadonnées + Word2Vec	0,74	0,79

Tab. 5.5 : Comparaison des résultats de l’identification des inférences avant et après le fusionnement des catégories “Absence” et “Incertitude”

Inférences énonciatives et discursives Comme le nombre des inférences énonciatives et discursives est plus faible que pour les autres types, nous avons fusionné ces deux types et comparé les résultats des modèles uni-label et multi-labels.

En comparant les F-mesure du modèle uni-label, le résultat de cette fusion n’a pas changé par rapport aux anciens modèles (Tableau 5.6). En revanche, cette fusion a effectivement amélioré la macro F-mesure du modèle de multi-labels avec une augmentation de 0,70 vers 0,75.

Type d’inférence	Précision	Rappel	F-mesure
Énonciative et discursive fusionnées (224)	0,91	0,54	0,56
Énonciative (185)	0,92	0,55	0,57
Discursive (68)	0,99	0,53	0,55

Tab. 5.6 : Comparaison des résultats de la classification multi-labels après et avant de fusionner l’inférence énonciative et discursive

5.4.3.3 Échantillons équilibrés

Au lieu de fusionner certains types, nous avons réorganisé le corpus d’entraînement en conservant le même nombre d’échantillons pour chaque type. Selon les scores de la

macro F-mesure, le modèle avec des échantillons équilibrés (0,82) a été amélioré par rapport à l'ancien modèle. Il a obtenu le meilleur résultat pour l'instant. Cependant, la taille du corpus a été réduite de 8 715 textes à 1 244 textes. Nous n'avons pas adopté cette méthode.

5.5 Conclusion

Dans ce chapitre, nous avons effectué les expériences pour identifier et classer les inférences avec SVM. En analysant des résultats, nous nous somme arrivés à quelques conclusions :

- Lors de l'identification des inférences, si l'on s'intéresse seulement à la catégorie "présence" dans le travail suivant, la meilleure solution est de fusionner la catégorie "absence" et "incertitude".
- En comparant les résultats de la classification des types d'inférence, nous avons décidé d'utiliser un modèle spécifique par type.
- En ce qui concerne le déséquilibre du corpus, la fusion de certains types n'a pas résolu ce problème. L'équilibre du corpus a amélioré le résultat, mais la réduction du corpus a perdu plein d'informations textuelles.

Le meilleur score de F-mesure pour l'identification des inférences sont de 0,79. Si ces résultats sont globalement bons pour déterminer la présence d'inférences dans un commentaire, ils restent faibles pour certains types d'inférences. Le repérage et le typage automatique des inférences est une tâche complexe.

Dans le prochain chapitre, nous présentons l'utilisation des informations de repérage et de typage d'inférence automatique pour améliorer les performances de la fouille d'opinion.

Chapitre 6

Inférences dans la fouille d'opinion

Contents

6.1 Introduction 73

6.2 Méthode lexicale pour la fouille d'opinion 74

6.2.1 Ressources : ontologie émotionnelle chinoise 74

6.2.2 Taux couverture de la ressource linguistique 75

6.2.3 Application de l'ontologie 76

6.2.4 Méthode insuffisante 77

6.3 Fouille d'opinion avec les inférences annotées manuellement 77

6.3.1 Prédiction automatique de la polarité avec SVM 77

6.3.2 Détection automatique du thème 82

6.4 Fouille d'opinion avec les inférences annotées automatiquement 83

6.4.1 Prédiction de la polarité 84

6.4.2 Détection des thèmes 84

6.5 Conclusion 86

6.1 Introduction

Le traitement des inférences est à la fois indispensable et complexe, mais utile pour la fouille d'opinion pour trois raisons principales.

Premièrement, les thèmes des opinions ne sont pas toujours explicites dans les messages des utilisateurs. Par exemple, "proche de la tour Eiffel" implique une localisation positive de l'hébergement, dans un contexte touristique. La proximité d'une station de métro est encore plus complexe à interpréter. Cette localisation est-elle positive du point de vue de l'accès aux transports, ou négative du fait des nuisances engendrées ?

Deuxièmement, il n'est pas toujours possible de dégager des indices solides pour repérer facilement les phrases contenant des inférences. Les mots porteurs de sentiments ou les formes morpho-syntaxiques ne permettent pas une identification avec certitude.

Le lecteur doit alors mobiliser ses connaissances personnelles du monde et ses compétences linguistiques pour décoder ces inférences. Ce travail se s'avère encore plus complexe pour une machine, même en mobilisant des moyens de traitement automatique des langues.

Troisièmement, pour le domaine spécifique de l'hôtellerie, des inférences peuvent également être faites par le lexique touristique et des noms propres. Le traitement du lexique spécifique fait partie de l'analyse d'inférences.

Dans la première série (logique-pragmatique-lexicale), nous relevons que ces trois types peuvent apparaître indépendamment de tout autre sous-type d'inférence (c'est-à-dire sans combinaison avec le sous-type énonciatif ou discursif ni avec un sous-type de déduction, d'induction ou de rétroduction). La forte proportion d'inférences de type pragmatique (50,3 % des inférences identifiées dans notre corpus) met en évidence l'intérêt de prendre en compte les informations culturelles pour mener des recherches de la fouille d'opinion en chinois. De plus, 29,8 % des inférences sont lexicales. Il est donc nécessaire d'établir un lexique des termes utilisés dans le domaine touristique ou indicateur de sentiments.

Du point de vue de la combinaison des inférences, la présence d'un seul type représente seulement 37,1 % des cas, tandis que la combinaison de trois sous-types affecte jusqu'à 50 % des cas.

Dans notre corpus, toutes les inférences que nous avons identifiées expriment une opinion dont nous pouvons déterminer la polarité. Cela démontre que l'analyse des inférences est une question importante pour la fouille d'opinion en chinois.

Dans ce chapitre, nous démontrons qu'une méthode lexicale n'est pas suffisante pour la fouille d'opinion. Notre méthode avec métadonnées et surtout inférences apporte une amélioration significative. Ensuite, nous comparons les résultats des inférences identifiées et classées manuellement et automatiquement afin de trouver une solution équilibrée entre l'annotation manuelle coûteuse et l'annotation automatique avec erreurs.

6.2 Méthode lexicale pour la fouille d'opinion

6.2.1 *Ressources : ontologie émotionnelle chinoise*

La base de données que nous avons utilisée est l'ontologie du vocabulaire affectif chinois qui est organisée et annotée par des membres du département d'extraction d'information sous la direction du professeur Lin Hongfei de l'Université de technologie de Dalian (Xu et al. 2008). Cette ressource décrit un vocabulaire ou une phrase chinois sous différents angles, tels que les caractéristiques des parties du discours, la catégorie émotionnelle, l'intensité émotionnelle et la polarité.

Le système de classification des sentiments d'ontologie est basé sur les 6 catégories de (Ekman 1999). À partir de cette base, l'ontologie du vocabulaire ajoute la catégorie d'émotion « bon » (好) pour classer plus en détail la signification des émotions. Les

émotions de l'ontologie lexicale finale sont divisées en 7 catégories et 21 sous catégories. L'intensité émotionnelle est divisée en cinq niveaux de 1, 3, 5, 7 et 9, 9 étant le plus intense et 1 le moins intense. Les parties du discours sont divisées en 7 catégories, y compris le nom (nom), le verbe (verbe), l'adjectif (adj), l'adverbe (adv), le mot de réseaux sociaux (nw), l'idiome (idiome), la phrase prépositionnelle (prép). Chaque mot correspond à une polarité sous chaque type d'émotion. 0 signifie neutre, 1 signifie positif, 2 signifie négatif et 3 signifie que ce mot peut être à la fois positif et négatif. Lors de l'annotation, la polarité d'un mot est déterminée conjointement par le sens lui-même et la catégorie d'émotion. Il peut avoir une polarité positive (1) dans certains mots, neutre (0) dans d'autres mots. Cette ontologie émotionnelle chinoise contient un total de 27 466 mots émotionnels. La taille du fichier est de 1,22M. Les exemples sont présentés dans le tableau 6.1.

词语 mot	词性种类 partie du discours	词义数 nombre de sens	词义序号 identifiant de sens	情感分类 catégorie émotionnelle	强度 intensité	极性 polarité
脏乱 désordre	adj	1	1	NN	7	2
肮里肮脏 sale	idiom	1	1	NN	5	2
不干不净 malpropre	idiom	1	1	NN	3	2

Tab. 6.1 : Exemple de l'ontologie

6.2.2 *Taux couverture de la ressource linguistique*

Afin de calculer la distribution de l'ontologie émotionnelle dans notre corpus annoté, nous l'avons divisé en trois parties : opinion avec inférence, opinion sans inférence et opinion incertaine (au sens où les annotateurs ne sont pas sûrs qu'il y a une inférence). Dans le corpus avec inférence, seuls 3,84% des mots sont détectés par cette ontologie, 10% dans le corpus sans inférence et 4% dans le corpus incertain. Après avoir supprimé les doublons, le taux de couverture augmente légèrement (Tableau 6.2).

Par rapport aux 27 466 mots définis dans l'ontologie, le corpus contient seulement 375 mots, soit 9% de la ressource, parmi lesquels " 不错" (pas mal) est le seul mot dont la fréquence est supérieure à 100. Les 15 premiers fréquents mots sont " 安全"(sécurité), " 干净"(propreté), " 热情"(chaleureux), " 不好"(mauvais), " 推荐"(recommander), " 舒服"(confortable), " 不过"(mais), " 舒适"(être à l'aise), " 五星级"(cinq étoiles), " 希望"(espérer), " 便利"(pratique), " 喜欢"(aimer), " 安静"(calme), " 值得"(mériter). Ces mots sont étroitement liés aux commentaires du domaine de l'hôtellerie.

Au niveau des catégories émotionnelles, les opinions positives concernent plus PH (compliment, 820), NN (blâmer, 242), PA (joie, 159), PB (préférence, 141), NL (douter, 140), NE (ennuyé, 127), PG (confiance, 122), PE (Tranquillité d'esprit, 120). Même si les opinons sont jugées positives, il est possible que certains commentaires plutôt positifs mentionnent des points négatifs. Cela explique la présence de la catégorie NN et NE parmi les catégories émotionnelles les plus fréquentes.

En ce qui concerne les parties du discours, le corpus avec inférence comprend un total

de 7 parties du discours : 1 229 adjectifs, 542 verbes, 257 noms, 37 adverbes, 32 idiomes, 25 locutions prépositives, 10 nouveaux mots. La plupart des mots de sentiment sont des adjectifs. Cette ressource linguistique capture plus facilement les opinions exprimées par des adjectifs. Lorsqu'une opinion contient des idiomes ou les néologies, il est plus difficile de déterminer la polarité, car l'utilisation des néologismes varie fortement entre utilisateurs des réseaux sociaux. Dans notre travail, nous considérons que les idiomes et les néologismes sont des inférences lexicales. Pour les opinions qui ne contiennent aucun adjectif de sentiment, il est très probable de trouver une ou plusieurs inférences dans ce type d'opinion. Nous présenterons plus en détail dans la section suivante.

Si une opinion subjective ne contient pas d'inférence et que cette opinion est compréhensible (sans erreurs d'orthographe ou de segmentation), nous observons qu'elle contient fréquemment un mot de sentiment. Mais il est possible que la prédiction de la polarité et de l'intensité de sentiment se réalisent avec l'ordre des mots employés par un utilisateur. Nous constatons que sur les 27 466 mots de la ressource ontologique, seuls 109 sont présents dans les commentaires sans inférence, et 184 sont présents parmi les commentaires pour lesquels la présence d'une inférence est incertaine. Quel que soit le corpus, avec inférence, sans inférence ou incertain, les statistiques du tableau 6.2 montrent qu'une ressource linguistique générale n'est pas suffisante, du moins pour analyser un corpus du domaine spécifique.

Présence d'inférence	fréquence (mots de sentiments) / nombre de mots total	fréquence (mots de sentiments dédoublonnés) / nombre de mots total dédoublonnés
Présence	2132/55510 = 0.0384	375/4165 = 0.0900
Absence	783/7825 = 0.1000	109/798 = 0.1365
Incertitude	380/9497 = 0.0400	184/2149 = 0.0856

Tab. 6.2 : Taux de couverture des mots de sentiments dans le corpus

6.2.3 *Application de l'ontologie*

Après avoir identifié la présence des mots de sentiment dans notre corpus, nous allons ensuite prédire les polarités en appliquant le système de l'émotion dans cette ontologie. À partir du score de polarité et des cinq intensités, nous calculons le score de l'opinion : négative si inférieur à -1, neutre entre -1 et 1, et positive si supérieur à 1.

Parmi 7 168 opinions, 5 069 (71%) ne correspondent à aucun mot de sentiment. Cependant, 82% des 5 060 opinions (soit 4 153) non traitées par l'ontologie émotionnelle contiennent au moins une inférence. Autrement dit, la polarité de ces opinions n'est pas directement exprimée par un mot de sentiment, mais par des phrases avec les inférences. En revanche, pour les opinions avec des mots de sentiment détectés, mais les résultats de prédiction sont incorrects selon le calcul des scores. 758 (79%) de ces opinions contiennent des inférences. Même si une opinion contient des mots de sentiment, il est possible que la polarité de cette opinion soit l'inverse de la polarité des mots de sentiment

détectés. L'interprétation de ce genre d'opinions nécessite l'introduction des inférences. Montré dans la Figure 6.1, l'avantage des inférences est significatif dans l'analyse des sentiments.

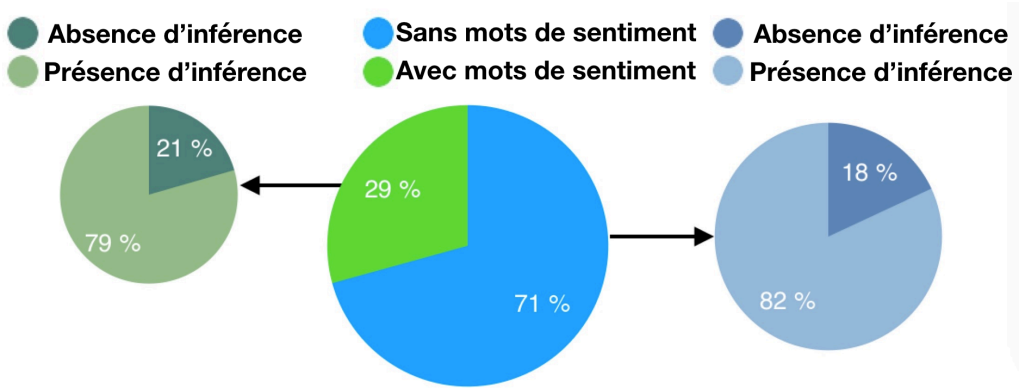


Fig. 6.1 : Proportion d'inférences présentes dans les opinions avec et sans mots de sentiment

6.2.4 *Méthode insuffisante*

Les expériences ci-dessus prouvent que l'application seule d'une ressource linguistique générale n'est pas satisfaisante, au moins pour un corpus spécifique. La fouille d'opinion et l'analyse des sentiments demandent un traitement plus spécifique ou une méthode plus puissante. C'est l'une des raisons pour laquelle nous introduisons le traitement des inférences dans nos travaux.

6.3 Fouille d'opinion avec les inférences annotées manuellement

6.3.1 *Prédiction automatique de la polarité avec SVM*

Après avoir testé l'application de l'ontologie des mots de sentiment, nous introduisons notre solution pour détecter la polarité des opinions de l'hôtellerie, en utilisant SVM (Support Vector Machine). Dans une première étape, nous transformons les opinions en vecteurs. Selon les résultats du chapitre précédent, nous avons prouvé que les vecteurs de Word2Vec avaient le meilleur résultat par rapport aux vecteurs de N-gram, TF-IDF et sac de mots. Dans ce chapitre, nous conservons directement les vecteurs Word2Vec et n'expérimentons plus avec les autres vecteurs. Dans l'étape suivante, nous utilisons seulement les métadonnées comme caractéristiques de l'entraînement. Dans la troisième

étape, nous ajoutons les informations sur les inférences. Les résultats de la détection de polarité sont améliorés à la suite de ces expérimentations.

6.3.1.1 SVM avec les vecteur

Comme SVM demande le nombre de caractéristiques identiques, nous commençons par transformer les textes en vecteurs en utilisant Doc2vec avec une dimension des vecteurs de 250. Même si la longueur des opinions est différente, il est possible d'obtenir la même taille de la matrice. Le résultat est très faible avec une macro F-mesure de 0,17. L'une des raisons pour laquelle la performance du modèle est pauvre est que le corpus provient du même domaine. Les vecteurs de Doc2Vec permettent de mettre en évidence des documents relatifs à différents thèmes. Parce que notre corpus ne couvre que les commentaires d'hôtels à Paris, la fouille d'opinion à l'aide des vecteurs Doc2vec n'est pas possible. Notre corpus est basé sur les commentaires des hôtels à Paris. Il n'est pas commode de distinguer la nuance des opinions par les vecteurs de Doc2Vec.

Polarité	Précision	Rappel	F-mesure
Négative (905)	0,78	0,65	0,71
Neutre (98)	1,00	0,02	0,04
Positive (999)	0,67	0,94	0,78
Inconnue (498)	0,55	0,37	0,44
Macro-moyenne (2500)	0,75	0,50	0,49

Tab. 6.3 : Résultats de la prédiction de la polarité seulement avec les textes en vecteur

Ensuite, nous avons remplacé Doc2Vec par Word2Vec pour mettre en évidence le contexte linguistique des mots. L'entraînement vectoriels est une combinaison de LDA et Word2Vec. La méthode a été présentée dans le chapitre précédent. Ces nouveaux vecteurs sont fournis comme caractéristiques du SVM. Le modèle s'entraîne sur les vecteurs de 868 dimensions. Nous observons que les résultats sont améliorés par rapport à Doc2Vec, avec une macro F-mesure de 0,49 (Tableau 6.3).

6.3.1.2 SVM avec les métadonnées

Déjà utilisées comme des caractéristiques du SVM dans le chapitre précédent, les métadonnées désignent d'une part les informations sur l'hôtel, et d'autre part les catégories morpho-syntaxiques.

Les métadonnées morpho-syntaxiques décrivent chaque opinion par la présence de la négation, sa longueur et les catégories des parties du discours. Il y a 54 catégories. Nous comptons le nombre des étiquettes présentées dans chaque opinion. Les métadonnées d'un hôtel contiennent le score total d'un hôtel, son nombre d'étoiles, sa localisation, le niveau et l'âge de l'utilisateur rédigeant ce commentaire, le score de ce commentaire donné par l'utilisateur.

D'abord, le modèle SVM est produit uniquement à partir de ces métadonnées. Les résultats augmentent avec une F-mesure de 0,77 (Tab. 6.4). Mais le modèle n'a détecté aucune opinion neutre (F-mesure de 0,0) et n'a pas bien prédit les cas inconnus (F-mesure de 0,55). Ensuite, nous avons utilisé les vecteurs et les métadonnées comme des caractéristiques en même temps. Le score de la macro F-mesure a été encore amélioré à 0,58, mais les résultats de la catégorie neutre et inconnue sont toujours modérés. Pour les opinons neutres, le nombre des échantillons est bien moindre par rapport aux autres catégories. Ce déséquilibre conduit à de mauvais résultats. Quant aux opinons inconnues, le faible score de rappel montre que les vecteurs et les métadonnées à eux seuls ne sont pas suffisants pour identifier les opinions dont la polarité est inconnue. Par conséquent, nous ajoutons les inférences à l'étape suivante afin de voir si ces faiblesses peuvent être améliorées.

Polarité	Précision	Rappel	F-mesure
Négative (905)	0,76	0,80	0,78
Neutre (98)	0,00	0,00	0,00
Positive (999)	0,83	0,96	0,89
Inconnue (498)	0,62	0,49	0,55
Macro-moyenne (2500)	0,55	0,56	0,55

Tab. 6.4 : Résultats de la prédiction de la polarité avec les métadonnées

6.3.1.3 SVM avec les inférences

Le modèle a été développé par l'ajout des caractéristiques sur l'indice de la présence d'inférences. Pour chaque opinion, nous l'avons distribué des étiquettes "présence", "absence" et "incertitude" selon les résultats d'annotation manuelle. Parmi 8 713 opinons, 5 582 ont été étiquetées de "présence", 1 584 d'"absence" et 1 547 d'"incertitude". Les caractéristiques du modèle contenaient les vecteurs de Word2Vec, les métadonnées et l'indice de la présence d'inférences.

En ajoutant uniquement l'indice de la présence des inférences, le score de la macro F-mesure du modèle est passé à 0,70. Cette amélioration s'explique notamment par les catégories "inconnue" et "négative". Par rapport aux scores de la F-mesure de l'étape précédente, le score de l'"inconnue" a augmenté de 0,55 à 0,96 et le score de la catégorie "négative" de 0,78 à 0,88 (Tableau 6.5).

Si une opinion est négative ou difficile de déterminer la polarité, il est plus probable que les locuteurs énoncer leurs avis avec des inférences. Surtout pour les mauvaises expériences, au lieu de résumer avec quelques mots explicites, les locuteurs décrivent les expériences en détail. Par exemple, dans ce commentaire, "La porte de la chambre fait en face au lit. Si quelqu'un apporte quelque chose dans la chambre, il aura une vue dégagée sur le lit dès que la porte sera ouverte. Ce n'est pas du tout en sécurité. Une grande partie

de la raison de la réservation de cet hôtel était les photos sur Booking.com. D'après les photos, l'hôtel dispose d'un toit qui permet d'avoir une vue sur la tour Eiffel et l'Arc de Triomphe. Mais après l'enregistrement à l'hôtel, quand je demandais au personnel, je savais que le toit était la terrasse de la chambre en dernier étage, il n'était pas ouvert à tous les clients." ("房门是正对着床的、如果送东西到客房、一开门床上便一览无遗、完全没有安全感，一开始订这家酒店很大一部分原因是看了 Booking.com 上的该酒店照片，从照片显示来看，该酒店有一个天台可以直接□望到铁塔和凯旋门，但是到达酒店办理住之后，询问员工了才知道那个天台是顶层房间的露台，并不是对所有住客开放的。") C'est vraiment en interprétation ces descriptions que nous pouvons déterminer la polarité négative. Cette interprétation adopte inévitablement des inférences, par exemple l'inférence logique pour pouvoir comprendre le flux de texte, l'inférence discursive pour déduire les énonces qui manquent de mots sentimentaux, ou même l'inférence pragmatique si un locuteur mentionne un point précis dans le cadre d'un hôtel ou revêtit les caractéristiques d'une communauté culturelle (la tour Eiffel et l'Arc de Triomphe).

Même si l'indice de présence d'inférences contribue globalement à la prédiction de la polarité, aucun cas de la catégorie "neutre" n'a été correctement détecté. Premièrement, le nombre total de cette catégorie est très petit par rapport aux autres catégories (98 parmi 2 500). Deuxièmement, la frontière entre neutre et positif/négatif est floue. Comme nous n'avons pas spécifié les critères de jugement, chaque annotateur a sa propre interprétation de ce qu'est une inférence neutre. Il est probable que les annotateurs auront qualifié une inférence de neutre lorsqu'ils ne pouvaient pas prendre de décision, conduisant à une catégorie complexe pour des systèmes automatiques. Il manque possiblement d'une distinction significative de la présence d'inférences entre les opinions neutres et les autres.

À l'exception de la catégorie neutre, les scores prouvent que l'ajout des étiquettes de présence d'inférences servit à la prédiction de la polarité avec une F-mesure supérieure à 0,90. Dans l'étape suivante, nous allons continuer à améliorer les résultats sur chaque type d'inférence.

Polarité	Précision	Rappel	F-mesure
Négative (905)	0,89	0,88	0,90
Neutre (98)	1,00	0,03	0,06
Positive (999)	0,85	0,97	0,90
Inconnue (498)	1,00	0,93	0,96
Macro-moyenne (2500)	0,93	0,70	**0,70**

Tab. 6.5 : Résultats de la prédiction de la polarité avec les vecteurs, les métadonnées et les indices de la présence des inférences

6.3.1.4 SVM avec tous les paramètres en ajoutant les 5 types d'inférence

Sur la base des caractéristiques utilisées dans les étapes précédentes, nous avons ensuite ajouté les cinq types d'inférence comme caractéristiques. En plus d'indiquer si une opinion contient des inférences, les types d'inférence correspondants ont été précisés. Il est possible, voire vraisemblable qu'une opinion contienne à la fois plusieurs types d'inférence. Le cas où un seul type d'inférence est présent dans une opinion ne représente que 23,8 % des opinions totales, soit 2 526 parmi 10 609 (voir Tableau 6.6). En particulier les inférences énonciatives et discursives, elles coexistent avec les autres types.

Afin de représenter le cas où plusieurs inférences sont présentes dans une opinion, nous avons ajouté 5 nouvelles caractéristiques correspondant à 5 types. Les valeurs de ces caractéristiques sont 1 et 0 indiquant la présence ou l'absence de ce type. Les 5 types ne sont pas exclusifs.

	Logique	Pragmatique	Lexicale	Énonciative	Discursive	Total	Pourcentage
Apparition seul	709	292	1521	4	0	2526	23.8%
Nombre total	3587	2947	3374	547	154	10609	100%

Tab. 6.6 : Nombre de chaque type d'inférence d'une apparition tout seul et total

Le nouveau score de la F-mesure est le meilleur pour le moment. En d'autres termes, l'ensemble de tous les paramètres, en particulier la partie des inférences, améliore les performances du modèle. Il est capable de prédire les opinions positives et négatives avec une F-mesure minimale de0,90 et de repérer les opinions avec une polarité assez floue pour que les travaux futurs puissent les analyser. En revanche, ce modèle n'a toujours pas réussi à traiter la catégorie neutre avec une précision et un rappel faibles.

Polarité	Précision	Rappel	F-mesure
Négative (905)	0,90	0,90	0,90
Neutre (98)	0,00	0,03	0,06
Positive (999)	0,87	0,97	0,91
Inconnue (498)	0,99	0,94	0,97
Macro-moyenne (2500)	0,94	0,71	**0,71**

Tab. 6.7 : Résultats de la prédiction de la polarité avec les vecteurs, les métadonnées, les indices de la présence des inférences et les cinq types d'inférence

6.3.1.5 Résumé de l'expérimentation de la prédiction de la polarité

L'expérimentation de la prédiction de la polarité contient 5 étapes :

1. une simple application de l'ontologie émotionnelle chinoise
2. SVM avec une accumulation des caractéristiques
   a) texte en vecteur

b) texte en vecteur et métadonnées
c) texte en vecteur, métadonnées et présence d'inférence
d) texte en vecteur, métadonnées, présence d'inférence et 5 types d'inférence

La performance du modèle s'améliore tout au long des 5 étapes. Le score de la macro F-mesure augmentait de 0.17 à 0.71.

Fig. 6.2 : Procédures de l'expérimentation de la prédiction de la polarité

En fonction de l'évolution des caractéristiques, nous avons dessiné leurs taux d'apprentissage correspondants. Les 5 courbes du bas vers le haut présentent une expérience de l'application de l'ontologie jusqu'à l'introduction des 5 types d'inférences dans les paramètres. Pour chaque étape, les caractéristiques sont cumulées de l'étape précédente.

Les courbes se divisent en 3 classes. La courbe bleu clair représente la performance de l'application de l'ontologie émotionnelle chinoise qui se trouve en bas du diagramme. La deuxième classe qui est au milieu du diagramme représente une première amélioration du modèle en ajoutant les vecteurs et les métadonnées, y compris morpho-syntaxiques et métadonnées de l'hôtel. La troisième classe en haut du diagramme regroupe 2 courbes : la présence des inférences, les 5 types d'inférence. Il est évident que les informations sur l'inférence jouent un rôle important dans la prédiction de la polarité.

6.3.2 *Détection automatique du thème*

À partir de nos expériences de prédiction de polarité, nous avons prouvé que les caractéristiques des inférences étaient appropriées. Dans cette section, nous appliquons le même processus pour la détection automatique des thèmes. C'est-à-dire le numéro 2 de la figure 6.2. Les résultats sont présentés dans le tableau 6.8.

	V			V+M			V+M+Inf			V+M+Inf+Type		
Thèmes	P	R	F	P	R	F	P	R	F	P	R	F
Clientèle (5)	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Équipement (788)	0,74	0,85	0,80	0,72	0,93	0,81	0,72	0,94	0,81	0,72	0,94	0,82
Général (156)	0,57	0,19	0,29	0,73	0,29	0,41	0,70	0,31	0,43	0,69	0,33	0,44
Localisation (454)	0,84	0,92	0,88	0,85	0,94	0,89	0,87	0,94	0,91	0,87	0,94	0,90
Mixte (170)	0,50	0,61	0,55	0,51	0,66	0,58	0,51	0,66	0,58	0,51	0,66	0,57
Personnel (164)	0,72	0,52	0,60	0,80	0,58	0,67	0,83	0,60	0,70	0,83	0,59	0,69
Prix (41)	0,50	0,24	0,33	0,53	0,24	0,33	0,53	0,24	0,33	0,53	0,24	0,33
Propreté (33)	0,43	0,09	0,15	0,50	0,12	0,20	0,50	0,12	0,20	0,57	0,12	0,20
QualitéService (160)	0,86	0,04	0,07	0,65	0,19	0,30	0,69	0,22	0,33	0,67	0,24	0,35
Sécurité (33)	0,50	0,15	0,23	0,42	0,24	0,31	0,47	0,24	0,32	0,50	0,24	0,33
Macro-moyenne (2500)	0,56	0,39	0,40	0,60	0,46	0,49	0,62	0,47	0,51	0,62	0,48	0,51

Tab. 6.8 : Résultats comparatifs de la détection des thèmes selon des caractéristiques variantes (V : vecteurs, M : métadonnées, Inf : présence d'inférence, Type : type d'inférence)

6.4 Fouille d'opinion avec les inférences annotées automatiquement

Dans cette section, au lieu d'utiliser les inférences annotées manuellement, nous utilisons les inférences automatiquement identifiées et classées du chapitre précédent. Cette étape a un double objectif :

1. En tenant compte des erreurs des expériences automatiques des inférences, les inférences sont-elles toujours bénéfiques pour la fouille d'opinion ?
2. En comparant les résultats des expériences avec les différentes caractéristiques, nous envisageons de trouver une solution équilibrée entre l'annotation manuelle et automatique.

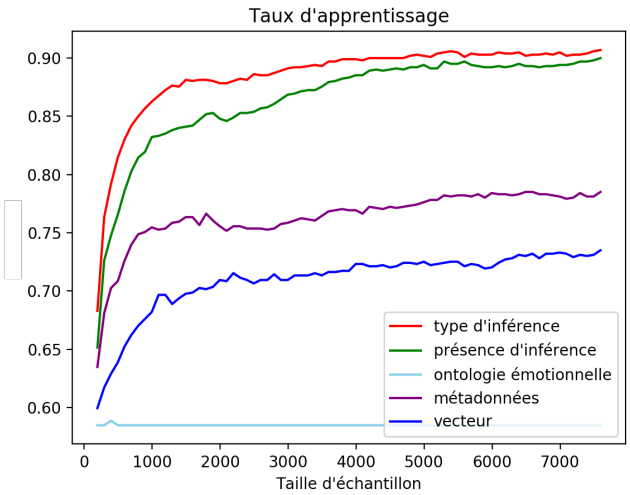


Fig. 6.3 : Taux d'apprentissage comparatifs de la prédiction de la polarité en utilisant les différents paramètres

6.4.1 *Prédiction de la polarité*

Parmi les caractéristiques, les vecteurs et les métadonnées n'ont pas été modifiés. Nous avons remplacé la présence d'inférence et les types d'inférence par les résultats de l'annotation automatique du chapitre précédent.

En comparant les résultats du tableau 6.9, les résultats de l'annotation manuelle sont évidemment meilleurs que ceux de l'annotation automatique, car la classification automatique des inférences n'est pas correcte à 100%. Les erreurs de l'annotation automatique sont présentées dans la prédiction de la polarité. D'une part, la faiblesse des résultats automatiques nous oblige à améliorer le modèle de classification automatique ou à trouver un équilibre entre l'annotation manuelle et automatique. D'autre part, l'écart entre les résultats manuels et automatiques confirme à nouveau l'importance des inférences dans la fouille d'opinion. Plus les inférences données sont précises, meilleurs sont les résultats obtenus.

6.4.2 *Détection des thèmes*

Quant à la détection de thème, nous avons également remplacé les inférences annotées manuellement par les résultats du modèle d'apprentissage automatique. Le score de la macro F-mesure descend de 0,51 à 0,43. D'un point de vue global, les inférences fonctionnent moins bien pour la tâche de la prédiction des thèmes. De plus, les erreurs menées par les résultats de l'apprentissage automatique aggravent la performance de la détection

	V + M + Présence Inf						V + M + Présence Inf + Type					
	Ann. manuelle			Ann. auto			Ann. manuelle			Ann. auto		
Polarité	P	R	F	P	R	F	P	R	F	P	R	F
Négative (905)	0,89	0,88	0,88	0,79	0,77	0,78	0,90	0,90	0,90	0,79	0,75	0,77
Neutre (98)	1,00	0,03	0,06	1,00	0,03	0,06	1,00	0,03	0,06	1,00	0,03	0,06
Positive (999)	0,85	0,97	0,90	0,80	0,95	0,87	0,87	0,97	0,91	0,82	0,94	0,87
Inconnue (498)	1,00	0,93	0,96	0,64	0,57	0,60	0,99	0,94	0,97	0,61	0,60	0,61
Macro-moyenne (2500)	0,93	0,70	0,70	0,81	0,58	0,58	0,94	0,71	0,71	0,80	0,58	0,58
Moyenne pondérée (2500)	0,90	0,89	0,88	0,78	0,77	0,75	0,91	0,90	0,89	0,77	0,77	0,75

Tab. 6.9 : Comparaison des résultats de la polarité avec l'annotation manuelle et automatique (Ann. auto : annotation automatique, Ann. manuelle : annotation manuelle)

des thèmes. D'autres expériences sur le thème feront partie de nos futurs travaux.

6.5 Conclusion

Dans ce chapitre, nous avons d'abord prouvé qu'une simple application des mots de sentiment ne suffisait pas pour la fouille d'opinion. Ensuite, en appliquant les inférences annotées manuellement dans la fouille d'opinion, nous avons aperçu que l'introduction des inférences aidait effectivement à la prédiction de la polarité et du thème, même si cet effet soit moins évident dans la deuxième tâche. En revanche, les inférences annotées automatiquement ont entraîné des erreurs qui ont rendu le modèle moins efficace. Par conséquent, nous proposons deux solutions pour améliorer les résultats :

- Comme le résultat manuel est meilleur, mais que l'annotation manuelle est assez coûteuse en terme de temps et de financier, nous souhaiterions identifier le bon équilibre entre l'annotation manuelle et automatique. La situation envisagée est de combiner le travail automatique et une partie du travail manuel afin d'obtenir un meilleur résultat.
- On peut considérer l'ensemble de la prédiction de la polarité et la classification des inférences comme un pipeline. Les vecteurs, les métadonnées et les inférences sont tous une partie du pipeline. Afin d'obtenir un meilleur résultat final, il est indispensable d'améliorer chaque chaînon. Autrement dit, pour que le résultat de la prédiction de la polarité soit le meilleur, chaque modèle du pipeline soit le meilleur.

Nous envisageons de trouver un équilibre entre l'annotation manuelle et automatique et d'analyser les relations entre inférences et certains phénomènes linguistiques dans le chapitre suivant.

Chapitre 7

Ajustement et application du modèle

Contents

7.1 Introduction 87

7.2 Équilibre entre l'annotation manuelle et automatique des inférences 87

7.3 Analyses linguistiques des résultats 89

7.3.1 Discussion des résultats par rapport aux théories linguistiques sur les inférences 90

7.3.2 Phénomène linguistique observé dans les résultats 95

7.4 Conclusion 98

7.1 Introduction

L'inférence et ses types se sont avérés efficaces pour la fouille d'opinion selon les expériences précédentes. Dans ce chapitre, nous nous concentrons sur l'amélioration du modèle. Les premières expériences visent à trouver une meilleure solution en tenant compte du prix et de la qualité de l'annotation manuelle. Ensuite, le modèle est ajusté en prenant en compte trois aspects : le corpus, les caractéristiques et les paramètres. Enfin, nous envisageons d'appliquer le modèle à un autre corpus du même domaine (commentaires d'hôtels à Paris) mais provenant d'une autre source (Mafengwo.cn).

7.2 Équilibre entre l'annotation manuelle et automatique des inférences

L'annotation manuelle des inférences a été effectuée par trois annotatrices pendant un mois (environ 80 heures). Le corpus annoté contient 38 000 tokens. Compte tenu du coût de l'annotation manuelle, on aimerait établir des modèles automatiques pour le travail d'annotation. Cependant, comme nous l'avons vu au chapitre 6, les résultats ne sont pas convaincants. Par conséquent, nous intégrons une partie de l'annotation manuelle dans l'apprentissage automatique pour prédire la polarité.

En raison du coût élevé de l'annotation manuelle, nous présentons dans cette section les expériences que nous avons réalisées pour mesurer l'utilité des informations manuelles dans les prédictions automatiques si les modèles ne sont jamais fournis qu'avec une partie de ces annotations manuelles. L'objectif est de déterminer quels sont les types d'information devraient faire l'objet d'un travail d'annotation manuelle pour optimiser les performances du système tout en réduisant les coûts humains.

7.2.0.1 Expériences

Dans toutes nos expériences, nous avons conservé la même répartition entre les corpus d'entraînement et d'évaluation, ainsi que les vecteurs et les métadonnées. En revanche, nous avons ajusté les 6 classes : "présence d'inférence" ainsi que les types d'inférence "logique", "pragmatique", "lexicale", "énonciative" et "discursive".

Afin de mesurer l'importance des types d'informations fournies par les humains les plus utiles aux modèles, nous avons réalisé une série d'expériences basées soit sur la combinaison de deux classes d'informations, soit sur une seule classe.

De plus, chaque modèle devait contenir les 6 catégories. En triant les scores de la macro F-mesure et de la F-mesure pondérée, nous avons répertorié les 5 meilleurs modèles respectivement pour une seule catégorie manuelle sélectionnée et deux catégories manuelles sélectionnées (Tableau 7.1).

	Annotation manuelle						Macro-moyenne			Moyenne pondérée		
	Présence	Logique	Pragmatique	Lexicale	Énonciative	Discursive	P	R	F	P	R	F
Plafond	**X**	**X**	**X**	**X**	**X**	**X**	**0,94**	**0,71**	**0,71**	**0,91**	**0,90**	**0,89**
2-1	X	X					0,94	0,71	0,71	0,90	0,90	0,88
2-2	X		X				0,93	0,70	0,70	0,90	0,89	0,88
2-3	X			X			0,93	0,70	0,70	0,90	0,89	0,88
2-4		X	X				0,89	0,65	0,66	0,86	0,85	0,83
2-5		X		X			0,89	0,65	0,65	0,85	0,85	0,83
1-1	X						0,93	0,70	0,70	0,90	0,89	0,87
1-2		X					0,88	0,64	0,65	0,85	0,84	0,82
1-3			X				0,85	0,62	0,62	0,81	0,81	0,79
1-4				X			0,82	0,60	0,59	0,79	0,78	0,77
1-5					X		0,81	0,59	0,59	0,78	0,78	0,76
Plancher							**0,80**	**0,58**	**0,58**	**0,77**	**0,77**	**0,75**

Tab. 7.1 : Meilleures combinaisons entre l'annotation manuelle et automatique

7.2.0.2 Analyse des résultats

Les 10 meilleurs modèles sont indiqués dans le tableau 7.1. Le "plafond" représente le modèle en utilisant uniquement les inférences annotées manuellement. À l'inverse, le "plancher" représente le modèle entièrement automatique. Chaque modèle se voit distribuer un identifiant à deux chiffres. Le premier numéro fait référence au nombre de catégories d'informations utilisées, le second numéro est un numéro de série des expériences. Par exemple, "2-1" signifie que ce modèle utilise deux catégories manuelles et

se classe en premier. “1-2” représente une catégorie utilisée. Parmi les modèles avec une catégorie, ce modèle est au deuxième rang.

Présence d’inférence Les meilleurs résultats sont majoritairement obtenus par des modèles contenant deux catégories manuelles. La seule exception est la présence d’inférence (identifiant 1-1 dans le tableau 7.1). En plus de la catégorie “présence d’inférence”, les cinq types sont annotés par le modèle automatique. Le score est déjà supérieur aux modèles 2-4 et 2-5 qui ne contiennent pas la présence d’inférence manuelle. De plus, les trois autres modèles avec deux catégories contiennent tous la présence d’inférence. Cela prouve que la qualité de l’identification des inférences est très importante pour prédire la polarité. Nous avons ainsi reconstruit un modèle en utilisant la présence d’inférence annotée manuellement et en omettant les cinq types. Ce modèle avait des caractéristiques telles que des vecteurs, des métadonnées et des indices de présence d’inférence. Son score de la macro F-mesure et de la F-mesure pondérée sont de 0,70 et 0,88. Par rapport au modèle 1-1, les résultats sont presque les mêmes. Dans le cas où les types d’inférences ne sont pas demandés, l’une des méthodes les plus efficaces est d’identifier manuellement la présence d’inférence dans chaque texte et d’omettre tout le traitement des types.

Types d’inférence Dans le tableau 7.1, parmi les 10 meilleurs modèles, l’inférence logique se produit 5 fois et l’inférence pragmatique apparaît 4 fois. Comparées aux trois autres types, elles jouent un rôle plus important dans le modèle de prédiction de la polarité. Cependant, la différence entre le modèle 2-5 et le modèle 1-2 n’est pas significative. En d’autres terme, le modèle avec les inférences logiques et pragmatiques annotées manuellement ne s’est pas beaucoup amélioré par rapport au modèle avec l’inférence logique seule. Par conséquent, la deuxième solution consiste à reconnaître manuellement l’inférence logique dans les commentaires. Les autres catégories concernant l’inférence peuvent être classifiées automatiquement.

En conclure, nous observons surtout que fournir des informations sur la présence d’une inférence permet d’obtenir les meilleurs résultats. La fourniture d’une deuxième catégorie d’informations est plus marginale : présence + logique ou présence + pragmatique ou présence + lexicale obtiennent quasiment les mêmes scores (les différences ne sont d’ailleurs peut-être pas significatives). Cette observation est confirmée lorsqu’une seule catégorie est fournie, c’est avec les informations de présence que l’on obtient les meilleurs résultats. Par contre, dès que la présence est supprimée, les résultats baissent : logique + pragmatique ou logique + lexicale, ou une seule catégorie parmi les autres.

7.3 Analyses linguistiques des résultats

En comparant les résultats de la fouille d’opinion, le modèle qui utilise la classification manuelle est meilleur que celui qui utilise la classification automatique. Nous

pouvons l'interpréter comme plus les résultats de la classification des inférences sont précis, meilleure est la performance de la recherche d'opinion. Nous revenons donc à une analyse linguistique des inférences qui ne sont pas correctement identifiées par le modèle. Ces analyses sont à la fois un retour d'expérience du corpus et des pistes de travaux futurs.

7.3.1 *Discussion des résultats par rapport aux théories linguistiques sur les inférences*

Nous sommes d'abord revenus sur l'état de l'art des inférences. Nous avons catégorisé les inférences sur deux aspects :

7.3.1.1 Réalisation sémantique

Au niveau sémantique, nous avons distingué les inférences logiques des inférences pragmatiques (Kintsch 1998 ; Rossi et Campion 1999 ; Dufaye 2001 ; Martin 2004 ; Duchêne 2008 ; Bouquiaux et Leclercq 2017). Cet aspect met l'accent sur la manière d'accéder aux sens exprimés dans l'inférence. Nous distinguons un raisonnement logique des contextes textuels d'un autre raisonnement pragmatique basé sur l'ensemble de connaissances personnelles. Cette distinction nous aide d'une part à reconnaître les inférences selon les langues et les cultures, d'autre part à comprendre le schéma narratif du texte.

7.3.1.2 Mode de production

En termes de mode de production, nous avons distingué les inférences lexicales, énonciatives et discursives (Doucy et Massoussi 2012).

À ce niveau, nous nous intéressons à la manière dont l'émetteur du message a produit l'inférence. Selon le type d'inférence qui nous intéresse, nous observons des variations dans la taille des portions textuelles intégrant une inférence, et dans la richesse sémantique du contenu. Cette observation permet une analyse sur deux points :

- repérer des mots sentimentaux spécifiques dans le domaine de l'hôtellerie. La signification étymologique des mots peut être neutre. Pourtant, dans les commentaires d'hôtels, ces mots expriment une polarité. Le dictionnaire traditionnel des mots de sentiment n'est pas toujours suffisant pour la fouille d'opinion, en particulier dans un domaine spécifique. La définition de l'inférence lexicale permet de localiser ces types de mots.
- traiter de longs commentaires. Dans notre corpus, nous avons très souvent observé des commentaires composés de plusieurs phrases narratives. Ils contiennent peu de mots de sentiment. L'expression de l'opinion se réalise à travers toute la narration. Les techniques existantes de fouille d'opinion ne sont pas adéquates pour traiter cette catégorie. De ce fait, nous avons proposé d'analyser l'inférence énonciative et discursive.

Le résultat montre que la performance de la fouille d'opinion est améliorée après l'application de cette classification des inférences. Le score de F-mesure passe de 0,75 à 0,89.

	Commentaire	Traduction	Type d'inférence	Polarité
1	竟号称五星级	Il a osé prendre un hôtel de 5 étoiles	logique, pragmatique	négative
2	真不明白猫头鹰上的高分怎么来的	Je ne comprends pas vraiment d'où viennent les scores élevés sur TripAdvisor.	logique, pragmatique	négative
3	我带着三个老人在马路边等， 在预订的时间近一小时后 才有人拿钥匙来开门。	J'ai attendu au bord de la route avec trois personnes âgées, c'était presque une heure après l'heure prévue que quelqu'un a apporté la clé pour ouvrir la porte.	logique, pragmatique, énonciative, discursive	négative
4	电梯从一楼才有	L'ascenseur n'est disponible qu'à partir du premier étage.	pragmatique, lexicale, énonciative	négative
5	预订上下楼，后来发现是地下室	J'ai réservé un loft, mais j'ai découvert plus tard que c'était le sous-sol.	logique, pragmatique	négative
6	房间真的有 20 平米吗	La pièce fait-elle vraiment 20 mètres carrés ?	logique, énonciative	négative
7	以为可以看到完整凯旋门， 结果需要探出头才看的到一个边边角角	Je pensais pouvoir voir l'Arc de Triomphe en entier, mais j'ai dû sortir la tête pour voir un coin.	logique, pragmatique, lexicale, énonciative, discursive	négative
8	大部分景点不用换线或只换一次	Il n'est pas nécessaire de changer de ligne (métro) pour accéder aux sites pittoresques, ou alors il n'y a qu'un seul changement.	pragmatique, énonciative	positive
9	如果到达时不是前台开放时间， 可在门外的机器输入预订信息， 拿到房卡，直接进入。	Si l'heure d'arrivée ne correspond pas aux heures d'ouverture de la réception, vous pouvez saisir les informations de réservation à la machine devant la porte, obtenir la carte de chambre et entrer directement.	logique, pragmatique, énonciative, discursive	positive
10	晚上 11 点出门都不担心。	Ne vous inquiétez pas de sortir à 11 heures du soir.	pragmatique	positive

Tab. 7.2 : Exemple de polarité correctement détectée avec la classification des inférences

Le tableau 7.2 donne des exemples de commentaires dont la polarité a été correctement détectée après l'intégration des inférences dans la fouille d'opinion. En calculant les commentaires dont la polarité a été corrigée après l'application des inférences, le nombre de commentaires négatifs est 6 fois plus de commentaires positifs. L'effet des inférences est le plus évident pour les commentaires négatifs. Cependant, il y a encore des erreurs. Nous observons que le nombre d'erreurs est plus important sur les commentaires négatifs (173) que sur les commentaires positifs (59).

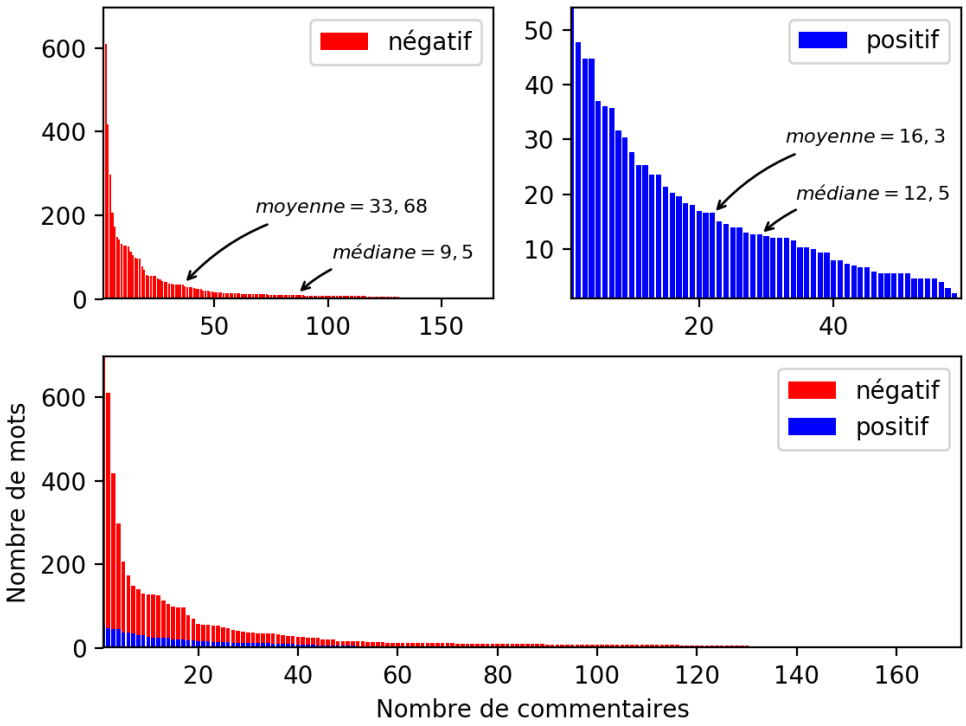


Fig. 7.1 : Nombre de commentaires positifs en fonction du nombre de mots

Nous avons comparé la distribution de longueur des commentaires de deux polarités dans la figure 7.1. La longueur moyenne est de 16,3 pour des commentaires positifs, 33,68 pour des commentaires négatifs. Mais le point médian des commentaires positifs (12,5) est plus élevé que celui des commentaires négatifs (9,5). Nous avons découvert deux caractéristiques des commentaires négatifs avec une polarité mal détectée :

1. La longueur des commentaires est soit très longue, soit très courte.
2. Le nombre de courts commentaires est considérable. Ces commentaires courts sont

souvent des phrases ou des syntagmes simples.

Sur la base de ces réflexions, nous revenons au troisième aspect de l'inférence dans la section suivante.

7.3.1.3 Direction textuelle et mécanisme mental

Dans l'état de l'art, une autre dimension majeure de l'inférence est la prise en compte de la direction et du nombre d'éléments dans un texte. La direction renvoie à la fois à l'ordre des informations présentes dans le texte pour comprendre l'inférence, et d'autre part à la procédure de raisonnement mise en œuvre.

Le nombre désigne le nombre des prémisses disposées par lesquelles la conclusion est atteinte. Cette dimension de l'inférence fait une distinction parmi l'inférence médiate, l'inférence immédiate, l'induction, la déduction et la rétroduction (Peirce 1958 ; Deledalle et Peirce 1994 ; Kintsch 1998 ; Van den Broek et al. 1999 ; Rossi et Campion 1999 ; Dufaye 2001 ; Khemlani et al. 2012).

Nous envisageons d'introduire le concept d'ordre dans les travaux futurs pour deux raisons :

1. Comme nous l'avons analysé dans la section précédente, la classification de cinq types améliore la fouille d'opinion, mais il y a quelques erreurs. Si nous analysons plus concrètement comment une inférence est produite et son ordre de production, nous pouvons trouver une nouvelle solution aux erreurs actuelles. La première hypothèse est qu'avec la précision du nombre de prémisses, il est possible pour le modèle de séparer les commentaires avec une phrase simple (une prémisse) des commentaires avec des phrases complexes (plus de deux prémisses). Cela correspond à la distinction entre l'inférence médiate et immédiate.
2. Dans le cas de l'annotation manuelle, nous avons déjà pris en compte la relation des syntagmes ou des phrases à l'intérieur d'un commentaire. Outre les annotations concernant l'ensemble d'un commentaire, il a été segmenté en syntagme par virgule et point-virgule, segmenté en phrase par point. Chaque élément d'un commentaire a également été annoté de la même manière. Par conséquent, nos travaux futurs ont la possibilité d'analyser le flux des éléments d'un commentaire et de localiser des inférences à différents niveaux textuels.

Par exemple, "和前台说定 2 天 (1)，结果前台英语没听懂只给定了一天 (2)，第二天被赶出酒店 (3)，晚上 7 点满巴黎大街再找新酒店 (4)。" ("J'ai dit à la réception de réserver pour 2 jours, mais la réception ne comprenait pas l'anglais et n'avait réservé qu'un jour, et on a été expulsé de l'hôtel le lendemain, et j'ai dû rechercher un nouvel hôtel partout dans la rue de Paris à 19h.") Ce commentaire n'a pas été correctement détecté comme négatif par le modèle. Même si l'inférence logique montre la séquence de chaque syntagme, les inférences ne sont pas toujours en position finale. Il est possible qu'un client commence immédiatement en prononçant le syntagme 4, puis en expliquant les raisons avec les syntagmes 1, 2 et 3. Cependant, nous ne pouvons pas identifier l'emplacement exact de la production de l'inférence. Si des inférences déductives, inductives

et rétroductives sont appliquées dans cet exemple, nous pouvons désigner l'ensemble de la production de l'inférence de syntagme 1 à 4. La production de l'inférence se trouve à la fin du syntagme 4. De plus, puisque ce commentaire contient plusieurs phrases, le nombre des prémisses est supérieur à 1. Il contient également l'inférence médiate.

7.3.2 *Phénomène linguistique observé dans les résultats*

7.3.2.1 Intensifieurs

Les intensifieurs font partie des mots grammaticaux en chinois. Ils correspondent à des adverbes de degré. La première caractéristique que nous avons observée est la présence d'intensifieurs dans des inférences non identifiées dont 65,73% contiennent des intensifieurs. Lorsqu'une opinion contient à la fois des inférences et des intensifieurs, la fonction des intensifieurs n'est pas simplement de "renforcer le contexte émotionnel du mot" (Bhaskar et al. 2014), mais aussi d'apaiser les tensions ou d'exprimer des espoirs.

Atténuation des attentes Premièrement, les intensifieurs atténuent la polarité négative. Par exemple, "价格比较便宜, 就不能要求太高了." (Le prix est assez avantageux, alors ne demandez pas trop.) Les intensifieurs "assez" et "trop" atténuent la tension, mais expriment également une opinion insatisfaite.

Espérance Deuxièmement, les intensifieurs aident à exprimer les attentes. Par exemple, "价格再便宜一点就更好" (Si le prix était plus bas, ce serait encore mieux.) L'intensifieur "更" (plus, encore) indique que le prix est le point relativement faible de cet hôtel. Il s'exprime comme une attente des clients. Ce genre d'opinion contient aussi des inférences logiques et énonciatives.

Degré Tao (2014) propose une catégorisation des intensifieurs en 5 degrés : degré superlatif, degré excessif, degré élevé, degré comparatif, degré de suspicion.

- Le degré superlatif qui représente un degré extrême utilise des adverbes comme "最" (le plus), "至" (extrêmement), "无比" (incomparable), "极" (le plus haut degré), "顶" (sommet) , etc. Si le degré superlatif est présent dans un commentaire avec inférences, la force est extrême. Toutes les opinions dans ce commentaire expriment la même polarité, qu'elle soit très négative soit très positive. Il est rare que ce genre de commentaire contienne des opinions de polarité différente. Par exemple, "电梯极其狭窄，房间极小，被子有汗味，不把窗户关上楼下摩托车的声音让人无法入睡。" (L'ascenseur est extrêmement étroit, la pièce est extrêmement petite, la couette sent la sueur et le bruit de la moto en bas m'empêche de dormir sans fermer les fenêtres.)
- Le degré excessif représente un degré qui dépasse la limite normale acceptable par l'être humain. Nous utilisons souvent "太" (trop) ou "过于" (outre mesure) pour exprimer ce degré. En général, des opinions avec un degré excessif sont négatives.

- Le degré élevé est une catégorie vague. “很” (très), “挺” (très), “非常”(extraordinairement), “格外” (spécialement), “特别” (particulièrement), “十分” (totalement), “相当”(relativement) sont tous des adverbes de degré élevé. Par exemple, “电梯非常小, 大型的行李根本进不去。” (L’ascenseur est très petit, les grandes valises ne rentrent pas du tout.) Même si nous combinons les différents intensifieurs de cette catégorie avec le même adjectif, le degré est différent. Il faut le contexte pour distinguer la nuance entre eux.
- Le degré comparatif est légèrement inférieur à la catégorie précédente. Les intensifeurs représentatifs sont “较” (comparativement), “比较” (relativement), “较为” (assez). Les commentaires contiennent à la fois des intensifieurs comparatifs et des inférences sont observés des cooccurrences des opinions positives et négatives. Par exemple, “wifi 连接过于繁琐，经常断，总要重新输入邮箱注册。” (La connexion Wi-Fi est trop lourde, souvent déconnectée, demande toujours à nouveau l’enregistrement des é-mails.)
- Le degré de suspicion est au niveau le plus bas parmi les 5 catégories. Il contient des intensifieurs comme “稍” (légèrement), “略” (légèrement), “有点” (un peu), “微” (faiblement), “稍微” (un peu), etc. De même que les intensifieurs comparatifs, ils sont utilisés pour atténuer les termes. Cependant, les commentaires avec ces intensifieurs contiennent souvent deux polarités. La polarité de l’ensemble du commentaire est plus proche du neutre. Par exemple, ” 唯一的遗憾就是离各大景点稍微远了一些，但我们并不介意。” (Le seul regret est que c’est un peu loin des principales attractions, mais cela ne nous dérange pas.)

La présence des inférences ayant un rapport avec des intensifieurs, nous envisageons d’établir de nouvelles caractéristiques du modèle autour des intensifieurs dans nos travaux futurs.

7.3.2.2 Connecteur

Étant une catégorie de mots grammaticaux, les connecteurs sont également liés à l’apparition des inférences. Parmi 5402 commentaires avec inférences, 3592 (soit 66,5%) contiennent au moins un connecteur. En particulier, 82,6% des commentaires dont la polarité n’est pas correctement détectée contiennent des connecteurs.

Certains connecteurs d’opposition comme “但是” (mais, pourtant), “然而” (cependant) sont surreprésentés dans les inférences logiques. Ces connecteurs connectent deux opinions de polarité opposée. Cependant, les opinions antérieures et postérieures des connecteurs d’addition (“和” (et),“且” (de plus)) ou alternatives (“或” (soit), “或者” (soit)) sont généralement identiques. L’interprétation du message est partiellement construite à partir des connecteurs.

7.3.2.3 Ironie

L'ironie est un phénomène linguistique et une figure de rhétorique par laquelle on dit le contraire de ce que l'on veut exprimer (Xiang et al. 2020).

La plupart des études sur l'ironie chinoise se concentrent d'une part sur la recherche de caractéristiques et de modèles linguistiques liés à l'ironie qui peuvent être utilisés pour identifier l'ironie (Y.-j. Tang et H.-H. Chen 2014 ; Xiang et al. 2020), et d'autre part sur la détection automatique de l'ironie (Y.-H. Huang et al. 2017 ; A. Li et C.-r. Huang 2019 ; Pan et al. 2020).

Nous avons adapté et appliqué les modèles proposés par Y.-j. Tang et H.-H. Chen (2014). Par exemple :

- modèle : 可以 / 能再 + adjectif négatif + 一点 (Il n'y a rien de mal à être plus + adjectif négatif)
- exemple : 前台英语还能再烂一点吗? (L'anglais du personnage peut-il être pire ?)

Cependant, il existe encore des commentaires ironiques qui ne sont pas reconnus par les modèles. Par exemple, "也许酒店认为丢点钱并不是什么大事" (Peut-être que l'hôtel pense qu'il est normal de perdre de l'argent). Dans les travaux futurs, il est nécessaire d'intégrer des méthodes automatiques pour identifier l'ironie. Nous nous intéressons ici à la relation entre l'ironie et l'inférence.

Comme l'ironie est souvent utilisée pour critiquer ou exprimer le mécontentement, les textes ironiques sont majoritairement négatifs (Benamara et al. 2017). Cela correspond à la statistique d'inférence. Les commentaires négatifs avec inférences sont plus nombreux que les commentaires positifs.

Si un texte contient de l'ironie, il contient également une inférence, car le sens du texte n'est pas littéral. L'ironie est une condition suffisante pour l'inférence. Autrement dit, les méthodes de détermination de l'ironie peuvent être adaptées et appliquées pour détecter l'inférence. Nous avons essayé d'identifier les inférences dans les commentaires ironiques avec le modèle SVM. Tous les commentaires ironiques n'étaient pas considérés comme des commentaires avec inférences. Par conséquent, l'identification de l'ironie est un moyen pertinent de repérer les inférences.

7.3.2.4 凡尔赛文学 Littérature de Versailles

Depuis l'année 2020, un nouveau mot du web "凡尔赛文学" (traduction littérale : littérature de Versailles)a fait surface (Guo 2021). Il fait référence à des expressions qui semblent se plaindre, mais se vanter. Les gens utilisent des expressions euphémiques de plainte ou d'insatisfaction pour révéler par inadvertance leur condescendance. "好羡慕你们那些轻轻松松就长胖了的人，我这一个月吃了好多东西以为能 90 斤结果…太不公平了，我想哭！" (Je veux vraiment que vous preniez du poids facilement, j'ai mangé beaucoup de choses ce mois-ci et je pensais pouvoir accrocher 45 kilos... C'est tellement injuste, j'ai envie de pleurer !)

Puisque la vraie signification est implicite, ces types d'expressions contiennent des inférences. Les deux exemples ci-dessous expliquent comment le phénomène connu sous le nom de *Littérature de Versailles* influence la fouille d'opinion :

- 虽然酒店的地下停车位是收费的但是车位太小我的加长版的车子几乎不能掉头 (Bien que la place de parking de l'hôtel soit payante, la place de parking est trop petite, ma voiture de luxe inclinable peut difficilement faire demi-tour). À première vue, ce commentaire critique la petite superficie du parking. En effet, ce client aimerait montrer sa voiture extraordinaire.)
- 升级了最贵的顶楼套房，也无法看到凯旋门夜景，只能看到埃菲尔铁塔。(Après avoir choisi la suite penthouse la plus chère, je ne vois toujours pas la vue nocturne de l'Arc de Triomphe, seulement la tour Eiffel). L'intention d'écrire ce commentaire n'est pas de se plaindre du manque de vue, mais de se vanter d'avoir pris la suite où l'on peut voir la tour Eiffel.)

En comparaison avec les commentaires laissés par les autres clients de ces hôtels, la valence négative identifiée dans ces messages ne se retrouve pas dans les autres messages. Pour le premier exemple, avoir un parking est un plus pour la majorité. Pour le deuxième exemple, les autres commentaires du même hôtel apprécient plutôt la jolie vue du soir et la bonne localisation.

Après avoir appliqué notre modèle, ces commentaires sont catégorisés comme inférentiels et négatifs. Tout d'abord, le modèle a la capacité d'identifier l'inférence dans ces commentaires. Cependant, la polarité négative ne représente pas la véritable intention de ces commentaires. Par la suite, pour distinguer les opinions hétérogènes, nous allons soit désigner une nouvelle catégorie de polarité spécifique, soit catégoriser plus en détail les types d'inférences.

7.4 Conclusion

Dans ce chapitre, nous avons essayé de trouver une solution pour équilibrer l'annotation manuelle et automatique. Sur la base de la comparaison des résultats, l'étiquette la plus importante est la présence d'inférence. Parmi les 5 types d'inférence, les inférences logiques et pragmatiques jouent un rôle plus important dans la prédiction de la polarité, mais leur rôle est moins évident par rapport à la présence d'inférence. Par conséquent, la méthode la plus efficace consiste à annoter manuellement la présence d'inférence et à classer automatiquement les types d'inférence.

Ensuite, nous avons pris du recul sur l'état de l'art des inférences et des résultats obtenus. Cela nous permet de réfléchir à la pertinence de notre typologie des inférences et d'ajouter de nouveaux types dans les travaux futurs. De plus, nous avons analysé la relation entre les inférences et certains phénomènes linguistiques s'observent davantage dans notre corpus, comme les intensifieurs, les connecteurs et l'ironie.

Chapitre 8

Conclusion

Contents

8.1 Contribution 99

8.1.1 Corpus annoté avec les inférences, la polarité et les thèmes 99

8.1.2 Classification automatique des inférences 100

8.2 Perspectives 101

8.2.1 À court terme 101

8.2.2 À long terme 102

Dans ce travail de thèse, nous avons combiné les inférences linguistiques avec la fouille d'opinion afin de proposer une méthode d'identification des opinions implicites. Tous nos travaux prouvent que les inférences sont pertinentes pour améliorer la performance de la fouille d'opinion.

En étudiant des recherches existantes concernant différents types d'inférence dans le chapitre 2, nous avons constaté que les inférences sont peu traitées dans la fouille d'opinion parmi la littérature existante (chapitre 3). Nous nous sommes inspirés des travaux existants pour proposer une typologie des inférences pour le chinois, autour de 5 types : logique, pragmatique, lexicale, énonciative et discursive. Puis, sur la base de cette typologie, nous avons produit un guide d'annotation que nous avons ensuite utilisé pour annoter un corpus (chapitre 5), dans l'objectif de construire un corpus d'entraînement pour la classification automatique des inférences (chapitre 6) et de comparer les expériences de la fouille d'opinion avec et sans le traitement d'inférence (chapitre 7).

8.1 Contribution

8.1.1 *Corpus annoté avec les inférences, la polarité et les thèmes*

Notre première contribution est la réalisation d'un corpus de commentaires rédigés en chinois par des touristes en visite à Paris annotés avec les 5 types d'inférence (logique, pragmatique, lexicale, énonciative et discursive), la polarité (positive, négative, neutre et

inconnue) et les 9 thèmes. Ce corpus est composé de 1391 commentaires. Nous mettons à disposition ce corpus annoté sur notre compte GitHub[1].

Compte tenu des scores des accords inter-annotateurs, les inférences discursives et énonciatives sont plus difficiles à identifier au niveau du type et des frontières. Les annotatrices avaient moins de difficulté à typer l'inférence lexicale. Les scores des inférences logique et pragmatique sont tous autour de la moyenne. Cependant, le nombre d'inférences discursive et énonciative est faible par rapport aux trois autres types. Le manque de candidats ne permet pas au modèle statistique d'effectuer un bon apprentissage.

Afin d'identifier les frontières des inférences et de localiser les inférences dans un commentaire, nous avons proposé quatre niveaux d'annotation : commentaire, phrase, syntagme et lexical. Ces quatre niveaux nous permettent de mieux localiser les inférences et d'analyser la distribution des inférences dans chaque niveau. Au niveau du commentaire et de la phrase, presque tous les candidats annotés contiennent des inférences (100% et 92%), tandis que la présence d'inférence diminue au niveau du syntagme (46%). La distribution variée des inférences dans les quatre niveaux montre que le long commentaire est plus susceptible d'avoir des inférences. En même temps, il y a plus de types d'inférence dans ce genre de commentaire.

Pour conclure, parmi les cinq types d'inférence, l'inférence discursive est la plus difficile à traiter. Elle est généralement présente dans de longs commentaires, c'est-à-dire au niveau du commentaire. Pourtant, parmi les quatre niveaux textuels, le niveau du commentaire contient plus de complexité linguistique. Cela rend le problème plus difficile à résoudre.

8.1.2 *Classification automatique des inférences*

La deuxième contribution réside dans la classification automatique des inférences en fonction des caractéristiques linguistiques, des métadonnées du domaine et des vecteurs du plongement de mots.

Pour l'identification des inférences, la meilleure solution est de fusionner les catégories "absence" et "incertitude". Cette fusion d'un côté transforme la tâche de classification multiple en binaire, et de l'autre côté obtient le meilleur résultat. Pour classer les types d'inférence, un modèle spécifique à un type est meilleur qu'un modèle de multi-labels.

Lors de l'analysant des résultats, les inférences discursive et énonciative restent toujours complexes en raison du manque d'exemples. Après avoir équilibré l'échantillonnage, le résultat a été amélioré, mais le corpus de réduction a perdu beaucoup d'informations textuelles. Cette tâche nécessite une augmentation du nombre d'exemples et une maîtrise des informations retenues ou l'utilisation d'algorithmes plus performants.

Bien que certains types restent encore difficiles à traiter, l'intégration de la classification des inférences dans la fouille d'opinion a amélioré la détection de la polarité et

1. `https://github.com/liyunyan/ChineseHotelReviewAnnotation`

du thème. Les résultats d'expériences ont montré que le traitement des inférences est pertinent pour la fouille d'opinion, en particulier pour les opinions implicites.

De plus, compte tenu du coût élevé de l'annotation manuelle, l'objectif est de maximiser la classification automatique dans la fouille d'opinion. Nous avons trouvé une solution équilibrée entre la classification manuelle coûteuse et la classification automatique : annoter manuellement la présence d'inférence et utiliser directement les types d'inférence automatiquement classés .

8.2 Perspectives

Tout au long de nos recherches, nous avons mis l'accent sur certaines perspectives en distinguant les pistes à court et à long terme.

8.2.1 *À court terme*

8.2.1.1 Quatre niveaux textuels

Dans notre travail, nous n'avons analysé que la distribution des inférences dans les quatre niveaux textuels (commentaire, phrase, syntagme et lexical). Par la suite, nous prévoyons d'étudier les composants (qui sont l'ensemble des phrases ou syntagmes séparés par la ponctuations et qui construisent une inférence) d'une inférence dans les différents niveaux, en particulier l'enchaînement des composants. Pour l'inférence discursive qui est souvent construite à partir de plusieurs énoncés, ses composants couvrent certainement plusieurs syntagmes ou phrases. Premièrement, nous pouvons reproduire les inférences en prenant en compte du processus de raisonnement. Cela nous aide à déterminer les types d'inférences. Deuxièmement, il est plus clair de définir les frontières des inférences en définissant chaque composant d'une inférence. Enfin, puisque la production des inférences n'est pas toujours dans le dernier composant, l'ordre du processus de raisonnement devient important. En conséquence, nous pouvons introduire de nouveaux types d'inférence comme la déduction, l'induction et la rétroduction qui prennent en compte l'ordre du processus.

8.2.1.2 Domaine spécifique vs. domaine général

Notre analyse des inférences porte pour l'instant sur un corpus de commentaires touristiques à Paris. Il serait intéressant d'adapter des inférences à d'autres types de corpus ou à un corpus général. Nous proposons deux pistes à suivre :

1. reproduire les vecteurs avec différents plongements lexicaux. Les vecteurs actuels des expériences sont basés sur l'ensemble de notre corpus de deux sites (Booking.com et Mafengwo.cn). Nous testerons successivement le plongement de mots sur la base d'un corpus spécifique de commentaires touristiques plus large, d'un corpus de commentaires de produits ou de services et d'un grand corpus général

en chinois tel que BERT (Y. Wang et al. 2020 ; M. Tan et J. Jiang 2020 ; M. Li et al. 2021).

2. appliquer le modèle actuel au corpus du site Mafengwo.cn. Nous avons déjà annoté manuellement un petit corpus de 100 commentaires provenant de Mafengwo.cn. Après l'application du modèle construit à partir du corpus de Booking.com, la performance du modèle ne s'est pas beaucoup dégradée. Nous continuerons à tester notre méthode dans un corpus de Mafengwo.cn plus grand puis dans d'autres types de corpus.

8.2.1.3 Thématique

Nous envisageons également d'analyser plus en détail la détection des thèmes abordés dans les corpus. D'abord, comme la catégorisation des thèmes est une tâche subjective(Mukherjee et B. Liu 2012), nous démontrerons si la catégorisation de 10 thèmes est pertinente. Ensuite, en réalisant les expériences d'inférence, nous pouvons analyser les relations entre les types d'inférence et les thèmes.

8.2.2 *À long terme*

Le traitement des opinions implicites constitue une problématique importante pour la fouille d'opinion qui nécessite des travaux complémentaires.

8.2.2.1 Phénomènes linguistiques

Les inférences sont fortement liées aux phénomènes linguistiques. Dans notre thèse, nous avons uniquement analysé la relation entre inférences et intensifieurs, connecteurs et ironie au chapitre 8, car nous avons constaté que ces trois phénomènes sont plus fréquents dans notre corpus. Cependant, il existe nécessairement d'autres phénomènes linguistiques dans d'autres types de corpus.

Au niveau linguistique, nous allons continuer à caractériser et modéliser les autres phénomènes qui font partie des expressions implicites, par exemple, la construction des verbes en série (Müller et Lipenkova 2009 ; Waltraud 2008). Étant une caractéristique de certaines langues asiatiques, les verbes en série[2] ne sont pas beaucoup observés dans notre corpus. Néanmoins, les occurrences trouvées contiennent souvent des inférences. Par exemple, "免费停车" (traduction littérale : être gratuit + stationner). Il est possible que dans d'autres types de corpus, les verbes en série deviennent une caractéristique plus évidente pour repérer les inférences.

2. La construction des verbes en série est un phénomène syntaxique dans lequel deux ou plusieurs verbes ou phrases verbales sont enchaînés en une seule clause.

8.2.2.2 Type de corpus varié

Au niveau du corpus, notre corpus de commentaire touristique est limité à un seul domaine. Nous envisageons également d'étendre notre travail aux inférences présentes dans différents types de corpus, par exemple des mini blogs provenant de réseaux sociaux ou des échanges sur le même sujet dans lesquels les allers-retours contiennent certainement différents types d'inférence.

Nous continuons à étudier comment adapter les inférences dans d'autres domaines. Comme un phénomène linguistique, les inférences ne se limitent pas à un seul domaine. Nous souhaitons adapter les inférences dans d'autres domaines (Garcia-Fernandez et al. 2014 ; W. Wang et al. 2017).

8.2.2.3 Algorithmes

Dans les expériences automatiques, le but est de vérifier si les inférences sont pertinentes pour la fouille d'opinion. Au lieu de comparer la performance des algorithmes, nous avons utilisé des SVM dans toutes nos expériences. Nous envisageons d'étudier l'impact du paramétrage des modèles en fonction de la taille des échantillons traités et des caractéristiques disponibles, comme souligné par Basari et al. (2013), Khairnar et Kinikar (2013) et Sabuj et al. (2017).

Nous estimons également que des algorithmes plus récents que les SVM seraient appropriés pour combiner les inférences dans la fouille d'opinion. Nous envisageons d'emprunter les méthodes traitant les polarités et les aspects implicites (Zeng et F. Li 2013 ; H.-Y. Chen et H.-H. Chen 2016a) : distinguer les caractéristiques des opinions explicites et implicites (B. Liu et al. 2005 ; Hai et al. 2011), catégoriser les indicateurs des aspects implicites (Cruz et al. 2014).

Bibliographie

Basari, Abd. Samad Hasan, Burairah Hussin, I. Gede Pramudya Ananta et Junta Zeniarja (2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. In : *Procedia Engineering* 53. Malaysian Technical Universities Conference on Engineering ; Technology 2012, MUCET 2012, p. 453-462. doi : https://doi.org/10.1016/j.proeng.2013.02.059.

Beaupré, Sophie (2009). L'approche dialectique pragmatique dans l'analyse des arguments. Mém. de mast. Montréal, Canada : UQAM.

Benamara, Farah, Cyril Grouin, Jihen Karoui, Véronique Moriceau et Isabelle Robba (2017). Analyse d'opinion et langage figuratif dans des tweets : présentation et résultats du Défi Fouille de Textes DEFT2017. In : *Atelier TALN 2017 : Défi Fouille de Textes (DEFT 2017)*. Orléans, France, pp. 1-12.

Bhaskar, J., K. Sruthi et Prema Nedungadi (2014). Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers. In : *International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)*, p. 1-6.

Bouquiaux, Laurence et Bruno Leclercq (2017). *Logique formelle et argumentation*. 2e. De Beock.

Bres, Jacques (2013). Catherine Kerbrat-Orecchioni, L'implicite. In : *Cahiers de praxématique*. doi : 10.4000/praxematique.3497.

Calvo, Manuel G., Enrique Meseguer et Manuel Carreiras (2001). Inferences about predictable events : eye movements during reading. In : *Psychological Research* 65.3, p. 158-169. doi : 10.1007/s004260000050.

Cambria, E., B. Schuller, Y. Xia et C. Havasi (2013). New Avenues in Opinion Mining and Sentiment Analysis. In : *IEEE Intelligent Systems* 28.2, p. 15-21. doi : 10.1109/MIS.2013.30.

Cambridge Advanced Learner's Thesaurus (2013). 4e éd. Cambridge University Press. isbn : 978-1107035157.

Chan, Samuel W. K. et Benjamin K. TŚou (1999). Semantic Inference for Anaphora Resolution : Toward a Framework in Machine Translation. In : *Machine Translation* 14.3/4, p. 163-190.

Chen, Huan-Yuan et Hsin-Hsi Chen (2016a). Implicit Polarity and Implicit Aspect Recognition in Opinion Mining. In : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, p. 20-25. doi : 10.18653/v1/P16-2004.

– (jan. 2016b). Implicit Polarity and Implicit Aspect Recognition in Opinion Mining. In : p. 20-25. doi : 10.18653/v1/P16-2004.

Cruz, I., A. Gelbukh et G. Sidorov (2014). Implicit Aspect Indicator Extraction for Aspect-based Opinion Mining. In : *International journal of computational linguistics and applications* 5.2, p. 135-152.

Cuel, Louis (2014). Discrete geometric inference. Theses. Université de Grenoble.

Dave, Kushal, Steve Lawrence et David Pennock (2003). Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. In : *Proceedings of WWW-03, 12th International Conference on the World Wide Web*. ACM Press, p. 519-528.

Davis, Wayne (2019). Implicature. In : *The Stanford Encyclopedia of Philosophy*. Sous la dir. d'Edward N. Zalta. Fall 2019. Metaphysics Research Lab, Stanford University.

Deledalle, Gérard et Charles S. Peirce (1994). Les ruptures épistémiologiques et les nouveaux paradigmes. In : *Travaux du Centre de Recherches Sémiologiques* 62.

Denhière, Guy et Serge Baudet (1992). *compréhension de texte et sciences cognitives*. Paris, P.U.F.

Dictionnaire de L'Académie française (1798). 5e éd.

Dong, Jian, Hongxiu Li et Xianfeng Zhang (2014). Classification of Customer Satisfaction Attributes : An Application of Online Hotel Review Analysis. In : *13th Conference on e-Business, e-Services and e-Society (I3E)*. Sous la dir. d'Hongxiu Li, Matti Mäntymäki et Xianfeng Zhang. T. AICT-445. Digital Services and Information Intelligence. Part 3 : Digital Business. Sanya, China : Springer, p. 238-250. doi : 10.1007/978-3-662-45526-5_23.

Doucy, Geoffrey et Viavoo Massoussi (juil. 2012). Sémantique inférentielle et compréhension des verbatim clients. In : *SHS Web of Conferences* 1. doi : 10.1051/shsconf/20120100230.

Doussau, Constance et Sabine Rigal (2011). Étude du développement de la production d'inférences de liaison en compréhension écrite du CE1 au CM1. Mém. de mast. Université Claude Bernard Lyon1 : Université Claude Bernard Lyon1.

Drew, Paul (2018). Inferences and Indirectness in Interaction. In : *Open Linguistics* 4, p. 241-259. doi : 10.1515/opli-2018-0013.

Du, Weifu et Songbo Tan (2009). An Iterative Reinforcement Approach for Fine-Grained Opinion Mining. In : *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado : Association for Computational Linguistics, p. 486-493.

Duchêne, Annick (2008). Les inférences dans la communication : cadre théorique général. In : *Actes de Réducation orthophonique*. Fédération Nationale des Orthophonistes.

Ducrot, Oswald (1969). Présupposés et sous-entendus. In : *Langue française, n°4, 1969. La sémantique*. doi : 10.3406/lfr.1969.5456.

Ducrot, Oswald et Jean-Marie Schaeffer (1999). *Nouveau Dictionnaire Encyclopédique des Sciences du Langage*.

Dufaye, Lionel (2001). Les Modaux et la négation en anglais contemporain. In : *Cahiers de Recherche*. Ophrys.

Durand-Guerrier, Viviane (1994-1995). Logique et raisonnement mathématique Exemple d'analyse de tâches à l'aide de la logique formelle. fr. In : *Publications mathématiques et informatique de Rennes* 3, p. 1-13.

Eensoo, Egle et Mathieu Valette (2012). Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments. In : *TALN 2012*. Sous la dir. d'Antoniadis, Georges, Blanchon, Hervé et Sérasset et Gilles. T. 2. Grenoble, France : GETALP-LIG, p. 367-374.

Ekman, Paul (1999). Basic emotions. In : *Handbook of cognition and emotion* 98.45-60, p. 16.

Fayol, Michel (2003). La compréhension : évaluation, difficultés et interventions. In : *Actes de Conférence de Consensus*. Paris.

Garcia-Fernandez, Anne, Olivier Ferret et Marco Dinarelli (2014). Evaluation of different strategies for domain adaptation in opinion mining. In : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland : European Language Resources Association (ELRA), p. 3877-3880.

Gjelsvik, Olav (2015). Rationality, Capacity and Inference. In : *Teorema : Revista Internacional de Filosofía* 34.2, p. 105-116.

Gombert, Jean-Emile, Michel Fayol, Daniel Zagar, Pierre Lecocq et Liliane Sprenger-Charolles (1992). *Psychologie cognitive de la lecture*. Sous la dir. de PUF.

Graesser, Arthur, Murray Singer et Tom Trabasso (août 1994). Constructing Inferences During Narrative Text Comprehension. In : *Psychological review* 101, p. 371-95. doi : 10.1037/0033-295X.101.3.371.

Grice, H. Paul (1975). Logic and Conversation. In : *Syntax and Semantics* 3, p. 41-58.

Guermeur, Yann (2007). SVM Multiclasses, Théorie et Applications. Habilitation à diriger des recherches. Université Henri Poincaré - Nancy I.

Guo, Jie (2021). Viewing "Versailles Literature" from the Perspective of Interactive Linguistics. In : *Modern Linguistics* 09, p. 120-124. doi : 10.12677/ML.2021.91019.

Hai, Zhen, Kuiyu Chang et Jung-Jae Kim (2011). Implicit Feature Identification via Co-occurrence Association Rule Mining. In : *Computational Linguistics and Intelligent Text Processing - 12th International Conference*. Tokyo, Japan, p. 393-404. doi : 10.1007/978-3-642-19400-9_31.

Halpern, Jack (2006). The Role of Lexical Resources in CJK Natural Language Processing. In : *Proc of Multilingual Language Resources and Interoperability Work*. Sydney, Australia, p. 9-16.

Han, Zhong-Ming, Meng-Qi Li, Wen Liu, Meng-Mei Zhang, Da-Gao Duan et Chong-Chong Yu (2018). Survey of Studies on Aspect-Based Opinion Mining of Internet. In : *Journal of Software* 29.2, p. 417-441.

Horn, Laurence (1984). Toward a new taxonomy for pragmatic inference : Q-based and R-based implicature. In : *Meaning, form, and use in context : linguistic applications*. Sous la dir. de Deborah Schiffrin. Georgetown University Press, p. 10-42.

Hu, Minqing et Bing Liu (2004). Mining and Summarizing Customer Reviews. In : *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Dis-*

covery and Data Mining. KDD '04. Seattle, WA, USA : ACM, p. 168-177. doi : 10.1145/1014052.1014073.

Huang, Yu-Hsiang, Hen-Hsen Huang et H. Chen (2017). Irony Detection with Attentive Recurrent Neural Networks. In : *ECIR*.

Itkes, Oksana et Nira Mashal (2015). Processing negative valence of word pairs that include a positive word. In : *Cognition and Emotion*. doi : 10.1080/02699931.2015.1039934.

Jiang, Long, Ming Zhou, Lee-Feng Chien et Cheng Niu (2007). Named Entity Translation with Web Mining and Transliteration. In : *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI'07. Hyderabad, India : Morgan Kaufmann Publishers Inc., p. 1629-1634.

Joshi, Anju et Anubhooti Papola (2017). Aspect Level Opinion Mining on Customer Reviews using Support Vector Machine. In : *International Journal of Advanced Research in Computer and Communication Engineering*. T. 6.

Kern-Isberner, Gabriele et Christian Eichhorn (2014). Structural Inference from Conditional Knowledge Bases. In : *Studia Logica : An International Journal for Symbolic Logic* 102.4, p. 751-769.

Khairnar, Jayashri et Mayura Kinikar (2013). Machine Learning Algorithms for Opinion Mining and Sentiment Classification.

Khemlani, Sangeet, J Trafton, Max Lotstein et P Johnson-Laird (jan. 2012). A process model of immediate inferences. In : p. 151-156.

Kim, S., J. Zhang, Z. Chen, A. Oh et S. Liu (2013). A hierarchical aspect-sentiment model for online reviews. In : p. 526-533.

Kim, Soo-Min et Eduard Hovy (2004). Determining the Sentiment of Opinions. In : *Proceedings of the 20th International Conference on Computational Linguistics*. COLING '04. Geneva, Switzerland : Association for Computational Linguistics. doi : 10.3115/1220355.1220555.

– (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In : *Proceedings of the Workshop on Sentiment and Subjectivity in Text*. SST '06. Sydney, Australia : Association for Computational Linguistics, p. 1-8.

Kintsch, Walter (1998). *Comprehension : A paradigm for cognition*. Sous la dir. de Cambridge University Press.

Kispal, Anne (jan. 2008). Effective Teaching of Inference Skills for Reading. Literature Review. Research Report DCSF-RR031. In : *National Foundation for Educational Research*.

Larousse, éd. (1994). *Dictionnaire De Linguistique Et Des Sciences Du Langage*.

– éd. (2015). *Dictionnaire Larousse maxipoche plus*.

Lavigne, Judith (2008). Les mécanismes d'inférence en lecture chez les élèves de sixième année du primaire. Thèse de doct. Québec, Canada : Université Laval.

Lazhar, Farek (jan. 2019). Implicit Feature Identification for Opinion Mining. In : *Int. J. Bus. Inf. Syst.* 30.1, p. 13-30.

Levinson, Stephen C. (2016). Turn-taking in Human Communication –Origins and Implications for Language Processing. In : *Trends in Cognitive Sciences* 20.1, p. 6-14. doi : https://doi.org/10.1016/j.tics.2015.10.010.

Li, Anran et Chu-ren Huang (2019). A Method of Modern Chinese Irony Detection. In : *From Minimal Contrast to Meaning Construct : Corpus-based, Near Synonym Driven Approaches to Chinese Lexical Semantics*.

Li, Mingzheng, Lei Chen, Jing Zhao et Qiang Li (2021). A Chinese Stock Reviews Sentiment Analysis Based on BERT Model. In : *Applied Intelligence*. doi : 10.21203/rs.3.rs-69958/v1.

Li, ShengYu, JunBo Gao et LiLi Xu (2017). Sentiment Analysis Solution Based on Hotel Product Reviews. In : *Computer Systems & Applications* 26, p. 227-231. doi : 10.15888/j.cnki.csa.005511.

List, Johann-Mattis (2019). Automatic Inference of Sound Correspondence Patterns across Multiple Languages. In : *Computational Linguistics* 45.1, p. 137-161. doi : 10.1162/coli_a_00344.

Liu, Bing (2010). Sentiment Analysis and Subjectivity. In : *Handbook of Natural Language Processing*.

– (2012). Sentiment Analysis and Opinion Mining. In : *Synthesis Lectures on Human Language Technologies*.

Liu, Bing, Minqing Hu et Junsheng Cheng (2005). Opinion Observer : Analyzing and Comparing Opinions on the Web. In : *Proceedings of the 14th International Conference on World Wide Web*. Chiba, Japan : Association for Computing Machinery, p. 342-351. doi : 10.1145/1060745.1060797.

Liu, Huan et Lei Yu (2005). Toward integrating feature selection algorithms for classification and clustering. In : *IEEE Transactions on Knowledge and Data Engineering* 17.4, p. 491-502. doi : 10.1109/TKDE.2005.66.

Lou, De-Cheng et Tian-Fang Yao (2006). Semantic analysis and opinion mining on Chinese review sentences. In : *Journal of Computer Applications* 26.11, p. 2622.

Lu, Yanan, Yue Zhang et Donghong Ji (2016). Multi-prototype Chinese Character Embedding. In : *Proc of LREC*. Sous la dir. de Nicoletta Calzolari (Conference Chair) et al. Portorož, Slovenia : European Language Resources Association (ELRA).

Marcel Adam, Just et H. Clark Herbert (1973). Drawing inferences from the presuppositions and implications of affirmative and negative sentences. In : *Journal of Verbal Learning and Verbal Behavior* 12.1, p. 21-31. doi : 10.1016/S0022-5371(73)80057-X.

Markus, Keith A. (2014). An incremental approach to causal inference in the behavioral sciences. In : *Synthese* 191.10, p. 2089-2113.

Martin, Robert (1976). *Inférence, antonymie et paraphrase - Eléments pour une théorie sémantique*. Klincksieck. isbn : 2-252-01906-9.

– (2004). *Comprendre la linguistique : épistémologie élémentaire d'une discipline*. 2e. Presses universitaires de France. isbn : 2-13-054549-1.

Martins, Daniel et Brigitte Le Bouëdec (1998). La production d'inférences lors de la compréhension de textes chez des adultes : une analyse de la littérature. In : *L'Année psychologique* 98.3, p. 511-543. doi : 10.3406/psy.1998.28581.

McGarrity, K. S., J. Sietsma et G. Jongbloed (2014). Nonparametric Inference in a Stereological Model with Oriented Cylinders Applied to Dual Phase Steel. In : *The Annals of Applied Statistics* 8.4, p. 2538-2566.

McHugh, Mary L. (2012). Interrater reliability : The kappa statistic. In : *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB* 22, p. 276-82. doi : 10.11613/BM.2012.031.

McMullin, Ernan (2013). The Inference that Makes Science. In : *Zygon*® 48.1, p. 143-191. doi : 10.1111/j.1467-9744.2012.01319.x.

Mejri, Salah (2011). Phraséologie et traduction. In : *Équivalences, 38e année-n°1-2, 2011. L'enseignement de la traduction, sous la direction de Christian Balliu.* doi : 10.3406/equiv.2011.1363.

Meredith, M. Pike, A. Barnes Marcia et W. Barron Roderick (2010). The role of illustrations in children' s inferential comprehension. In : *Journal of Experimental Child Psychology* 105.3, p. 243-255. doi : https://doi.org/10.1016/j.jecp.2009.10.006.

Mukherjee, Arjun et Bing Liu (juil. 2012). Aspect Extraction through Semi-Supervised Modeling. In : *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Jeju Island, Korea : Association for Computational Linguistics, p. 339-348. url : https://www.aclweb.org/anthology/P12-1036.

Müller, Stefan et Janna Lipenkova (2009). Serial Verb Constructions in Chinese : An HPSG Account. In : *Proceedings of the 16th International Conference on Head-Driven Phrase Structure Grammar*. University of Göttingen, Germany : CSLI Publications, p. 234-254.

Oxford English Dictionary (2000). Oxford University Press. OED online.

Pan, Hongliang, Zheng Lin, Peng Fu, Yatao Qi et Weiping Wang (2020). Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection. In : *Findings of the Association for Computational Linguistics : EMNLP 2020*. Online : Association for Computational Linguistics, p. 1383-1392. doi : 10.18653/v1/2020.findings-emnlp.124.

Pang, Bo et Lillian Lee (2008). Opinion Mining and Sentiment Analysis. In : *Found. Trends Inf. Retr.* 2.1-2, p. 1-135. doi : 10.1561/1500000011.

Pang, Bo, Lillian Lee et Shivakumar Vaithyanathan (2002). Thumbs up ? Sentiment Classification using Machine Learning Techniques. In : *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. University of Pennsylvania, Philadelphia, PA, USA : Association for Computational Linguistics, p. 79-86. doi : 10.3115/1118693.1118704.

Patron, Sylvie (2011). Enunciative Narratology : a French Speciality. In : *Current Trends in Narratology*. Sous la dir. de Greta Olson. Narratologia. Berlin, Walter de Gruyter, pp. 267-289.

Pedregosa, Fabian et al. (2011). Scikit-learn : Machine Learning in Python. In : *Journal of Machine Learning Research* 12.85, p. 2825-2830.

Peirce, Charles S. (1958). The Collected Papers of Charles Sanders Peirce. In : *Cambridge : Harvard University Press*. T. 1-6.

Prabowo, Rudy et Mike Thelwall (2009). Sentiment analysis : A combined approach. In : *J. Informetrics* 3, p. 143-157.

Ranger, Graham (2013). MIND YOU : an enunciative description. In : *Colloque "Modalité et commentaire / Modalisation a posteriori"*. Paris, France.

Recanati, François (2002). *Literal Meaning*. Cambridge University Press.

Rinaldi, Ekki et Aina Musdholifah (2017). FVEC-SVM for opinion mining on Indonesian comments of youtube video. In : doi : 10.1109/ICODSE.2017.8285860.

Rodriguez, Abel et Peter Müller (2013). Nonparametric Bayesian Inference. In : *NSF-CBMS Regional Conference Series in Probability and Statistics* 9, p. i-110.

Rossi, Jean-Pierre et N Campion (1999). Inférences et compréhension de texte. fre. In : *L'Année psychologique* 99.3, p. 493-527. doi : 10.3406/psy.1999.28518.

Rubin, Victoria L. (2014). TALIP Perspectives, Guest Editorial Commentary : Pragmatic and Cultural Considerations for Deception Detection in Asian Languages. In : *ACM Transactions on Asian Language Information Processing* 13.2. doi : 10.1145/2605292.

Ruph Porte, Catherine (2011). Inférence lexicale et sens figuré : une entrée didactique. Mém. de mast. Université Stendhal Grenoble 3, p. 126.

Sabuj, Mir Shahriar, Zakia Afrin et K. M. Hasan (2017). Opinion Mining Using Support Vector Machine with Web Based Diverse Data. In : *International Conference on Pattern Recognition and Machine Intelligence*, p. 673-678. doi : 10.1007/978-3-319-69900-4_85.

Saleh, M. Rushdi, M.T. Martín-Valdivia, A. Montejo-Ráez et L.A. Ureña-López (2011). Experiments with SVM to classify opinions in different domains. In : *Expert Systems with Applications* 38.12, p. 14799-14804. doi : https://doi.org/10.1016/j.eswa.2011.05.070.

Samha, Amani K., Yuefeng Li et Jinglan Zhang (2014). Aspect-Based Opinion Extraction from Customer reviews. In : *CoRR* abs/1404.1982.

Schmalhofer, Franz, Mark Mcdaniel et Dennis Keefe (2002). A Unified Model for Predictive and Bridging Inferences. In : *DISCOURSE PROCESSES* 33, p. 105-132. doi : 10.1207/S15326950DP3302_01.

Sheng, Steve (2011). *Report on Chinese Variants in Internationalized Top-Level Domains*. Rapp. tech. Marina Del Rey, CA : ICANN.

Silins, Nicholas (2013). Introspection and inference. In : *Philosophical Studies : An International Journal for Philosophy in the Analytic Tradition* 163.2, p. 291-315.

Simonin, Olivier (sept. 2018). Sens implicite, implicatures et principes d'inférence. In : *CORELA - COgnition, REprésentation, LAngage* HS-25.

Spotorno, Nicola, Corey Mcmillan, Katya Rascovsky, David Irwin, Robin Clark et Murray Grossman (2015). Beyond words : Pragmatic inference in behavioral variant of frontotemporal degeneration. In : *Neuropsychologia* 75. doi : 10.1016/j.neuropsychologia.2015.07.002.

Su, Qi, Xinying Xu, Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen et Zhong Su (2008). Hidden Sentiment Association in Chinese Web Opinion Mining. In : *Proceedings of the 17th International Conference on World Wide Web*. Beijing, China : Association for Computing Machinery, p. 959-968. doi : 10.1145/1367497.1367627.

Sun, Weiwei et Xiaojun Wan (2016). Towards Accurate and Efficient Chinese Part-of-Speech Tagging. In : *Computational Linguistics* 42.3, p. 391-419. doi : 10.1162/COLI_a_00253.

Tan, Minghuan et Jing Jiang (2020). A BERT-based Dual Embedding Model for Chinese Idiom Prediction. arXiv : 2011.02378 [cs.CL].

Tanaka, Koji (2011). Inference in the Mengzi 1A : 7. In : *Journal of Chinese Philosophy* 38.3, p. 444-454. doi : 10.1111/j.1540-6253.2011.01665.x.

Tang, Yi-jie et Hsin-Hsi Chen (2014). Chinese Irony Corpus Construction and Ironic Structure Analysis. In : *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*. Dublin, Ireland : Dublin City University et Association for Computational Linguistics, p. 1269-1278.

Tang, Mingjun (2015). Materials for the Study of Xuanzang's Inference of Consciousness-only (wei shi bi liang). In : *Wiener Zeitschrift für die Kunde Südasiens / Vienna Journal of South Asian Studies* 56/57, p. 143-198.

Tao, Aili (2014). Magnitude Classification of the Scope of Degree-Adverbs/ 刍议程度范畴的量级划分. In : *J. CENT. SOUTH UNIV. (SOCIAL SCIENCE)* 20.

Thibaud, Elodie et Justine Viviant (2014). Compréhension de mots nouveaux et mécanismes d'inférences chez des enfants atteints de dysphasie âgés de 8 à 11 ans. Mém. de mast. Université Claude Bernard Lyon 1.

Tsou, B.K., H.L. Lin, H.C. Ho et B.Y. Lai (1996). From Argumentative Discourse to Inference Trees : Using Syntactic Markers as Cues in Chinese Text Abstraction. In : *Journal of Chinese Linguistics Monograph Series* 9, p. 199-222.

Turney, Peter D. et Michael L. Littman (2003). Measuring Praise and Criticism : Inference of Semantic Orientation from Association. In : *ACM Trans. Inf. Syst.* 21.4, p. 315-346. doi : 10.1145/944012.944013.

Van den Broek, Paul, Michael Young, Y Tzeng et Tracy Linderholm (1999). The landscape model of reading : Inferences and the one-line construction of memory representation. In : *The construction of mental representations during reading*, p. 71-98.

Varghese, R. et M. Jayasree (2013). Aspect based Sentiment Analysis using support vector machine classifier. In : *2013 International Conference on Advances in Computing,*

Communications and Informatics (ICACCI), p. 1581-1586. doi : 10.1109/ICACCI.2013.6637416.

Vittori, Girolamo (1609). *Thrésor des trois langues : francoise, italiene, et espagnolle*. Genève, Philippe Albert et Alexandre Pernet.

Vlad, Monica (2011). Médiation du sens et interférence dans la lecture scolaire en français langue étrangère. In : *Synergies Pologne* 8, p. 107-115.

Wagener-Wender, Monika et K. Wender (1990). Expectations, Mental Representations, and Spatial Inferences. In : t. 25. Academic Press, p. 137-157. doi : https://doi.org/10.1016/S0079-7421(08)60253-4.

Waltraud, Paul (2008). The serial verb construction in Chinese : A tenacious myth and a Gordian knot. In : *Linguistic Review* 25.3-4, p. 367-411. doi : 10.1515/TLIR.2008.011.

Wang, Wei, Guanyin Tan et Hongwei Wang (2017). Cross-domain comparison of algorithm performance in extracting aspect-based opinions from Chinese online reviews. In : *International Journal of Machine Learning and Cybernetics* 8. doi : 10.1007/s13042-016-0596-x.

Wang, Yile, Leyang Cui et Yue Zhang (2020). Does Chinese BERT Encode Word Structure ? In : *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online) : International Committee on Computational Linguistics, p. 2826-2836. doi : 10.18653/v1/2020.coling-main.254.

Wichmann Søren ; Brown, Cecil H (2003). Contact among some Mayan languages : Inferences from loanwords. English. In : *Anthropological linguistics*.

Wilson, Deirdre et Dan Sperber (2012). *Meaning and Relevance*. Cambridge University Press.

Wilson, Theresa, Janyce Wiebe et Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In : *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada : Association for Computational Linguistics, p. 347-354. doi : 10.3115/1220575.1220619.

Wright, Crispin (2014). Comment on Paul Boghossian, "What is inference". In : *Philosophical Studies : An International Journal for Philosophy in the Analytic Tradition* 169.1, p. 27-37.

Wu, Shih-Jung et Rui-Dong Chiang (2015). Using syntactic rules to combine opinion elements in Chinese opinion mining systems. In : *JCIT*. T. 10, p. 137-144.

Wu, Yueh-Cheng et Shu-Kai Hsieh (2010). PyCWN : a Python Module for Chinese Wordnet. In : *Proc of COLING, Demo*. Beijing, China, p. 5-8.

Xiang, Rong, Xuefeng Gao, Yunfei Long, Anran Li, Emmanuele Chersoni, Qin Lu et Chu-Ren Huang (2020). Ciron : a New Benchmark Dataset for Chinese Irony Detection. In : *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France : European Language Resources Association, p. 5714-5720.

Xu, L., Hongfei Lin, Y. Pan, H. Ren et J. Chen (2008). Constructing the affective lexicon ontology. In : *Journal of the China Society for Scientific and Technical Information* 27, p. 180-185.

Yan, Liyun (2018). Analyse des inférences pour la fouille d'opinion en chinois. In : *CORIA-TALN-RJC 2018 - Conférence sur le Traitement Automatique des Langues Naturelles*.

Yang, Bishan et Claire Cardie (2013). Joint Inference for Fine-grained Opinion Extraction. In : *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Sofia, Bulgaria : Association for Computational Linguistics, p. 1640-1649.

Yao, Tian-fang, Xi-wen Cheng, Fei-yu Xu, Uszkoreit Hans et Rui Wang (2008). A Survey of Opinion Mining for Texts. In : *Journal of Chinese Information Processing* 22.3, p. 71.

Yi, J. et W. Niblack (2005). Sentiment mining in WebFountain. In : *21st International Conference on Data Engineering (ICDE'05)*, p. 1073-1083. doi : 10.1109/ICDE.2005.132.

Zeng, Lingwei et Fang Li (2013). A Classification-Based Approach for Implicit Feature Identification. In : *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. T. 8202, p. 190-202. doi : 10.1007/978-3-642-41491-6_18.

Liste des tableaux

3.1 Chiffres comparatifs de deux sites 27
3.2 Statistique du dictionnaire intégré dans la segmentation Jieba 30
3.3 Catégories et les interprétations de la partie du discours pour l'annotation manuelle 33
3.4 Nombre d'étiquettes et temps d'annotation 34
3.5 Données statistiques des hôtels dans chaque catégorie 35

4.1 Définition des thèmes et des sous-thème 39
4.2 Comparaison de la distribution des arrondissements 43
4.3 Scores des entités après l'accord 46
4.4 Scores des attributs après l'accord 47
4.5 F-mesure entre deux versions et version consensuelle 47
4.6 Distribution des étiquettes de la présence d'inférence selon les niveaux syntagme, phrase et commentaire 49
4.7 Distribution des étiquettes des types d'inférence selon les niveaux syntagme, phrase et commentaire 49
4.8 Distribution des polarités selon les niveaux syntagme, phrase et commentaire 51
4.9 Distribution des thèmes selon les niveaux syntagme, phrase et commentaire 52

5.1 Résultats de l'identification des inférences selon les différentes caractéristiques 66
5.2 Comparaison des résultats de la classification des 5 types avec de différentes caractéristiques 69
5.3 Résultats de la classification multi-labels des types d'inférence 69
5.4 Résultats du modèle de multi-labels avec les textes en pinyin 70
5.5 Comparaison des résultats de l'identification des inférences avant et après le fusionnement des catégories "Absence" et "Incertitude" 71
5.6 Comparaison des résultats de la classification multi-labels après et avant de fusionner l'inférence énonciative et discursive 71

6.1 Exemple de l'ontologie 75
6.2 Taux de couverture des mots de sentiments dans le corpus 76
6.3 Résultats de la prédiction de la polarité seulement avec les textes en vecteur 78
6.4 Résultats de la prédiction de la polarité avec les métadonnées 79

6.5 Résultats de la prédiction de la polarité avec les vecteurs, les métadonnées et les indices de la présence des inférences 80
6.6 Nombre de chaque type d'inférence d'une apparition tout seul et total . 81
6.7 Résultats de la prédiction de la polarité avec les vecteurs, les métadonnées, les indices de la présence des inférences et les cinq types d'inférence 81
6.8 Résultats comparatifs de la détection des thèmes selon des caractéristiques variante (V : vecteurs, M : métadonnées, Inf : présence d'inférence, Type : type d'inférence) 83
6.9 Comparaison des résultats de la polarité avec l'annotation manuelle et automatique (Ann. auto : annotation automatique, Ann. manuelle : annotation manuelle) 85

7.1 Meilleures combinaisons entre l'annotation manuelle et automatique . . 88
7.2 Exemple de polarité correctement détectée avec la classification des inférences 92

Table des figures

2.1 Extrait du Dictionnaire de l'Académie française 9

4.1 Comparaison de la distribution des arrondissements 44
4.2 Schéma d'annotation 45
4.3 Distribution des 5 types selon les niveaux syntagme, phrase et commentaire 50
4.4 Liste de spécificités des commentaires positifs, négatifs, neutres et inconnus 53

6.1 Proportion d'inférences présentes dans les opinions avec et sans mots de sentiment 77
6.2 Procédures de l'expérimentation de la prédiction de la polarité 82
6.3 Taux d'apprentissage comparatifs de la prédiction de la polarité en utilisant les différents paramètres 84

7.1 Nombre de commentaires positifs et négatifs en fonction du nombre de mots 93

Annexe A

Guide d’annotation (version française)

Guide d’annotation des inférences

Guide d’annotation des inférences 1
I. Introduction 2
II. Description du corpus 2
III. Niveaux d’analyse 2
II. Objectifs 3
IV. Annotation 3
1. Qu’est-ce qu’une inférence? 3
2. Annotation de type d’inférence 4
3. Comment faire une annotation avec l’outil Brat? 8
Installation Brat 8
Une phrase ou un syntagme? 9
Présence d’inférence, quels types et quel thème? 9
Exemple d’annotation 9
V. Un exercice avant de commencer 12

I. Introduction

Dans la fouille d'opinion, la machine interprète mieux des opinions explicites qui soit contiennent des mots des opinions, soit expriment d'une manière explicite. Mais plus souvent, une opinion est 1) une description des expériences dont la totalité exprime l'attitude d'un utilisateur ou 2) une opinion qui contient des inférences. C'est-à-dire qu'elle demande une interprétation avant d'aboutir à une conclusion. Afin de détecter des opinions avec des inférences à l'aide de l'apprentissage automatique, nous avons besoin d'un corpus de référence avec une annotation manuelle des inférences et leurs types.

II. Description du corpus

Afin de diversifier les sources, nous travaillons sur les commentaires touristiques de 2 sites :

- booking.com
- mafengwo.cn

Bien que le site donne accès aux hôtels de nombreuses villes dans le monde, nous n'avons conservé que les hôtels de Paris, de manière à disposer d'un corpus homogène puisque les commentaires intègrent parfois des informations sur la proximité de lieux touristiques (Tour Eiffel, Montmartre, etc.). Nous comparons les hôtels à Paris de ces deux sites en prenant compte des informations comme localisation géographique, score, étoile, prix, et commentaire. La partie principale du corpus est des commentaires des utilisateurs avec leurs propres expériences. Les messages laissés par les anciens clients témoignent de différences culturelles et sociales, dans le choix des critères d'évaluation (taille des chambres, présence d'équipements et services dans l'hôtel), dans l'utilisation du vocabulaire (terme générique vs terme spécifique au domaine), et dans la manière d'exprimer une information (en particulier les éléments jugés négatifs). En particulier, cette annotation est dans le but de mettre en évidence les rôles des inférences dans une opinion exprimée en distinguant leurs types.

III. Niveaux d'analyse

Le corpus d'annotation contient 1000 commentaires que nous annoterons avec 3 niveaux d'analyse différents:

1. Commentaire: est un ensemble une/des phrases écrites par un utilisateur. Il peut contenir des opinions sur de différents aspects. Par exemple, un commentaire va décrire en même temps la propreté de la chambre, la localisation de cet hôtel, et la qualité des services.
2. Phrase: est un des morceaux dans un commentaire, segmenté par des ponctuations de fin de phrase, comme '。？ ！… . Si la phrase n'est qu'une phrase simple, nous nous arrêtons le découpe. Si la phrase contient plusieurs propositions, nous allons continuer à le séparer dans l'étape suivante.
3. Syntagme: est une proposition d'une phrase complexe, segmentée par des ponctuations comme « , ； . Un syntagme contient une opinion complète ou une partie de l'opinion.

II. Objectifs

Le travail d'annotation portera sur 3 niveaux d'analyse : (i) le commentaire complet, (ii) chaque phrase de chaque commentaire, et (iii) chaque syntagme de chaque phrase. Cette annotation a plusieurs objectifs:

1. pour chaque niveau d'analyse, détecter s'il existe une/des inférence(s), en commençant par le deuxième niveau ; si une ou plusieurs inférences ne peuvent être détectées qu'entre plusieurs phrases, alors on reviendra au niveau supérieur
2. préciser le(s) type(s) d'inférence;
3. localiser la/les inférences dans le premier niveau d'analyse;
4. analyser la relation des syntagmes constituant une inférence.

IV. Annotation

1. Qu'est-ce qu'une inférence?

Si un commentaire contient un sens implicite et qu'il demande d'une interprétation pour aboutir à une relation binaire de **(thème(sous-thème), polarité)**, nous le considérons d'avoir une inférence. Le « thème » est des sujets auxquels des utilisateurs portent une opinion. La « polarité » est un sentiment positif, négatif ou neutre. Par exemple, « proche de la tour Eiffel » est possible être transformé en cette relation binaire: (localisation(site touristique), positive) via une interprétation inférentielle. Contrairement au commentaire « la localisation est bonne », nous le transformons directement en (localisation+positive). Un commentaire contient une/des inférence(s) seulement si nous avons besoin d'une interprétation inférentielle pour extraire cette relation binaire. Nous avons catégorisé les thèmes comme suivants:

Thème	**Sous-thème**							
Localisation	Géographie	Site touristique	Transport	Course et shopping	Restauration	Sécurité	Environnement	
Équipement	Chambre	Salle de bain	Restaurant (petit-déjeuner)	Région commune	Wifi	Gratuitement	Isolation	Spa
Personnel	Qualité service	Attitude	Compétence des langues					
Propreté	Chambre	Salle de bain	Région commune					
Qualité du service	Prix de l'hôtel	Prix de repas	Navette (site, aéroport)	Autre service dans l'hôtel				

Au moment de la segmentation des phrases en syntagmes, le sens complet d'une opinion peut aussi être séparé. C'est-à-dire qu'un syntagme seul ne contient pas d'inférence alors que l'ensemble des syntagmes le contiennent. C'est pour telle que nous allons annoter l'existence des inférences au niveau des phrases et au niveau des syntagmes. Il y a 3 annotations: avec-inférence, sans-inférence et inférence-incertaine.

A. “**avec-inférence**” est la présence d’une inférence. L’inférence est présentée dans un commentaire seulement si nous avons besoin d’une interprétation pour préciser la polarité d’un commentaire. Par exemple, « La localisation de cet hôtel est bonne » ne contient pas d’inférence, alors que « L’hôtel est à côté de la tour Eiffel » le contient, car le sens littéral du 2e exemple ne montre pas directement un sens positif.

B. ‘**sans-inférence**’ est l’absence de l’inférence. Si un mot de sentiment ou d’opinion est présent dans un commentaire, ceci est considéré comme une absence d’inférence. Par exemple, comme « La localisation de cet hôtel est bonne » contient un mot d’opinion « bonne », nous n’avons pas besoin d’une interprétation pour déterminer la polarité de l’opinion (positive ou négative). Cette phrase ne contient pas d’inférence. Si on peut comprendre l’opinion par une interprétation littérale, cette opinion ne contient pas non plus une inférence. Aussi une interprétation est littérale si elle est sémantiquement équivalente ou synonyme d’une partie du texte, ce qui peut être démontré à l’aide de la grammaire, de la syntaxe et de la connaissance des synonymes. Par exemple, « le réseau wifi reste à renforcer » est traduite littéralement par « le réseau wifi ne fonctionne pas très bien ». Donc il ne contient pas d’inférence.

C. ‘**inférence-incertaine**’ représente « on ne sait pas s’il contient une inférence » (incertain), qui comprend des situations suivantes:

   a) Erreurs orthographiques qui sont incompréhensibles.
   b) Un syntagme qui n’a pas de sens.
   c) Un syntagme qui ne contient pas un sens complet.

Les exemples sont présentés dans le tableau ci-dessous.

2. Annotation de type d’inférence

En basé sur un corpus annoté des inférences, nous allons annoter 5 types d’inférence: inférence logique, pragmatique, lexicale, énonciative, discursive:

I. **Inférences logiques**: se réalisent en découlant du texte. Elles reposent sur un raisonnement formel et mettent en oeuvre un processus logique qui contient induction ou déduction. Par exemple, « 房间 整理 及时 » (Ménage de la chambre est en temps opportun). Le mot “及时” infère que la service de ménage est en bon moment. Nous avons obtenu cette conclusion par le transfert entre des deux unités lexicales synonymes de « en temps opportun » et « bon moment ». Cette phrase contient donc une inférence logique. L’induction correspond à un processus qui permet de passer du particulier (faits observés, cas singuliers, données expérimentales, situations) au général (une loi, une théorie, une connaissance générale). La déduction correspond au processus presque inverse qui permet de conclure (déduire) une affirmation à partir d’hypothèses, de prémisses ou d’un cadre théorique : les conclusions résultent formellement de ces prémisses ou de cette théorie. Pour l’annotation on ne distingue pas entre les deux.

II. **Inférences pragmatiques**: sont possiblement vraies et communes à l’ensemble des lecteurs. Une inférence sera pragmatique si le lecteur moyen (comparé à son groupe d’appartenance), a

tendance à la donner après incitation. Elles s'appuient sur l'ensemble des connaissances acquises par un individu lors de ses expériences passées. Comme le corpus est des commentaires rédigés par les touristes chinois, on considère le groupe d'appartenance partage la même culture sociale chinoise. Par exemple, « 没有烧水壶 » (pas de bouilloire) et « 离老佛爷近 »（proche de Galeries Lafayette）contiennent des inférences pragmatiques, car les deux exemples reflètent la culture diététique et les habitudes d'achat des touristes chinois.

Une méthode optionnelle que nous utilisons pour distinguer entre les inférences logiques et pragmatiques est de relier l'ancien commentaire avec la nouvelle conclusion en ajoutant « mais » et « négation » pour la deuxième proposition. Par exemple:

<texte>: *La personne de l'accueil ne parlait pas l'anglais.*

<interprétation>: *La compétence des langues de la personne de l'accueil n'était pas bonne.*

<texte, mais interprétation en négation>: *La personne de l'accueil ne parlait pas l'anglais, mais la compétence des langues de la personne de l'accueil était bonne.*

Si cette combinaison est « acceptable » (c'est-à-dire qu'elle n'est pas paradoxale), on suppose que les inférences sont pragmatiques, sinon logiques. Ceci est seulement un critère pour distinguer les inférences logiques et pragmatiques. Nous l'appellerons « mais+négation » dans la suite.

III. **Inférences lexicales**: sont la phrase en dehors de tout cadre énonciatif et dépendent exclusivement de l'articulation entre les prédicats et leurs arguments. Le cadre minimal d'analyse des inférences lexicales est la phrase élémentaire, définie comme une relation entre un prédicat de premier ordre et ses arguments. C'est cette relation qui permet d'interpréter les unités lexicales, et d'expliciter toutes les inférences qui leur sont rattachées. Cela signifie, d'une part, que les propriétés sémantiques sont subordonnées à la syntaxe, et, d'autre part, que les unités lexicales sont étudiées strictement du point de vue de leurs propriétés linguistiques. Les inférences lexicales construisent le sens à partir des structures prédicatives (prédicats et arguments).

A. Métonymie
B. Unité lexicale spécifique dans le domaine de l'hôtellerie portant un sentiment ou une opinion: si un mot qui n'est pas un mot de sentiment ou d'opinion, mais qui porte une polarité dans le domaine de l'hôtellerie, nous le considère comme une inférence lexicale. Par exemple, « aération », « anglais » ou « cafard ».
C. Absence de connecteur reliant deux phrases simples (la parataxe). Cette absence est palliée par l'inférence. Très souvent, les connecteurs de cause sont absents et c'est l'inférence qui prend le relais. Par exemple, « Je n'arrive pas à me connecter. Mon téléphone est parfaitement fonctionnel. » Un « mais » implicite entre les deux propositions infère un connecteur de cause.

IV. **Inférences énonciatives**: sont actualisées en contexte. Les inférences énonciatives se fondent sur le sens ainsi construit pour l'inscrire dans une situation énonciative. C'est-à-dire que le cadre

d'une inférence est un énoncé au lieu d'un mot ou un entraînement des énoncés. Plusieurs mécanismes sont en jeu comme la négation, la polyphonie, les savoirs partagés, etc.

A. Dans les interactions clients, l'interprétation d'un dysfonctionnement ou d'un événement négatif dans le parcours du client s'explique par ces savoirs partagés. Cela repose très souvent sur une comparaison implicite entre une situation « normale » attendue et une situation présentée comme « perturbée » par le locuteur.

1. Des adverbes qui déterminer la polarité d'une phrase: "入住三天，只有第一天提供了瓶装水" (Parmi les trois jours, les bouteilles d'eau étaient offertes **seulement** le premier jour) implique la situation attendue est « chaque jour il y a des nouvelles bouteilles d'eau offertes ».
2. Des phrases interrogatives: "捡到东西难道不主动联系客人？" (Ne devriez-vous pas contacter immédiatement les clients une fois voir des objets trouvés ?)
3. Des déterminants d'occurrences multiples et de fraction: "找前台要了两次吹风机无果" (J'ai demandé **deux fois** à l'accueil pour avoir un sèche-cheveux, pas de résultat); "客房 价格 仅 是 旺季 的 三分之一" (Le prix des chambres ne représente qu'**un tiers** par rapport en saison).

B. Reproche ou appréciation implicite: 敲门敲了一下，还没来得及应门，员工就进来了。(Le personnel a frappé la porte une seule fois et je n'ai même pas eu le temps de répondre, ce personnel est entré.) Cette phrase ne mentionne pas directement que la qualité de service n'est pas bonne, mais elle implique un reproche du personnel de l'hôtel.

V. **Inférences discursives**: interfèrent avec des connaissances extralinguistiques au niveau du discours. Le niveau d'analyse de ces inférences n'est plus la phrase ni l'énoncé, mais ce que l'on peut appeler des situations et qui se définissent comme la conjonction de paramètres linguistiques directement extraits du texte et de connaissances pragmatiques. L'interprétation et la mise en place d'une situation font appel à des inférences linguistiques auxquelles s'ajoutent des connaissances extérieures au texte. Ces connaissances permettent notamment de moduler les premières inférences linguistiques et de leur ajouter de nouvelles significations. Dans notre cas, si des mots clés d'un thème ne sont pas mentionnés dans un commentaire et que nous impliquons le thème par le contenu hors linguistique, nous le considérons d'avoir une inférence discursive. Par exemple, selon le contenu de la phrase "酒店送了一小盒巧克力给女儿" (L'hôtel a offert une boîte de chocolat à ma fille), nous résumons que l'attitude du personnel est bonne.

Les exemples des types d'inférences sont présentés ci-dessous.

Type d'inférence	定义	Définition	Exemple	Traduction
Logique	结论是基于文本的字面意思的推断。没有引入文本以外的信息。 用mais+négation的方法，得到的结果**不可以**被接受。充分必要条件	Elles sont nécessairement vraies et reposent sur un raisonnement formel et mettent en oeuvre un processus logique (induction et déduction). Le résultat de la méthode « mais+négation » est inacceptable.	酒店在景点附近。 早餐 每天 都 一样 房间 真的 太小了， 在 里面 转 个身 都 困难 。	L'hôtel est à côté des sites touristiques. Le petit-déjeuner sont pareil tous les jours. La chambre est très petite. Il est même difficile de se tourner dedans.
Pragmatique	需要运用文本以外的个人认知体系作出判断。同一文化背景下的用户，大部分会给出统一结果。 用mais+négation的方法，得到的结论**可以**被接受。充分不必要条件	Elles sont possiblement vraies et communes à l'ensemble des lecteurs. Elles s'appuient sur l'ensemble des connaissances acquises par un individu lors de ses expériences passées. Le résultat de la méthode « mais+négation » est acceptable.	前台不讲英文。 附近 就 有 家乐福 和 多家 餐厅 。 离 卢浮宫 步行 10 分钟 的 距离 。	La personne de l'accueil ne parlait pas l'anglais. Il y a un carrefour et plusieurs restaurants à côté. 10 minutes à pieds à musée du Louvre
Lexicale	1) 借代和转喻等。 2) 只在酒店领域才含有情感意义的词汇 3) 简单句之间缺乏连接词，两个断句之间的关系需要inference推断。	1) Métonymie 2) Unité lexicale spécifique dans le domaine de l'hôtellerie portant un sentiment ou une opinion 3) Absence de connecteur pour la parataxe. Cette absence est palliée par l'inférence	1) 我连不上网。 2) 通风，蟑螂 3) 连不上网。手机其他运行都正常。	1) Je suis déconnecté. 2) aération, cafard 3) Je n'arrive pas à me connecter. Mon téléphone est parfaitement fonctionnel.

Type d'inférence	定义	Définition	Exemple	Traduction
Énonciative	1) 陈述中隐含与正常情况相反的情， （突出情感倾向的副词，反问句，倍数或分数的副词以突出与正常情况的比较） 2) 暗示责备或赞扬	1) comparaison implicite à une situation « normale » (des déterminants d'occurrences multiples et de fraction, des phrases interrogatives ou rhétorique, des adverbes qui déterminer la polarité d'une phrase) 2) Reproche ou appréciation implicite	1) 入住三天，只有第一天提供了瓶装水。 1) 捡到东西难道不主动联系客人? 早餐 是 面包店 现成 的 面包 ， 却 卖 贵 了 好几倍 的 价钱 2) 敲门敲了一下，还没来得及应门，员工就进来了。	1) Parmi les 3 séjours, il n'y avail que le premier jour où les bouteilles d'eau étaient offertes. 1) Ne devriez-vous pas contacter immédiatement les clients une fois voir des objets trouvés ? 2) Le personnel a frappé porte une seule fois et je n'ai même pas eu le temps de répondre, ce personnel est entré.
Discursive	用超语言学的认知进行推理。比如文中没有出现主题 (thème) 的任何相关关键字, 我们也能推断出主题。 特别是用来判断评论中的隐含主题。	interfèrent avec des connaissances extra-linguistiques au niveau du discours. En particulier, quand les mots-clés des thèmes ne sont pas présents, on peut interpréter ses thèmes selon des connaissances extra-linguistiques.	酒店送了一小盒巧克力给女儿。 我们 住 了 四天 从 没 碰到 小偷。 我们 五个 人 带 着 行李 得分 三批 上 楼 。	L'hôtel a offert une boîte de chocolat à ma fille. Nous avons habité pendant quatre jours, jamais rencontré un voleur。 Nous cinq avec les bagages devons séparés en 3 lots pour monter.

3. Comment faire une annotation avec l'outil Brat?

Installation Brat

Les étapes de l'installation sont présentées sur la page https://brat.nlplab.org/installation.html. Brièvement, nous allons d'abord téléchargé brat v1.3 (https://brat.nlplab.org/) et dans le même répertoire lancer :

```
tar xzf brat-v1.3_Crunchy_Frog.tar.gz
cd brat-v1.3_Crunchy_Frog
./install.sh
```

Ensuite, il faut décompresser l'archive (inference) que nous avons préparée dans le répertoire data/. Enfin, nous commençons Brat par lancer python standalone.py dans le terminal et coller http://127.0.0.1:8001 dans un navigateur.

Une phrase ou un syntagme?

Pour chaque commentaire, il faut vérifier la présence des inférences au deux niveaux: 1) si une phrase contient des inférences? 2) si chaque syntagme à l'intérieur de cette phrase contient des inférences? Il est possible que la phrase et un syntagme de cette phrase contiennent tous les inférences.

Présence d'inférence, quels types et quel thème?

Si une phrase ou un syntagme présente des inférences, il y a 3 démarches à faire sur Brat:

1. sélectionner la portion pour activer la fenêtre d'édition,
2. choisir un type:

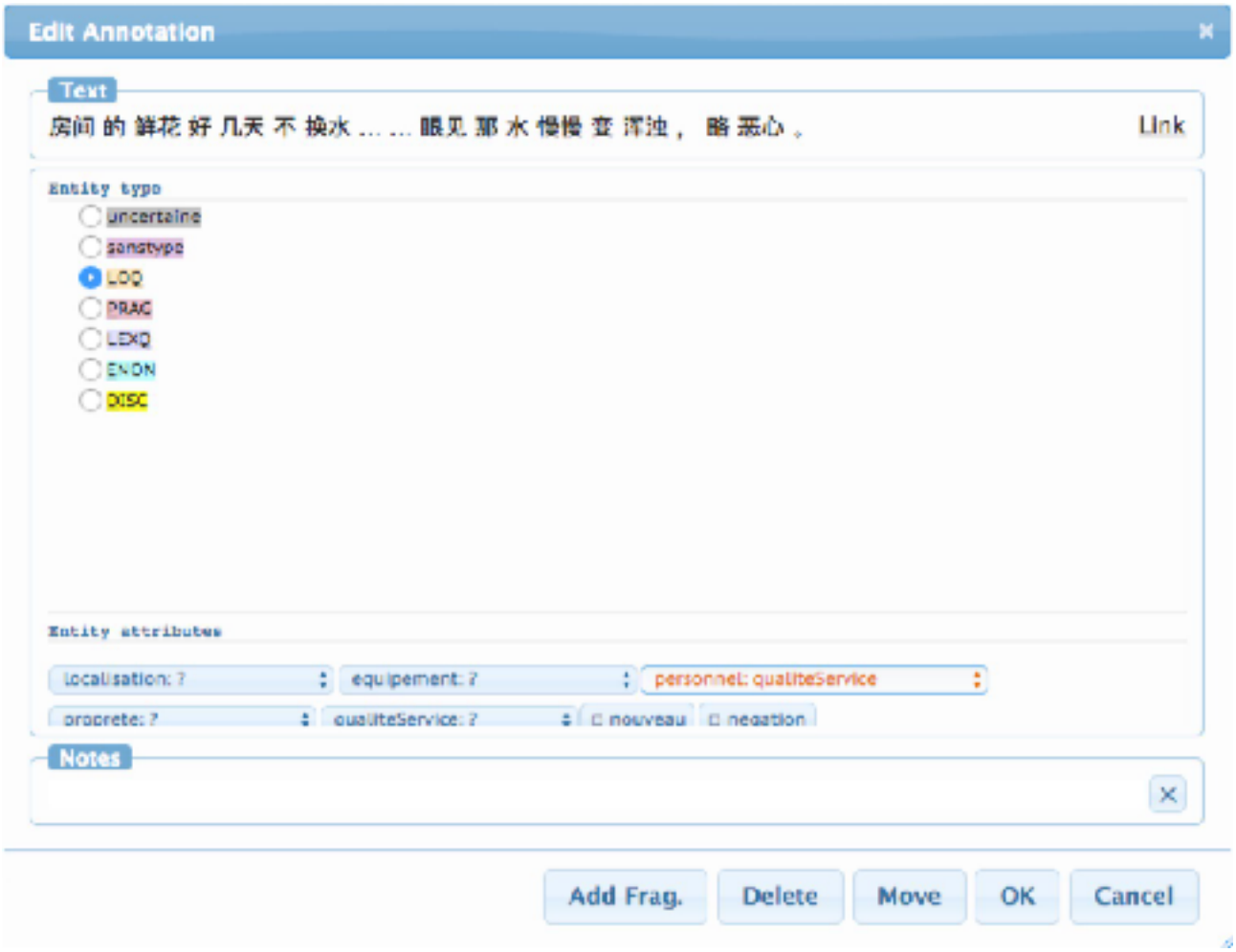
   LOGIQUE
   PRAGMATIQUE
   LEXICALE
   ENONCIATIVE
   DISCURSIVE
   INFERENCE-SANSTYPE: je suis sûr qu'il y a une inférence, mais je ne sais pas de quel type il s'agit parmi les types disponibles (possiblement, c'est une inférence d'un type nouveau, non prévu dans le guide)
   (attention: si cette portion contient plusieurs types d'inférence (c'est souvent le cas), il faut resélectionner cette portion pour ajouter un type chaque édition.)
   Pour faciliter l'annotation, vous pouvez utiliser les raccourcis de chaque type:

L	LOGIQUE
P	PRAGMATIQUE
C	LEXICALE
E	ENONCIATIVE
D	DISCURSIVE
S	INFERENCE-SANSTYPE

3. choisir un thème correspond à cette opinion. Si vous ne trouvez pas un thème convenable, vous pouvez choisir nouveau.

Exemple d'annotation

ID	Phrases/syntagmes	Inférences	Logiques	Pragmatiques	Lexicales	Énonciatives	Discursives
42-1	所在的区域有些乱，如果是女生单独旅行不建议住这。	1	1	-1	-1	1	-1
42-1-1	所在的区域有些乱，	-1					
42-1-2	如果是女生单独旅行不建议住这。	1	1	-1	-1	1	-1
42-2	旅舍的空间小，走廊没法并排走两个人；浴室也仅够转身。	1	1	-1	-1	-1	-1
42-2-1	旅舍的空间小，	-1					
42-2-2	走廊没法并排走两个人；	1	1	-1	-1	-1	-1
42-2-3	浴室也仅够转身。	1	1	-1	-1	-1	-1
42-3	房间里没有储物空间。	1	1	-1	-1	1	-1
42-4	wifi信号不稳定，速度慢。	-1					
42-4-1	wifi信号不稳定，	-1					
42-4-2	速度慢。	0					
42-5	我在的几天厨房坏了不能用。	1	1	1	-1	1	-1


Edit Annotation
Text
房间 的 鲜花 好 几天 不 换水 ... ... 眼见 那 水 慢慢 变 浑浊 ， 略 恶心 。
Link
Entity type
uncertaine
sanstype
LOQ
PRAG
LEXQ
ENON
DISC
Entity attributes
localisation: ?
equipement: ?
personnel: qualiteService
proprete: ?
qualiteService: ?
nouveau
negation
Notes
Add Frag.
Delete
Move
OK
Cancel

V. Un exercice avant de commencer

ID	Phrases/syntagmes	Inférences	Logiques	Pragmatiques	Lexicales	Énonciatives	Discursives
1-1	我们 8月 7日 上午 外出 、 下午 1点钟 回到 公寓 ， 打扫卫生 的 员工 已经 进入 公寓 整理 了 房间 。						
2-1	员工 特别 热情						
3-1	非常 惊喜 ， 住 的 房间 有 浴缸						
3-1-1	非常 惊喜 ，						
3-1-2	住 的 房间 有 浴缸						
4-1-1	房间 宽敞 干净 ， 前台 很 是 热情好客 ， 很 满意 。						
5-1	附近 很多 快餐店 和 酒吧 。						
6-1	酒店 浴池 卫生间 太小 ， 床单 ， 浴巾 随便 放 地上 ， 整间 房间 插头 插座 只有 2个						
7-1	早餐 简单 需 改进 。						
8-1	地理位置 优越 ， 距离 《 天使 爱美丽 》 电影 中 的 酒吧 很 近						

Les réponses:

ID	Phrases/ syntagmes	Inférences	Logiques	Pragmatiques	Lexicales	Énonciatives	Discursives
1-1	我们 8月 7日 上午 外出、 下午 1点钟 回到 公寓， 打扫卫生 的 员工 已经 进入 公寓 整理 了 房间 。	1	1, parce que la conclusion vient du déroulement du texte.	-1	1, le mot « 已经 » implique une surprise positive.	-1	-1
2-1	员工 特别 热情	-1, le syntagme contient explicitement la relation binaire (personnel(service), positive)					
3-1	非常 惊喜，住 的 房间 有 浴缸	1	-1	1, besoin de connaissance personnelle pour savoir que le baignoire n’est pas un équipement commun dans l’hôtel	1, le mot « baignoire » porte un sens positif	1, entre les deux propositions, on infère un « car » pour la causalité.	-1
3-1-1	非常 惊喜，	-1					
3-1-2	住 的 房间 有 浴缸	1	-1	1	1	-1, quand on segmente deux proposition. L’inférence énonciative est absente.	-1

ID	Phrases/ syntagmes	Inférences	Logiques	Pragmatiques	Lexicales	Énonciatives	Discursives
4-1-1	房间 宽敞 干净， 前台 很 是 热情好客， 很 满意 。	-1, même si la phrase contient 3 propositions, elles sont toutes explicites.					
5-1	附近 很多 快餐店 和 酒吧 。	1	-1	1, interprétation est positive car la description est sur l'environnement d'un hôtel	-1	-1	-1
6-1	酒店 浴池 卫生间 太小， 床单， 浴巾 随便 放 地上，整间 房间 插头 插座 只有 2 个	1	-1	1, seulement deux prises ne sont pas suffisantes.	-1	1, « les serviette sont par terre » implique la situation normale ou attendue est que les serviettes sont bien rangés.	-1
7-1	早餐 简单 需 改进 。	-1					
8-1	地理位置 优越， 距离 《 天使 爱美丽 》 电影 中 的 酒吧 很 近	1	-1	1, la connaissance du film est un savoir partagé extérieur du texte.	-1	-1	-1

Annexe B

Guide d’annotation (version chinoise)

inférence人工标注指南

inférence人工标注指南 1
I. 简介 2
II. 语料介绍 2
III. 分析层级 2
IV. 标注目标 2
V. 人工标注 2
V.I 什么是inférence? 2
V.II Inférence的类型标注 4
VI. 怎样用Brat标注 6
VI.I 句子和语段 6
VI.II 需要标注的内容 6
VI.III 运行Brat: 6
注意事项 7

I. 简介

信息挖掘与情感分析的领域中，机器能更好的诠释明确表达的观点，比如有明确的情感词，或者表达方式是明确的，不具备隐含意义的。 但是，一条评论经常是一段对于经历的称述，整段称述间接表达了作者的态度。或者是，评论中含有推理，需要读者的解读才能得到结论。为了在机器学习的过程中，让机器习得一个正确的样本语料，我们需要进行关于inférence和其类型的人工标注。

II. 语料介绍

语料来源于两个网站，booking和马蜂窝。所有的语料都是关于巴黎酒店的中文评论。之所以只保留了巴黎的酒店，是为了保证分析阶段文本的一致性，比如酒店评论中常会提及旅游景点的名称（埃菲尔铁塔，蒙马特高地），不同的城市，这一类的词汇也将不同。我们对比两个网站的文本语料的同时，也保留了关于地理位置，酒店评分，酒店星级等信息。当然主体文本还是用户留下的关于在此酒店中的体验和评价。由于用户的不同文化和教育背景，这些评论文本所提及的方面不同（比如房间大小，酒店设施，服务等），所使用的词汇等级也不相同（比如常见词汇和酒店业词汇或者用词的难度等）。此次标注任务旨在找出句子是否含有inférence以及它们的类型。

III. 分析层级

第一部分标注文本包含了1000条评论，我们将其分为三个层级：

1）评论层级（commentaire）：为一个用户写下的一段完整评论，是包含一个或多个句子的整条评论，可以同时提及多个主题。例如一条评论可以同时描述房间的清洁度，酒店地理位置和服务质量；

2）句子层级（phrase）：评论的其中一句话，被。？！……等标点符号分隔。如果只是一个简单句，我们不继续进入下一层级的分割。如果是复杂句，我们将进行下一层级的分隔。

3）语段层级（syntagme）：复杂句的其中一个从句，以，；分隔。一个语段可能包含一句完整评论或仅是一部分评论。

IV. 标注目标

标注工作只对其中两个层级（句子和语段）进行标注：1）评论中的每一个句子；2）每一个句子中的每个语段。目的在于：

1. 从第二层级开始，检测是否含有inférence. 如果inférence不能存在于较低的层级，我们就回到上一层级；
2. 明确inférence的类型；
3. 找到inférence在第一层级中的位置；
4. 分析共同组成一个inférence的不同语段间的关系。

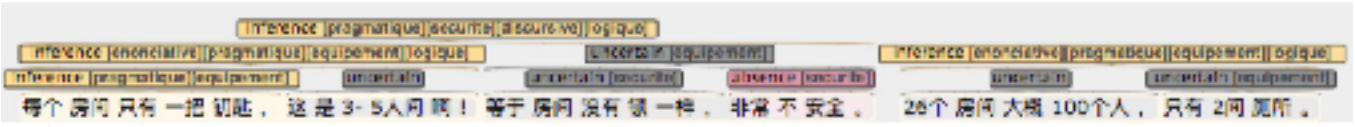
V. 人工标注

V.I 什么是inférence?

如果一条评论含有隐含意义，或者需要读者的解读才能得出两分结构的结论，比如（主题（副主题），褒贬性），我们就把这样的评论视作含有inférence。其中，“主题”是用户评论的方面（对象）。“褒贬性”指的评论是褒义，贬义还是中性的情感倾向。比如通过interprétation inférentielle, “靠近埃菲尔铁塔”可以被转化为（位置（旅游景点），好）。与“位置很好”不同，因为它可以被直接转化为（位置，好）。**一条评论含有inférence仅当它需要一次或多次推理来得到一个两分结构的关系。**我们把主题和副主题归纳如下：

在把句子分隔成语段时，完整的观点也可能被分割开来。也就是说，有时一个语段不含有inférence，但是几个这样的语段组成一句话时，就含有了。因此对于是否含有inférence，我们将进行不同层级的标注。

A. **INFERENCE**： 包含一个或多个inférence
B. **ABSENCE**: 不包含任何inférence。如果句子的褒贬意是通过文本中出现的情感词得到的，比如“服务热情”中的“热情”，那么该句中不包含inférence。或者得到二次结构的过程，没有经过任何的推理过程，则不包括inférence。
C. **UNCERTAIN**：指标注者不能确定该句中是否含有inférence. 比如由于拼写错误引起的歧义，被分割的语段意义不完整等。



Thème	Sous-thème							
Localisation	Géographie	Site touristique	Transport	Course et shopping	Restauration	Environnement		
Équipement	Chambre	Salle de bain	Restaurant (petit-déjeuner)	Région commune	Wifi	Gratuitement	Isolation	Spa
Personnel	Qualité service	Attitude	Compétence des langues					
Propreté	Chambre	Salle de bain	Région commune					
Qualité du service	Hôtel	Restauration	Navette (site, aéroport)	Autre service dans l’hôtel				
Prix	Hôtel	Restauration	Autre service					
Sécurité	Intérieur de l’hôtel	Extérieur de l’hôtel						
Général	Impression globale	Intégrité						

V.II Inférence的类型标注

基于上一阶段，即是否含有inférence的基础上，我们将inférence划分为5个类型：inférence logique, pragmatique, lexicale, énonciative, discursive.

I. Inférence logique: 基本文本的发展而形成；基于上下文的逻辑推理过程。比如“早餐 每天 都 一样”中没有提及任何情感评价词，但是根据对上下文逻辑的理解，推断出“早餐种类平淡”，包含了逻辑推理。

II. Inférence pragmatique：大多数读者都是这样理解的，但是推理的过程中，加入了读者自己的文化背景和理解，或者是，推理过程中需要首先知道评论的背景，比如要理解“在铁塔附近”为褒义，须知道评论的文本是巴黎的酒店，因此“铁塔附近”指的是“地理位置好”。再比如，读者与作者需要有统一的文化背景，才能理解其褒贬的评论，也含有inférence pragmatique。比如“没有烧水壶”，只有中国人才喝热的白开水，所以对于其他国家的人来说，这条评论只是中性，而对于中国游客来说，就是贬义。

III. Inférence lexicale：词汇层面。包括两种具体情况：
   A. 借喻，转喻 （很少出现）
   B. 在不考虑上下文的情况下，词汇本身不含褒贬情感，但是在酒店领域的评论里却含有情感色彩，比如“通风”，“英语”，“wifi”，“插座”，“老佛爷”，“蟑螂”以及所有的巴黎景点等等。

IV. Inférence énonciative：基于一个句子或语段层面的，而不是词汇或者几个句子组成的段落。也就是说一个简单句或者几个（，；或空格分隔开的）语段构成的一个句子。这样句子通常围绕一个主题进行阐述。包含以下几种情况：
   A. 用户对负面情绪（评价）的阐述，常会通过与正常情况（预期情况）的隐含比较表达出来。比如：
      1. 副词的运用：“入住三天，只有第一天提供了瓶装水”
      2. 反问句或疑问句：“捡到东西难道不主动联系客人？”
      3. 数量词的运用：“找前台要了**两次**吹风机无果”
   B. 责备或表扬的隐含表达：“敲门敲了一下，还没来得及应门，员工就进来了。”
   C. 两个简单句之间，没有连接词，但是这两个简单句构成因果从句等，缺少的连接词是读者通过inférence得出的。比如“我连不上网，手机是正常的“。这两个简单句之间的”但是“，是由读者自己得出的。

V. Inférence discursive：基于几个简单句，或者复杂句，（由于网络评论标点符号的不规范性，）有时也可以是几个语段，主要在于阐述的主题多样，并且产生inférence的过程是借助于extralinguistique实现的。
   A. 比如几个描述性的句子共同阐述了一种观点。这些句子可以是连续的，也可以是被分割开的；
   B. 关于主题的关键词没有出现在文本中，对于主题的推断是根据常识和上下文一起推断出来的。

注意：

1. **一个对象可以同时拥有几个不同类型的inférence。**
2. **区别inférence logique 和 inférence pragmatique 的方法之一：mais+négation**

   原句子：酒店在景点附近。

   结论：酒店地理位置好。

   mais+négation: 酒店在景点附近，但是酒店地理位置不好。

   悖论（不成立）：logique

   接受（成立）：pragmatique

Type d'inférence	**定义**	**Exemple**
Logique	结论是基于文本的字面意思的推断。没有引入文本以外的信息。 用mais+négation的方法，得到的结果**不可以**被接受。充分必要条件	酒店在景点附近。 早餐 每天 都 一样 房间 真的 太小 了， 在 里面 转 个 身 都 困难 。
Pragmatique	需要运用文本以外的个人认知体系作出判断。同一文化背景下的用户，大部分会给出统一结果。 用mais+négation的方法，得到的结论**可以**被接受。充分不必要条件	前台不讲英文。 附近 就 有 家乐福 和 多家 餐厅 。(pragmatique et logique) 离 卢浮宫 步行 10分钟 的 距离 。
Lexicale	1) 借代和转喻等。 2) 只在酒店领域才含有情感意义的词汇	2) 通风，蟑螂
Énonciative	1) 陈述中隐含与正常情况相反的情， （突出情感倾向的副词，反问句，倍数或分数的副词以突出与正常情况的比较） 2) 暗示责备或赞扬 3) 简单句之间缺乏连接词，两个断句之间的关系需要inference推断。	1) 入住三天，只有第一天提供了瓶装水。 1) 捡到东西难道不主动联系客人? 早餐 是 面包店 现成 的 面包， 却 卖 贵 了 好几倍 的 价钱 2) 敲门敲了一下，还没来得及应门，员工就进来了。 3) 连不上网。手机其他运行都正常。
Discursive	A. 比如几个描述性的句子共同阐述了一种观点。这些句子可以是连续的，也可以是被分割开的 B. 关于主题的关键词没有出现在文本中，对于主题的推断是根据常识和上下文一起推断出来的。	每个 房间 只有 一把 钥匙， 这 是 3- 5人间 啊！ 等于 房间 没有 锁 一样， 非常 不 安全 。 酒店送了一小盒巧克力给女儿。 我们 住 了 四天 从没 碰到 小偷。 我们 五个 人 带 着 行李 得分 三批 上楼 。

VI. 怎样用Brat标注

VI.I 句子和语段

对于每一行的评论，首先以句子为单位标注，然后以语段为单位标注。也就是说句子层级和语段层级可以同时含有inférence。当以语段为单位时，如句子意义不完整，即可标注为uncertain（表示不确定是否含有inférence）；如语段中出现了明确的情感词汇和主题，则标注为absence。

VI.II 需要标注的内容

1. 选中对象，标点符号和空格归前一句所有。
2. 选择是否含有inférence：

INFERENCE: 有

从五个类型中选择与选中对象符合的类型：

logique
pragmatique
lexicale
enonciative
discursive
sans_type: je suis sûr qu'il y a une inférence, mais je ne sais pas de quel type il s'agit parmi les types disponibles (possiblement, c'est une inférence d'un type nouveau, non prévu dans le guide)

ABSENCE: 没有

UNCERTAIN: 由于意思和主题的不完整性，不能确定是否包含inférence

选择主题：

localisation：geographie|touristique|transport|shopping|restauration|securite|environnement

equipement：chambre|salleDeBain|restauration|wifi|regionCommune|gratuitement|isolation|autre

personnel： qualiteService|attitude|competenceDesLangues

proprete： chambre|salleDeBain|regionCommune

qualiteService：hotel|repas|navette|autre

securite：intérieur| extérieur de l'hôtel

prix: repas | hôtel | service

general: 总体印象（评价），诚信

mixte: 同时包含了几个主题，主要是针对长句子

nouveau：选项中没有，可以把自己认为适合的主题输入Notes框内。

VI.III 运行Brat:

打开终端：terminal（konsole）

依次输入：

```
cd Bureau/brat-v1.3_Crunchy_Frog
python standalone.py
```

复制终端显示的网页地址至浏览器

右上角点击登陆，输入用户名和密码，用户名和密码均为ertim

双击进入文件夹inference/，双击1-50.txt文档

选中一段文字，编辑框将自动弹出。

关闭工具：在终端输入 ctrl+c，再关闭网页。

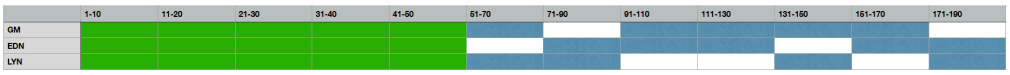
注意事项

1. 认真！如果标注结果不正确，会影响接下来的研究结果和标注文本的可用性。每天的工作，我都会做一个交叉对比。
2. 特别要注意标注的层级是句子和语段。不能标注完句子以后，忽略了其中的语段。除了，句子层级不包括inférence, 即ABSENCE时，可以忽略语段。
3. 由于三个人的标注文本都相同，标注过程中不能讨论，如有疑问，可以直接问我。标注结束后，我会总结结果，给出bilan时，再进行讨论。
4. 过程中如果遇到问题，请记下具体问题和对应的文本名和行数。
5. 每天的工作完成后，请将data/inference/*.ann文件发送至我的邮箱liyun.yan@inalco.fr。

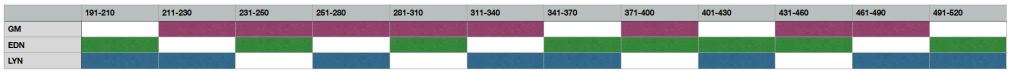
Annexe C

Exemple du planning des annotations

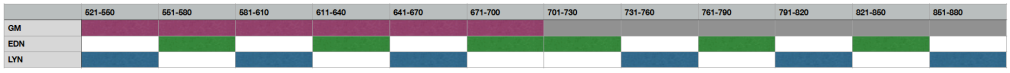
Semaine 1 (total 200 commentaires)



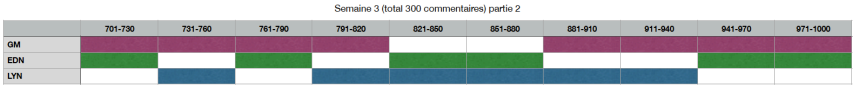
Semaine 2 (total 250 commentaires)



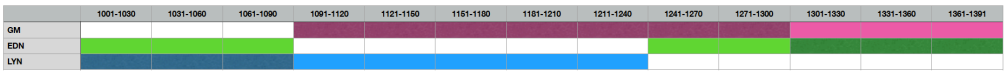
Semaine 3 (total 250 commentaires) partie 1



Semaine 3 (total 300 commentaires) partie 2



Semaine 4-5 (total 391 commentaires)



Annexe D

Publications

D.1 Analyse des inférences pour la fouille d'opinion en chinois, CORIA-TALN-RJC, May 2018, Rennes, France

Résumé La fouille d'opinion est une activité essentielle pour la veille économique, facilitée par les réseaux sociaux et forums dédiés. L'analyse repose généralement sur des lexiques de sentiments. Pourtant, certaines opinions sont exprimées au moyen d'inférences. Dans cet article, nous proposons une classification des inférences utilisées en chinois dans des commentaires touristiques, à des fins de fouille d'opinion, selon trois niveaux d'analyse (réalisation sémantique, modalité de réalisation, et mode de production). Nous démontrons l'intérêt d'analyser les différents types d'inférence pour déterminer la polarité des opinions exprimées en corpus. Nous présentons également de premiers résultats fondés sur des plongements lexicaux.

D.2 Inference Annotation of a Chinese Corpus for Opinion Mining, LREC, May 2020, Marseille, France

Abstract Polarity classification (positive, negative or neutral opinion detection) is well developed in the field of opinion mining. However, existing tools, which perform with high accuracy on short sentences and explicit expressions, have limited success interpreting narrative phrases and inference contexts. In this article, we will discuss an important aspect of opinion mining : inference. We will give our definition of inference, classify different types, provide an annotation framework and analyze the annotation results. While inferences are often studied in the field of Natural-language understanding (NLU), we propose to examine inference as it relates to opinion mining. Firstly, based on linguistic analysis, we clarify what kind of sentence contains an inference. We define five types of inference : logical inference, pragmatic inference, lexical inference, enunciative inference and discursive inference. Second, we explain our annotation framework which includes both inference detection and opinion mining. In short, this manual annotation determines whether or not a target contains an inference. If so, we then define inference type, polarity and topic. Using the results of this annotation, we observed several correlation relations which will be used to determine distinctive features for automatic inference

classification in further research. We also demonstrate the results of two preliminary classification experiments.

Liyun Yan

Le rôle des inférences pour la fouille d'opinion : applications aux réseaux sociaux en langue chinoise

Résumé

Cette thèse s'intéresse à l'inférence linguistique dans la fouille d'opinion dans un corpus des commentaires touristiques en chinois. Les techniques existantes qui sont bien développées sur des opinions courtes et explicites donnent des résultats limités dans l'interprétation des contextes implicites. De plus, l'expression de l'opinion met en œuvre des stratégies énonciatives différentes suivant les langues et les cultures. Notre hypothèse de travail consiste à étudier les inférences pour améliorer la fouille d'opinion. Dans cette perspective, notre première contribution propose une typologie des inférences pour le chinois en 5 types: logique, pragmatique, lexicale, énonciative et discursive (Rossi et Campion, 1999; Marin, 2004; Duchêne, 2008; Doucy et Massoussi, 2012). Nous avons appliqué cette typologie pour annoter un corpus, dans l'objectif de mener des expériences de fouille d'opinion avec et sans le traitement des inférences. Notre deuxième contribution se focalise sur la classification automatique des inférences en nous basant sur les caractéristiques linguistiques, les métadonnées du domaine et les vecteurs du plongement de mots. L'objectif d'une part est de prouver que le traitement des inférences améliore la performance de la fouille d'opinion et d'autre part de trouver une solution équilibrée entre la classification manuelle couteuse et la classification automatique. Dans ce travail de thèse, nous avons démontré l'intérêt d'étudier les inférences pour réaliser une fouille d'opinion en chinois. Toutefois, l'identification automatique des inférences reste complexe et nécessite une poursuite des travaux de recherche.

Mots-clés : Fouille d'opinion, Inférence, Langue chinoise, Réseaux sociaux

Résumé en anglais

This thesis is interested in linguistic inference in opinion mining in a corpus of tourist commentaries in Chinese. Existing techniques which are well developed on short and explicit opinions, give limited results in interpreting implicit contexts. In addition, expression of opinion implements different enunciative strategies according to languages and cultures. Our hypothesis consists in studying inferences to improve opinion mining. In this perspective, our first contribution proposes a typology of inferences for Chinese in 5 types: logical, pragmatic, lexical, enunciative and discursive (Rossi and Campion, 1999; Marin, 2004; Duchêne, 2008; Doucy and Massoussi, 2012). We applied this typology to annotate a corpus, with the objective of conducting opinion mining experiments with and without the processing of inferences. Our second contribution focuses on automatic classification of inferences based on linguistic characteristics, domain metadata and word embedding vectors. The objective on the one hand is to prove that the processing of inferences improves the performance of opinion mining and on the other hand to find a balanced solution between expensive manual annotation and automatic classification. In this thesis, we demonstrated the interest of studying inferences for opinion mining in Chinese. However, the automatic identification of inferences remains complex and requires further research.

Keywords : Opinion Mining, Inference, Chinese Language, Social Network