# Analyse des inférences pour la fouille d'opinion en chinois

Liyun Yan

# Analyse des inférences pour la fouille d'opinion en chinois

LiYun Yan
ERTIM, Inalco, 75007 Paris, France
liyun.yan@inalco.fr

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.

ABSTRACT

**Analysis of Inferences in Chinese for Opinion Mining**

Opinion mining is an essential activity for economic watch, made easier by social networks and ad hoc forums. The analysis generally relies on lexicon of sentiments. Nevertheless, some opinions are expressed through inferences. In this paper, we propose a classification of inferences used in Chinese in tourist comments, for an opinion mining task, based on three levels of analysis (semantic realization, modality of realization and production mode). We proved the interest to analyze the distinct types of inferences to identify the polarity of opinions expressed in corpora. We also present some results based on word embeddings.

MOTS-CLÉS : Inférences, fouille d'opinion, polarité.

KEYWORDS: Inferences, Opinion Mining, Polarity.

## 1 Introduction

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 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 raison du nombre élevé de commentaires disponibles

sur Internet, il est nécessaire de disposer d'outils automatiques de fouille d'opinion pour analyser le contenu et dégager les tendances exprimées. En outre, les commentaires contiennent une grande quantité d'inférences qui demandent une analyse plus profonde pour la fouille d'opinion. Dans cet article, nous présentons les différents types d'inférence (section 2). Puis, à partir de notre corpus (section 3), nous proposons une classification des types d'inférences fondée sur trois niveaux d'analyse et mettons en évidence les variantes linguistiques en chinois. Enfin, nous présentons les résultats que nous obtenons en corpus (section 4) et présentons les résultats de plongements lexicaux appliqués en corpus pour mettre en évidence la relation des éléments d'une inférence dans un espace vectoriel.

# 2 État de l'art

## 2.1 Définition

Une inférence est 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 deuxième assertion également vraie » (Larousse). 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 prenant appui sur un contenu vérifié ou supposé vérifié.

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). 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 de la vie quotidienne, Beaupré (2009) estime qu'il n'existe aucune loi absolue, mais que les inférences reposent sur un processus de généralisation et de règles.

## 2.2 Types d'inférence

Bien qu'il existe un nombre important d'études sur les inférences, il n'existe pas de consensus sur une classification uniforme des différents types d'inférences (Lavigne, 2008), dans la mesure où tout travail de classification dépend à la fois du domaine scientifique et des objectifs visés.

Au niveau du document, Graesser *et al.* (1994) ont proposé trois types d'inférence pour expliquer le processus de compréhension : locales (au niveau de la phrase ou du paragraphe), globales (à l'échelle du document) et explicatives (proposant une reformulation).

Au niveau linguistique, Dufaye (2001) se fonde sur la théorie du sens élaborée par Peirce (1958) pour proposer une distinction des inférences fondée sur le mécanisme mis en œuvre mentalement : déduction, induction, et rétroduction. Ces trois types sont considérées par Peirce comme les trois figures du syllogisme (Deledalle, 1994). Les exemples de ces trois types sont présentés dans le tableau 4. Pour aboutir 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 comme $(i)$ l'inférence immédiate qui ne possède qu'une seule prémisse et que la

conclusion est aboutie via cette prémisse ou $(ii)$ l'inférence médiale qui possède au moins 2 prémisses (Khemlani *et al.*, 2012). Les inférences rétroductives imposent de prendre en compte des connaissances antérieures (Deledalle, 1994).

| Type d'inférence | Exemple | Traduction |
|---|---|---|
| Déduction | 再来巴黎还会选择这里 | On va encore choisir cet hôtel la prochaine fois à Paris. |
| Induction | 唯一我不满的就是房间缺少一台咖啡机 | La seule chose qui n'était pas satisfaite est qu'il manque une machine à café dans la chambre. |
| Rétroduction | 洗澡地方对于小身材的亚洲人都有点拥挤，不知道歪果仁是怎么洗澡的在这么一个狭窄的地方。 | La salle de bain est étroite, même pour des asiatiques de petite taille. Imaginez comment des étrangers peuvent prendre la douche dans un entroit aussi petit. |

Tableau 1 – Exemples de déduction, induction et rétroduction

Duchêne (2008) distingue les inférences logiques des inférences pragmatiques. Les inférences logiques reposent sur un raisonnement formel et mettent en œuvre un processus logique alors que les inférences pragmatiques reposent sur un raisonnement inductif et s'appuient sur l'ensemble des connaissances acquises par un individu lors de ses expériences passées. Par exemple, il faut savoir que le Champ-de-Mars est à côté de la Tour Eiffel pour aboutir une conclusion positive pour le commentaire « *L'hôtel est à 5 min à pieds du Champ-de-Mars* ».

Doucy & Massous (2012) opèrent une distinction fondée sur le niveau d'analyse. 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 (l'enchaînement cohérent 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 l'inscrire dans une situation énonciative. Doucy & Massous (2012) soulignent également le fait que les connaissances extérieures au texte permettent de moduler le sens issu des inférences lexicales en apportant de nouvelles significations. Nous observons que cette opposition, fondée sur le niveau d'analyse, offre un cadre applicatif pertinent pour le traitement automatique des langues.

# 3 Matériel et méthodes

## 3.1 Description du corpus

Dans ce travail, nous nous intéressons aux inférences utilisées en chinois, dans un objectif de fouille d'opinion. A cet effet, nous avons rassemblé un corpus de commentaires postés sur trois 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], Mafengwo [2] et TripAdvisor [3]. Les trois sites fournissent tous une plateforme

1. http://www.booking.com/
2. http://www.mafengwo.cn/
3. http://www.tripadvisor.fr/

pour que les utilisateurs puissent partager leurs propres expériences sur des séjours aux hôtels de Paris. Booking et TripAdvisor sont utilisés par des utilisateurs internationaux, en différentes langues, alors que Mafengwo est un site chinois, utilisé par des internautes sinophones (Chine continentale, Hong Kong, Taïwan). Bien que tous les utilisateurs de ces sites soient des voyageurs qui les utilisent pour planifier leurs voyages, nous observons une différence de classes d'âges sur le site Mafengwo, davantage fréquenté par de jeunes utilisateurs qui rédigent des blogs de voyage et postent des annonces pour des voyages collectifs.

Les commentaires sont rédigés en chinois simplifié et traditionnel. A chaque hôtel est associé des métadonnées comme son nom, son URL, une note globale et des notes pour chacune des propriétés (localisation, service, équipement, WIFI, propreté, confort, etc). Le corpus contient 1 776 hôtels sur Booking avec 27 043 commentaires dont la longueur moyenne est 36 caractères ; 2 017 hôtels sur Mafengwo avec 24 025 messages dont la longueur moyenne est 83 caractères par message. Les commentaires en chinois de Booking et Mafengwo sont tous écrits par des natifs, alors que les commentaires sur TripAdvisor sont mélangés car le site permet aux volontaires de traduire des messages en chinois. Il est donc possible d'observer sur le site TripAdvisor un phénomène culturel différent du public chinois dans les traductions d'expériences.

## 3.2 Analyse des inférences

A partir de notre corpus, nous avons manuellement étudié les différents types d'inférence disponibles. De cette analyse, nous avons établi une nouvelle classification des inférences que nous estimons pertinente pour notre tâche de fouille d'opinion en chinois. Nous avons abandonné la distinction ici non pertinente entre inférences médiates et immédiates car seules les inférences immédiates sont présentes dans notre corpus. Notre étude met en évidence trois niveaux d'analyse des inférences.

- réalisation sémantique : désigne comment se fait l'accès au sens exprimé dans l'inférence (inférence logique, pragmatique, ou lexicale)
- modalité de réalisation : désigne le processus mental que le locuteur met en œuvre pour accéder au sens (déduction, induction ou rétroduction)
- mode de production : renvoie à la manière dont l'émetteur du message a produit l'inférence (inférence énonciative ou discursive)

## 3.3 Variantes linguistiques en chinois

En analysant le corpus, nous avons observé que des variantes chinoises jouent un rôle important dans l'analyse des inférences. Nous observons deux types de variantes : $(i)$ la polysémie, et $(ii)$ la conversion du chinois simplifié vers le traditionnel.

Wu & Hsieh (2010) et Sheng (2011) considèrent les variantes en chinois comme des caractères ayant des formes visuelles différentes, mais qui ont la même prononciation ou la même signification. En traitement automatique des langues, ces cas ne sont pas évidents car ils partagent le même code Unicode. Comme les termes chinois utilisent plus d'un caractère, il est difficile de distinguer ou d'extraire les informations. Par exemple, 行 (U+884C) désigne le terme « marcher » 行走(U+884C, U+8D70) mais aussi « banque » 银行(U+94F, U+884C) (Lu *et al.*, 2016). De même, le caractère 黄 renvoie à quatre sens : 黄色 « jaune », 姓黄 renvoie au nom de famille Huang, 这事儿黄了 signifie que « la chose a échoué », et 扫黄 « anti-pornographie ». Selon le CDNC (Chinese Domain Name

Consortium), environ 40 % des caractères ont des formes variantes, ce qui souligne l'importance de prendre en compte cet aspect dans notre analyse.

Le chinois simplifié et le chinois traditionnel sont toujours utilisés de nos jours. Le chinois simplifié est utilisé en Chine (RPC) alors que le chinois traditionnel est utilisé à Taïwan, Hongkong et Singapour. Comme les commentaires des forums mélangent le chinois simplifié et le chinois traditionnel, nous considérons également ce type de variantes, d'autant plus que différentes expressions peuvent être influencées selon les régions. Par exemple, Halpern (2006) relève que « taxi » est noté 出租汽车 en chinois simplifié, 計程車 en chinois traditionnel de Taïwan, et 的士 en chinois simplifié de Hong-Kong.

En analysant le corpus, nous avons fréquemment observé des variantes linguistiques sur des thématiques, des opinions et des objets. Il s'agit parfois de synonymes, mais aussi de variantes traditionnelles. Les variantes sont complexes à traiter car les commentaires postés sur Internet ne respectent pas strictement les traductions normalisées.

# 4 Résultats et discussion

## 4.1 Classification et combinaison des inférences

Nous renseignons dans le tableau 4 le nombre d'inférences pour chaque catégorie et donnons dans le tableau 3 des exemples pour chacun des types d'inférence que nous avons identifiés en corpus.

| Niveau d'analyse | Type | Nombre |
|---|---|---|
| Réalisation sémantique | logique | 36 (19,9 %) |
| | pragmatique | 91 (50,3 %) |
| | lexical | 54 (29,8 %) |
| Modalité de réalisation | induction | 17 (17,5 %) |
| | déduction | 75 (77,3 %) |
| | rétroduction | 5 (5,2 %) |
| Mode de production | discursif | 58 (52,3 %) |
| | énonciatif | 53 (47,7 %) |

Tableau 2 – Nombre d'inférences pour chacun des types définis

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

Premièrement, l'objet d'une opinion n'est pas toujours explicite dans le messages d'un utilisateur. Par exemple, « proche de la Tour Eiffel » implique de manière sous-entendue une localisation positive de l'hébergement, dans un contexte touristique. La proximité d'une station de métro est déjà plus complexe à interpréter. Cette localisation est-elle positive du point de vue de l'accès aux transports, ou négative en raison des nuisances engendrées ?

Deuxièmement, il n'est pas toujours possible de dégager des indices forts pour repérer facilement les phrases qui contiennent des inférences. Les mots porteurs de sentiments ou les formes morpho-syntaxiques ne permettent pas une identification de manière certaine. Le lecteur doit alors mobiliser

| Type | Exemple | Traduction | Polarité |
|---|---|---|---|
| logique, déduction, discursif | 我见过最小的卫生间，跟飞机上的差不多 | La plus petite salle de bain que j'ai rencontrée, presque comme en avion | négatif |
| logique, déduction, énonciatif | 酒店装修严重影响客户 | Les travaux de l'hôtel gênent beaucoup les clients | négatif |
| logique, induction, discursif | 退房时让酒店帮忙叫了车去机场，但我觉得价格贵了，可能被宰了 | On a commandé un taxi à l'hôtel au moment du check-out pour aller à l'aéroport, mais le prix était cher, c'était peut-être une anarque | négatif |
| logique, induction, énonciatif | 前台只有一个人，非常忙，每次都要排队等。 | Il n'y a qu'une personne à l'accueil qui est très occupée, il faut faire la queue chaque fois | négatif |
| logique, rétroduction, discursif | 巴黎人干什么都漫不经心，应该放在房间的手机到退房都没给，每天都说第二天。 | Les parisiens font n'importe quoi. On n'a pas eu accès au téléphone dans la chambre jusqu'au moment du départ. Tous les jours on nous a dit qu'on nous le donnerait le lendemain | négatif |
| pragmatique, déduction, discursif | 距离凯旋门约500米 | L'Arc de Triomphe est à environ 500m | positif |
| pragmatique, déduction, énonciatif | 离地铁站很近 | proche du métro | positif |
| pragmatique, déduction, discursif | 但铁塔景观并不理想，只能看到一些铁塔尖 | Cependant, la vue de la Tour Eiffel n'est pas idéale, on voit seulement le bout de la pointe | négatif |
| pragmatique, induction, énonciatif | 唯一我不满的就是房间缺少一台咖啡机。 | La seule chose qui n'est pas satisfaisante est qu'il manque une machine à café dans la chambre. | négatif |
| pragmatique, induction, discursif | 旁边有家乐福 | Carrefour City à côté | positif |
| pragmatique, rétroduction, discursif | 洗澡地方对于小身材的亚洲人都有点拥挤，不知道歪果仁是怎么洗澡的在这么一个狭窄的地方。 | La salle de bain est étroite, même pour des asiatiques de petite taille. Imaginez comment des étrangers peuvent prendre la douche dans un entroit aussi petit | négatif |
| lexical | 埃菲尔铁塔 | Tour Eiffel | positif |

Tableau 3 – Exemples des combinaisons d'inférences

des connaissances personnelles sur le monde et des compétences d'ordre linguistique pour décoder ces inférences. Ce travail se révèle encore plus complexe pour une machine, même en mobilisant des moyens du traitement automatique des langues.

Troisièmement, pour le domaine spécifique de l'hôtellerie, des inférences sont aussi représentées par le lexique du tourisme et des noms propres. Le traitement du lexique spécifique fait partie de l'analyse des inférences.

Dans la première série (logique-pragmatique-lexical), 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 déduction, induction ou 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 effectuer une fouille d'opinion en chinois. Par ailleurs, 29,8 % des inférences sont lexicales. Il est ainsi 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, alors que la combinaison de trois sous-types concerne jusqu'à 50 % des cas.

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

## 4.2 Plongements lexicaux

Comme il n'existe pas d'indice fort pour identifier automatiquement les inférences, nous avons essayé de les traiter au moyen d'un apprentissage automatique. Dans cet article, nous avons utilisé un modèle Word2Vec qui permet de classer les similarités d'un mot candidat dans un espace vectoriel même sans des étiquettes grammaticales.

### 4.2.1 Protocole expérimental

A partir d'un corpus de 1 238 989 tokens, segmenté par l'outil jieba [4] (sans charger des dictionnaires extérieurs), nous avons extrait les 2000 premiers termes les plus fréquents. A l'aide du module gensim [5] de Python, nous avons ensuite entraîné un modèle Word2Vec. En ce qui concerne les paramètres, nous avons défini que le seuil de fréquence, la longueur de la fenêtre et la taille de dimension sont respectivement 5, 5 et 400. Enfin, nous avons retenu les 50 mots cibles les plus similaires à chaque candidat. Une partie des résultats est présentée dans le tableau 4.

### 4.2.2 Analyse des résultats

Nous effectuons les constatations suivantes :

— Les mots cibles qui constituent une inférence avec le mot candidat ne sont pas toujours les plus proches dans un espace vectoriel. Cette observation explique pour quelle raison il n'est pas évident d'identifier les inférences en corpus. Par exemple, les mots les plus proches de 地理位置 (localisation) sont 地段 (secteur, 0,933), 环境 (environnement, 0,736), 治安 (sécurité publique, 0,710), et 景色 (vue, 0,603). Ils correspondent aux éléments fondamentaux d'une localisation. Mais les mots qui permettent d'identifier une inférence apparaissent également dans cette liste : Montparnasse (0,574), 凯旋门 (Arc de Triomphe, 0,569), 地铁口 (entrée du métro, 0,563) et 卢浮宫 (Louvre, 0,553). Le lien entre localisation et ces mots de cible constitue une inférence pragmatique. Par exemple, la phrase 酒店地理位置在凯旋门旁边 (« *La localisation de l'hôtel est à côté de l'Arc de Triomphe* ») combine des inférences

4. https://github.com/fxsjy/jieba
5. https://pypi.python.org/pypi/gensim

| N° | Token | Voisins distributionnels |
|---|---|---|
| 1 | 前台 (accueil) | 店员(personnel) 0,725 [...] 英语(personnel) 0,607 [...] 笑容 (sourire) 0,547) [...] |
| 2 | 房间 (chambre) | 屋子 (chambre) 0,741 [...] 卫生间 (toilette) 0,660 [...] 面积 (surface) 0,589 [...] |
| 3 | 方便 (pratique) | 便利 (pratique) 0,880 [...] 好找 (facile à trouver) 0,562 [...] 公交 (bus) 0,432 [...] |
| 4 | 地铁站 (station du métro) | 地铁口 (entrée du métro) 0,928 [...] 红磨坊 (moulin rouge) 0,724 [...] 巴黎圣母院 (Notre Dame de Paris)0,642 [...] 家乐福(Carrefour) 0,567 [...] |
| 5 | 电梯 (ascenseur) | 楼梯 (escalier) 0,721 [...] 行李箱 (valise) 0,575 [...] 缺点 (défaut) 0,451 [...] |
| 6 | 免费 (gratuit) | 下午茶 (goûter) 0,649 [...] 大厅 (hall) 0,621 [...] 打印机 (imprimante) 0,562 [...] 电热水壶 (bouilloire) 0,547 [...] |
| 7 | 工作人员 (personnel) | 员工 (employé) 0,928 服务员 (serveur) 0,910 店员 (vendeur) 0,861 服务生 (serveur) 0,801 人员 (personnel) 0,735 [...] |

Tableau 4 – Exemples des voisins distributionnels pour quelques candidats

pragmatique (besoin de connaissances extérieures au texte), énonciative (dans une situation énonciative) et déductive (avec une prémisse solide afin de définir une polarité positive).
Ce genre de lien est fréquent entre le mot candidat et ses voisins distributionnels. Le mot candidat 前台 (« accueil », exemple 1) dans le tableau 4, constitue une inférence à la fois pragmatique et énonciative avec son voisin 英语 (« anglais »[6]). Il en est de même pour 免费 (« gratuit ») et son voisin 水壶 (« bouilloire », exemple 6) qui représente un stéréotype des Chinois qui ont l'habitude de boire de l'eau chaude.
Ces voisins distributionnels n'apparaissent cependant pas parmi les premiers (par score décroissant), mais à partir de la dixième place dans la liste de voisins. Nous considérons que cela fournit néanmoins un indice pour l'identification des inférences dans l'étape suivante de notre recherche.

— La méthode des plongements lexicaux donne un moyen d'établir une base de lexique spécifique et de regrouper des variantes dans l'hôtellerie en chinois. Par exemple, la liste des similarités du candidat 埃菲尔铁塔 (« Tour Eiffel ») contient quasiment tous les sites à Paris, alors que la liste du candidat 工作人员 (« personnel ») regroupe toutes les variantes des métiers de l'hôtellerie, ce qui est listé dans la 7e ligne du tableau 4. Cela correspond aussi aux besoins de traiter des variantes linguistiques pour l'analyse des inférences.

— La plupart des résultats montre une relation forte entre le mot candidat et les voisins associés, qui constituent des inférences pragmatiques ou lexicales. Ce phénomène explique également la grande proportion de l'inférence pragmatique et lexicale observée dans la partie précédente.

6. Le fait que l'anglais soit parlé à l'accueil de l'hôtel constitue un élément positif pour les touristes chinois en visite à Paris. Il est donc normal de trouver ce qualificatif comme voisin distributionnel du terme « accueil ».

# 5 Conclusion

En étudiant des recherches existantes concernant différents types d'inférence, nous avons constaté que l'inférence est une question peu traitée dans la fouille d'opinion.Dans cet article, nous avons réalisé une analyse des inférences dans un corpus de commentaires touristiques rédigés en chinois sur l'hébergement à Paris. De cette analyse, nous avons mis en évidence la complexité et la nécessité du traitement des inférences pour la fouille d'opinion en chinois. Ces tâches sont compliquées car il n'existe pas de méthode optimale pour identifier facilement les inférences pragmatiques. D'autre part, nous avons observé que la répartition est irrégulière ; une inférence relève de un à trois sous-types d'une part, et la répartition entre sous-types n'est pas homogène. Cependant, nous considérons que la prise en compte des inférences offre une piste pertinente pour réaliser la fouille d'opinion en chinois dans un domaine spécifique. Nos travaux futurs vont consister à intégrer l'analyse des inférences comme caractéristiques dans les algorithmes d'apprentissages statistiques tels que les plongements lexicaux (Word2Vec), de manière à réaliser une fouille d'opinion. Aussi, nous allons ajouter un prétraitement de normalisation pour éviter les erreurs de translittération.

# Remerciements

J'adresse mes remerciements à M. Valette et M. Grouin, mes directeurs de thèse, pour leur précieuse aide à la relecture et à la correction de l'article.

# Références

BEAUPRÉ S. (2009). L'approche dialectique pragmatique dans l'analyse des arguments. Master's thesis, UQAM, Montréal, Canada.

DELEDALLE G. (1994). Charles S. Peirce. les ruptures épistémiologiques et les nouveaux paradigmes. *Travaux du Centre de Recherches Sémiologiques*, **62**.

DOUCY G. & MASSOUS T. (2012). Sémantique inférentielle et compréhension des verbatim clients. In *Congrès Mondial de Linguistique Française*, volume 1.

DUCHÊNE A. (2008). Les inférences dans la communication : cadre théorique général. In *Actes de Rééducation orthophonique*, number 234. Fédération Nationale des Orthophonistes.

DUFAYE L. (2001). Les modaux et la négation en anglais contemporain. In *Cahiers de Recherche*. Ophrys.

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

GOMBERT J.-E., FAYOL M., ZAGAR D., LECOCQ P. & SPRENGER-CHAROLLES L. (1992). *Psychologie cognitive de la lecture*.

GRAESSER A. C., SINGER M. & TRABASSO T. (1994). Constructing inferences during narrative text compréhension. *Psychological Review*, **101**(3).

HALPERN J. (2006). The role of lexical resources in CJK natural language processing. In *Proc of Multilingual Language Resources and Interoperability Work*, p. 9–16, Sydney, Australia.

KHEMLANI S., TRAFTON J. G., LOTSTEIN M. & JOHNSON-LAIRD P. N. (2012). A process model of immediate inferences. p. 151–156.

KISPAL A. (2008). *Effective Teaching of Inference Skills for Reading*. Rapport interne, Research Report DCSF-RR031.

LAVIGNE J. (2008). *Les mécanismes d'inférence en lecture chez les élèves de sixième année du primaire*. PhD thesis, Université Laval, Québec, Canada.

LU Y., ZHANG Y. & JI D. (2016). Multi-prototype chinese character embedding. In *Proc of LREC*, Portorož, Slovenia.

PEIRCE C. S. (1958). The collected papers of charles sanders peirce. In *Cambridge : Harvard University Press*, volume 1-6.

SHENG S. (2011). *Report on Chinese Variants in Internationalized Top-Level Domains*. Rapport interne, ICANN, Marina Del Rey, CA.

WU Y.-C. & HSIEH S.-K. (2010). PyCWN : a python module for Chinese Wordnet. In *Proc of COLING, Demo*, p. 5–8, Beijing, China.