TALN 2010, Montréal, 19–23 juillet 2010

Analyse d'opinion : annotation sémantique de textes chinois

Lei Zhang, Stéphane Ferrari
GREYC - Département Informatique
Université de Caen - Campus 2, 14032 Caen
{prenom.nom}@unicaen.fr

Résumé. Notre travail concerne l'analyse automatique des énoncés d'opinion en chinois. En nous inspirant de la théorie linguistique de l'*Appraisal*, nous proposons une méthode fondée sur l'usage de lexiques et de règles locales pour déterminer les caractéristiques telles que la Force (intensité), le Focus (prototypicalité) et la polarité de tels énoncés. Nous présentons le modèle et sa mise en œuvre sur un corpus journalistique. Si pour la détection d'énoncés d'opinion, la précision est bonne (94 %), le taux de rappel (67 %) pose cependant des questions sur l'enrichissement des ressources actuelles.

Abstract. Our work concerns automatic analysis of opinion in texts. Based on the Appraisal linguistic theory, our method uses lexical and syntactic resources to process such properties as the Force, the Focus and the polarity of an opinion. We present our model and its implementation on a journalistic corpus. The precision for detecting opinion expressions is high (94%), but the recall (67%) raises the question of how to enhance the resources.

Mots-clés : Analyse d'opinion, théorie de l'*Appraisal*.

Keywords: Opinion analysis, Appraisal theory.

1 Introduction

L'essor d'Internet a vu croître le nombre de documents mis à disposition du grand public. Depuis maintenant une dizaine d'années s'est développée en parallèle de ce phénomène la discipline communément appelée « fouille d'opinion », combinant des travaux en traitement automatique des langues et en fouille de données. Notre travail s'inscrit dans cette mouvance. Il vise à caractériser les énoncés d'opinion au sein de textes journalistiques en chinois, à en repérer les sources et les cibles ainsi que des propriétés telles que leur polarité, leur intensité, leur prototypicalité.[1]

Après un bref état de l'art, nous présentons notre approche qui s'articule sur une analyse syntaxique du chinois et une adaptation de la théorie linguistique de l'*Appraisal* du langage de l'évaluation en anglais (section 2). Nous décrivons notre méthode puis sa mise en œuvre informatique, en détaillant le corpus et les ressources exploitées. Nous présentons les premiers résultats obtenus, validant notre approche, et nous concluons en proposant quelques perspectives de poursuite de nos travaux visant notamment à traiter les cas encore problématiques et questionnant le problème de l'enrichissement des ressources.

[1]Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009 – projet OntOpiTex.

2 État de l'art

Fouille d'opinion en TAL Les travaux en TAL concernant la fouille d'opinion peuvent être classés sommairement en trois catégories principales : (i) constitution de ressources, (ii) classification, de textes ou de phrases, (iii) analyse des énoncés d'opinion au sein des textes. Nous renvoyons à la monographie (Pang & Lee, 2008) pour un panorama de ces approches, et au numéro « Fouille de Données d'Opinions » de RNTI[2] pour des travaux récents en français.
Dans (Hatzivassiloglou & McKeown, 1997), un des premiers travaux concernant la constitution de ressources, les auteurs s'attachent à déterminer la polarité d'adjectifs (positive ou négative). D'autres travaux cherchent à déterminer le caractère subjectif (versus objectif) d'entrées lexicales, comme la méthode proposée par (Baroni & Vegnaduzzo, 2004). Plus récemment, (Esuli & Sebastiani, 2006) proposent une adaptation de la ressource « SENTIWORDNET », combinant ces deux propriétés (subjectivité et polarité, positive ou négative) pour les associer aux *Synsets* de *WordNet*. De manière connexe, notons les travaux de (Whitelaw *et al.*, 2005) proposant une méthode d'apprentissage pour caractériser des expressions complexes selon des propriétés de l'*Appraisal*. La classification de textes et de phrases opère en général de la même manière, c'est-à-dire en terme de subjectivité ou de polarité, en exploitant souvent les ressources précédentes, et en faisant appel à des techniques de la fouille de données et de l'apprentissage. Les domaines applicatifs sont variés : analyses de critiques de film dans (Turney, 2002) ; pour le français, la campagne DEFT07 (Grouin *et al.*, 2009) concernait quatre domaines applicatifs distincts, critiques de produits culturels, critiques de jeux vidéo, relectures d'articles scientifiques et textes politiques. L'analyse au sein des textes consiste à préciser d'autres caractéristiques des énoncés d'opinion. Citons par exemple (Hu & Liu, 2004) qui étudient les critiques émises par des consommateurs sur les caractéristiques de divers produits. En français, (Ferrari *et al.*, 2009) proposent une approche fondée sur une étude linguistique des énoncés évaluatifs.

Notre étude se situe dans ce dernier axe, s'appuyant sur la théorie de l'*Appraisal*. À notre connaissance, c'est la première tentative d'adaptation de cette théorie pour le chinois. Les travaux en chinois concernent surtout la classification de texte, rarement l'analyse d'opinion au sein des textes. Notons toutefois le travail de (Wu & Oard, 2007), détectant la polarité de phrases dans des articles journalistiques.

Théorie de l'*Appraisal* La théorie linguistique de l'*Appraisal* (Martin & White, 2005), élaborée pour la langue anglaise, propose un système décrivant les propriétés des énoncés d'opinion selon différents aspects : l'attitude, l'engagement et la graduation. La figure 1 en présente les grandes lignes. L'attitude contient trois grandes catégories : affect, appréciation (plutôt esthétique) et jugement (plutôt étique), ainsi que la notion de polarité qui se combine avec les précédentes. L'engagement concerne les notions de prise en charge énonciative, dialogisme, polyphonie, *etc*. La graduation concerne la force, liée à l'intensité de l'opinion émise, et le focus, lié en quelque sorte à la prototypicalité des termes employés. La graduation peut agir tant sur l'attitude (un jugement très négatif) que sur l'engagement (une distanciation accentuée).

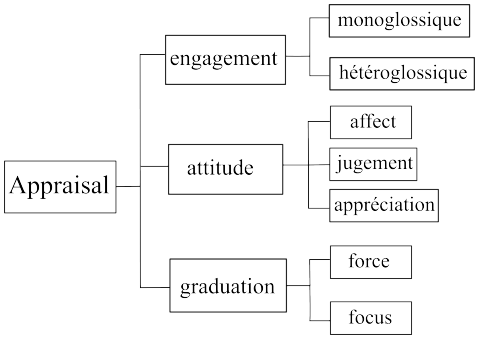


FIG. 1 – Système de l'*Appraisal*.

[2]Revue des Nouvelles Technologies de l'Information, RNTI-E-17, Cépaduès Éditions, 2009

3 Modèle d'analyse des énoncés d'opinion

Notre objectif est d'analyser les énoncés d'opinion afin d'en déterminer les caractéristiques suivantes : polarité, force et focus, et, à terme, la cible et le type d'attitude correspondant. Nous n'étudions pas à ce stade la notion d'engagement. Notre approche articule ressources lexicales et grammaires d'analyse locale. Nous en présentons dans cette section les grandes lignes, puis décrivons sa mise en œuvre et les résultats obtenus dans la section suivante.
Notre modèle s'appuie en premier lieu sur une analyse syntaxique locale, nécessaire pour établir comment les différents éléments du texte interviennent dans le calcul des valeurs des propriétés polarité, force et focus. La suite de l'analyse consiste à produire une fiche récapitulant les propriétés que nous pouvons capter pour chaque opinion exprimée, avec possiblement plusieurs fiches pour une même phrase, concernant éventuellement la même cible (*C'est un personnage plutôt antipathique, mais vraiment très intelligent.* donnera ainsi lieu à deux fiches, l'une pour *antipathique*, l'autre pour *intelligent*), avec ici les propriétés : polarité négative et force réduite (*plutôt* + *mais*) pour *antipathique* ; polarité positive, force accrue (*très*) et focus avivé (*vraiment*) pour *intelligent*.

Opinion, lexiques et syntaxe L'ensemble de notre approche exploite des ressources lexicales pour indiquer, *a minima*, le caractère potentiellement axiologique des mots du texte. Dès la première étape de notre modèle, pour alléger l'analyse syntaxique, seules les phrases d'opinion sont étudiées, c'est-à-dire celles contenant au moins une entrée lexicale présente dans nos lexiques.
L'analyse syntaxique que nous proposons pour le chinois se fonde sur des règles de composition de groupes de mots et de propositions. Nous renvoyons à (Zhang, 2010) pour plus de détails sur cette étape. La proposition y est choisie comme unité d'analyse, car nous pensons qu'elle constitue l'unité maximale du système de la grammaire du chinois. Nous décrivons plus particulièrement des règles de groupes de mots par rapport aux cinq catégories principalement impliquées dans l'expression d'opinion (adjectifs, noms, verbes, adverbes et expressions proverbiales lexicalisées. La grammaire qui en résulte se compose actuellement d'une soixantaine de règles simples. Nous présentons dans la section suivante des exemples de résultats (figure 2) de l'analyseur obtenu.

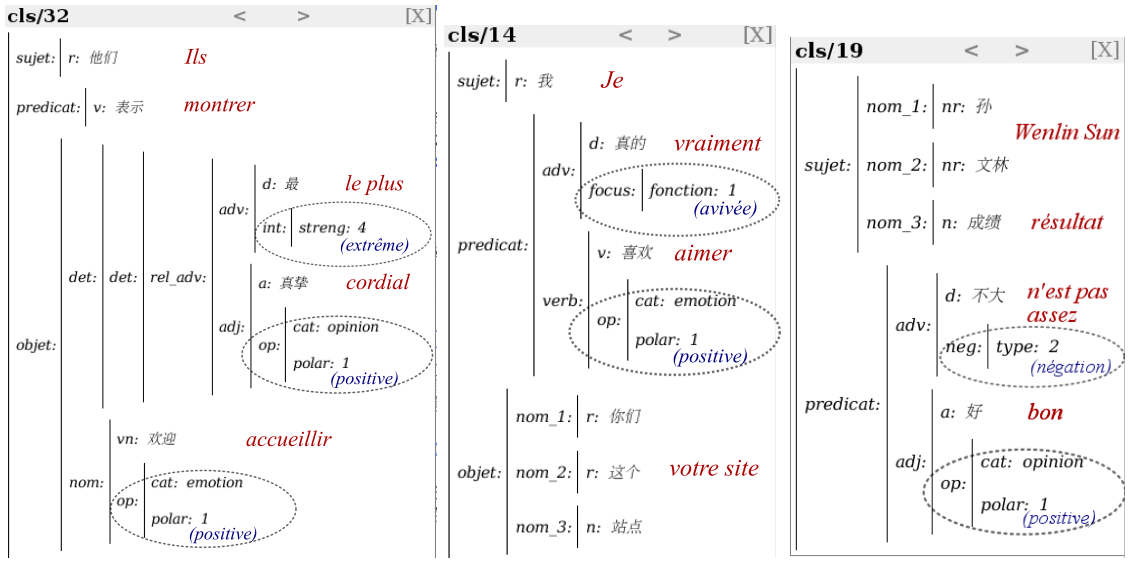
Par nature, certaines entrées lexicales sont directement liées à un type d'attitude et une polarité (*aimer* - affect positif, *beau* - appréciation positive, *opiniâtre* - jugement négatif). Les ressources lexicales exploitées codent aussi ce type d'information pour les entrées potentiellement subjectives, avec des ambiguïtés possibles pour certaines entrées, polysémiques ou sous-spécifiées. Enfin, les analyses de la polarité, de la force et du focus font intervenir d'autres indices codés eux aussi dans des ressources adaptées : modificateurs d'intensité (*très*, *extrêmement*), marqueurs de focus (*vrai*, *véritablement*, *sorte de*), négation, *etc.*

Fiches d'opinion La seconde étape du traitement consiste à établir pour chaque énoncé d'opinion une fiche synthétisant ses propriétés. Chaque proposition contenant au moins un mot d'opinion donne lieu à la construction d'autant de fiches que nécessaire. Chaque fiche est construite à partir de l'analyse syntaxique de la proposition dans laquelle apparaît le mot d'opinion qui a déclenché sa création.
Une fiche synthétise les informations suivantes : la catégorie de l'attitude (opinion et émotion), la polarité (positive ou négative), l'intensité (de 1 à 4) et le focus (avivé ou atténué), ainsi que la cible pour certaines configurations. L'ensemble de ces informations sont directement extraites de la structure syntaxique, à l'aide de règles locales pour calculer la polarité, la force et le focus résultant de la combinaison éventuelle de plusieurs indices. Ainsi, dans la phrase « 我 十分 高兴 » (trad. *Je [suis] extrêmement content*, citation

extraite de notre corpus d'étude, Le Quotidien du Peuple, 1998), l'analyse syntaxique en SujetNP PredicatAP (Adv Adj) permet de déduire directement les propriétés cible : *je* (structure de la proposition), intensité : *4* (maximum, adverbe et structure de la proposition), polarité : + et catégorie : émotion (Affect) directement liées à l'adjectif d'opinion *content*. Le focus reste neutre, inchangé par le contexte.
La négation est actuellement traitée de telle façon qu'elle peut modifier la polarité ou la force :

« Ce gâteau n'est pas **bon** » ⇒ polarité inversée
« Ce gâteau n'est pas *très* **bon** » ⇒ polarité inversée et force réduite

4 Mise en œuvre et résultats



(a) (b) (c)

FIG. 2 – Propositions analysées.

```
idfiche   : 32
texte     : 他们 表示 最 真挚 的 欢迎
catégorie : émotion
polarité  : positive
force     : 4 (extrême)
cible     : 最 真挚 的 欢迎
```

(a)

```
idfiche   : 14
texte     : 我 真的 喜欢 你们 这个 站点
catégorie : émotion
polarité  : positive
focus     : avivé
cible     : 你们 这个 站点
```

(b)

```
idfiche   : 19
texte     : 孙 文林 成绩 不大 好
catégorie : opinion
polarité  : négative
force     : 1 (réduite)
cible     : 孙 文林 成绩 ···
```

(c)

FIG. 3 – Fiches produites correspondant aux propositions analysées.

Corpus et lexiques Notre corpus d'étude est constitué d'un mois d'articles du journal chinois « Le quotidien du peuple » (janvier 1998). Il a été annoté manuellement par l'institut informatique et linguistique de Pékin.[3] Les mots sont séparés et annotés par leur classe grammaticale selon la fonction syntaxique qu'ils réalisent en contexte (un même mot chinois peut être verbe ou nom ou adjectif, c'est un emploi particulier qui fixe la catégorie). Nous avons segmenté le corpus en articles et adapté l'annotation au format XML[4]

[3] http://icl.pku.edu.cn/
[4] 3 293 fichiers XML (13,8Mo)

pour pouvoir l'exploiter avec *LinguaStream*,[5] une plateforme de TAL développée par notre équipe.
Nous avons adapté les lexiques chinois proposés par *HowNet*[6] pour l'analyse du sentiment (Zhang, 2010). Il y a six lexiques initiaux contenant 9 193 mots chinois : opinion positive, opinion négative, émotion positive, émotion négative, intensité et assertion. Ils peuvent être mis en correspondance avec une partie de la théorie d'*Appraisal* : jugement et appréciation réunis (opinion), affect (émotion) et intensité. Nous y avons adjoint un lexique pour la négation et un lexique pour le focus, créés par nos soins.
Les mots d'opinions s'inscrivent d'une façon générale dans les 5 classes grammaticales suivantes : adjectifs, noms, verbes, adverbes et expressions proverbiales. Un traitement particulier a été nécessaire pour gérer le problème de certains mots polycatégoriels qui portent une opinion selon leur classe. Par exemple, le même mot chinois « 安全 » est à la fois un nom (trad. « sécurité ») qui ne porte pas d'opinion et un adjectif (trad. « en sécurité ») qui porte une opinion positive.

Repérage des énoncés évaluatifs L'analyseur syntaxique n'est déclenché que pour les phrases contenant au moins un mot des lexiques *Opinion* ou *Émotion*. La proposition analysée est représentée sous la forme d'un schéma, qui contient les information suivantes : la structure de la proposition (sujet, prédicat) ; les relations de dépendance des groupes de mots ; les annotations sémantiques (polarité, intensité, *etc.*) issues des lexiques. La figure 2 montre l'analyse de trois propositions. Elles se traduisent : **(a)** Ils sont les plus accueillants ; **(b)** J'aime vraiment votre site ; **(c)** Le résultat de SunWenLin n'est pas assez bon. La figure 3 montre les trois fiches produites à l'étape suivante.

Évaluation Afin d'obtenir une première évaluation de nos outils, nous avons extrait manuellement les énoncés d'opinion dans 5 articles longs de notre corpus et les avons comparées avec les sorties de notre analyseur. 138 énoncés ont été repérés par nos soins, répartis dans 111 phrases. Notre analyseur a construit 98 fiches automatiquement, dont 92 correspondent à des énoncés repérés manuellement. Ceci correspond à un rappel de 66,7 % (92/138) et une précision de 93,9 % (92/98), soit une F-mesure harmonique de 78 %. Il y a deux causes principales de silence : (i) le mot d'opinion n'est pas dans notre lexique ; (ii) silence ou erreur de l'analyseur syntaxique – mot d'opinion marqué mais à l'extérieur des propositions traitées, propositions mal découpées (point-virgule, guillemets...). Pour synthétiser, la couverture du lexique et des règles reste limitée, la difficulté actuelle étant de savoir quelle quantité de mots et de règles ajouter pour quel gain.

Pour compléter cette étude de la qualité, nous avons évalué manuellement 282 fiches produites par notre système pour en vérifier la polarité (échantillonnage sur plusieurs articles au hasard). Nous obtenons sur ce point une précision 96,8 %. Les erreurs sont causées surtout par deux problèmes : pas de contexte pour juger la polarité d'une opinion (p.ex. la polarité du mot chinois « 骄傲 » (trad. « fier ») dépend de son utilisation : positive dans « être fier de quelqu'un » mais négative dans « faire le fier ») et lexique de négation encore incomplet (nous ne traitons que les adverbes de négation pour l'instant, mais il peut également y avoir d'autres tournures négatives : *manquer de*, *l'opposé de*).

5 Conclusion et perspective

Nous avons présenté une approche symbolique pour l'analyse de l'opinion dans des textes chinois. Fondée sur l'usage de lexiques et de règles d'analyse syntaxique locale, notre méthode vise à renseigner les

[5] `http://www.linguastream.org`
[6] HowNet Knowledge Database, `http://www.keenage.com`

propriétés d'énoncés d'opinion en s'inspirant de la théorie de l'*Appraisal*. Certaines configurations syntaxiques simples permettent déjà de mettre en œuvre des calculs pour déterminer la polarité, la force, le focus et la cible des opinions exprimées avec une bonne précision en ce qui concerne le caractère subjectif (94 %) et la polarité (93 %) mais un taux de rappel global de seulement 67 %.
Les erreurs actuelles et les cas non traités posent une question de fond. L'enrichissement tant du lexique que des règles peut difficilement se poursuivre manuellement, car un tel processus est coûteux et devient de moins en moins rentable. Il nous paraît plus utile de mettre en place des procédés de fouille automatique pour compléter nos ressources. C'est pourquoi nous envisageons désormais ce type d'approche, éventuellement de manière supervisée. Nous allons notamment étudier l'adaptation de ces techniques pour les propriétés qui nous intéressent (force, focus et aussi attitude), les travaux antérieurs s'étant focalisés sur celles de subjectivité et de polarité.

Références

BARONI M. & VEGNADUZZO S. (2004). Identifying subjective adjectives through web-based mutual information. In *KONVENS'04, 7th Konferenz zur Verarbeitung Natürlicher Sprache*, p. 613–619.

ESULI A. & SEBASTIANI F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC'06: 5th Conference on Language Resources and Evaluation*, p. 417–422.

FERRARI S., CHARNOIS T., MATHET Y., RIOULT F. & LEGALLOIS D. (2009). Analyse de discours évaluatif, modèle linguistique et applications. *RNTI*, **E-17**, 71–93.

GROUIN C., HURAULT-PLANTET M., PAROUBEK P. & BERTHELIN J. (2009). Deft'07 : une campagne d'évaluation en fouille d'opinion. *RNTI*, **E-17**, 1–24.

HATZIVASSILOGLOU V. & MCKEOWN K. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, p. 174–181, Morristown, NJ, USA: Association for Computational Linguistics.

HU M. & LIU B. (2004). Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 168–177.

MARTIN J. & WHITE P. (2005). *The Language of Evaluation: Appraisal in English*. Palgrave.

PANG B. & LEE L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, **2**(1-2), 1–135.

TURNEY P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the ACL (ACL'02)*, p. 417–424: Philadelphia, Pennsylvania, USA ACL.

WHITELAW C., GARG N. & ARGAMON S. (2005). Using appraisal groups for sentiment analysis. In *CIKM '05: 14th ACM international conference on Information and knowledge management*, p. 625–631, New York, NY, USA: ACM.

WU Y. & OARD D. (2007). Ntcir-6 at maryland: Chinese opinion analysis pilot taskdouglas w. oard. In *NTCIR-6 Workshop Meeting*.

ZHANG L. (2010). Analyse d'opinion : application à un corpus journalistique en chinois. In *CEDIL2010 : Colloque international des Etudiants chercheurs en Didactique des Langues et en Linguistique*: 29 juin au 2 juillet 2010, Grenoble, France. À paraître.