Extraction d'expressions-cibles de l'opinion : de l'anglais au français

Grégoire Jadi, Laura Monceaux, Vincent Claveau, Béatrice Daille

Extraction d'expressions-cibles de l'opinion : de l'anglais au français

Grégoire Jadi[1], Laura Monceaux[1], Vincent Claveau[2], Béatrice Daille[1]
(1) LINA - Univ. Nantes
gregoire.jadi beatrice.daille laura.monceaux@univ-nantes.fr
(2) IRISA–CNRS, France
vincent.claveau@irisa.fr

RÉSUMÉ

Dans cet article, nous présentons le développement d'un système d'extraction d'expressions-cibles pour l'anglais et sa transposition au français. En complément, nous avons réalisé une étude de l'efficacité des traits en anglais et en français qui tend à montrer qu'il est possible de réaliser un système d'extraction d'expressions-cibles indépendant du domaine. Pour finir, nous proposons une analyse comparative des erreurs commises par nos systèmes en anglais et français et envisageons différentes solutions à ces problèmes.

ABSTRACT

Opinion Target Expression extraction : from English to French

In this paper, we present the development of an Opinion Target Extraction system in english and transpose it to french. In addition, we realize an analysis of the features and their effectiveness in english and french which suggest that it is possible to build an Opinion Target Extraction system independant of the domain. Finally, we propose a comparative study of the errors of our systems in both english and french and propose several solutions to these problems.

MOTS-CLÉS : Fouille de l'Opinion, Extraction d'Expressions-Cibles, Étiquetage Multilingue.

KEYWORDS: Opinion Mining, Opinion Target Expression Extraction, Multingual Labelling.

1 Introduction

Avec l'expansion du web social, les internautes sont de plus en plus enclins à partager leurs avis sur les réseaux sociaux ou les sites spécialisés. Le domaine de la fouille d'opinion vise à désambiguïser automatiquement ces informations en ce qui concerne les émotions, les opinions, les sentiments et les attitudes exprimés, en les traduisant par une valence affective (polarité positive ou négative, score de subjectivité, etc.) ou par une classe. En plus de l'identification des opinions exprimées, les chercheurs en fouille de l'opinion cherchent également à identifier la ou les cibles de l'opinion.

Nos travaux ont pour objectifs de créer un système état de l'art d'extraction des expressions-cibles sur l'anglais puis de le transposer au français. La détection des expressions-cibles de l'opinion est un des objets de la sous-tâche 1 de la tâche 12 de l'édition 2015 de SEMEVAL (Pontiki *et al.*, 2015). L'objectif des systèmes est de repérer les expressions-cibles de l'opinion dans des commentaires portant sur des restaurants. Une expression-cible est une référence explicite à l'entité commentée

ou à un des aspects de l'entité commentée dans le texte. Cette référence peut être un mot simple ou une expression polylexicale. Par exemple dans la phrase *The lobster sandwich is good and the spaghetti with Scallops and Shrimp is great.*, les systèmes doivent extraire les expressions-cibles *lobster sandwich* et *spaghetti with Scallops and Shrimp*.

Jusqu'à présent, les données de SEMEVAL concernaient uniquement la langue anglaise, mais depuis l'édition 2016, des données en français ont été fournies. À notre connaissance, notre système, ainsi que les systèmes d'extraction d'expressions-cibles qui seront proposés pour cette édition de SEMEVAL pour le français sont les premiers du genre.

La suite de ce travail est structurée de la façon suivante. Nous dressons d'abord un état de l'art des méthodes employées (section 2). Puis nous présentons nos travaux sur l'anglais (section 3.1) et le français (section 3.2) ainsi que les méthodes utilisées. Enfin, nous terminons par une analyse de l'efficacité des traits utilisés (section 4.1) ainsi qu'une analyse des erreurs réalisées par nos systèmes (section 4).

2 État de l'art

Nous présentons brièvement les systèmes supervisés proposés pour l'extraction de cible à SEMEVAL 2015 ayant obtenus les meilleurs résultats (cf table 1).

TABLE 1 – Résultats SEMEVAL15 Extraction des expressions-cibles

Système	F1	Précision	Rappel
EliXa	0.700 5	0.689 3	0.712 2
NLANGP	0.671 2	0.705 3	0.640 2
IHS-RD	0.631 3	0.675 8	0.592 3
Lsislif	0.622 3	0.713 6	0.551 7

Au vu des données proposées, les participants ont opté pour des systèmes usant d'apprentissage supervisé, lesquels utilisent principalement des modèles CRF (Lsislif (Hamdan *et al.*, 2015), IHS-RD (Chernyshevich & Stankevitch, 2015) et NLANGP (Toh & Su, 2015)). Le système EliXa (San Vicente *et al.*, 2015) est l'exception avec un modèle Perceptron. Les traits utilisés sont assez similaires d'un système à l'autre. Parmi les traits utilisés par les différents systèmes, on retrouve souvent des traits morphologique tel le mot-forme, la forme du mot (casse, symbole, chiffre, etc.), le lemme, des bigrammes de mots, des préfixes et suffixes de 1 à n caractères. Des traits morphosyntaxiques sont également utilisés, que ce soit des étiquettes morphosyntaxiques, ou l'arbre syntaxique de la phrase (Toh & Su, 2015). Les lexiques sont également largement employés, soit sous la forme de lexiques de l'opinion (Toh & Su, 2015; Hamdan *et al.*, 2015), soit sous la forme de lexique des domaines étudiés (restaurants et hôtels) (Toh & Su, 2015). Enfin, différentes méthodes de partitionnement ont également été employées pour créer de façon non-supervisée des ressources. (San Vicente *et al.*, 2015; Toh & Su, 2015) créent tous deux des partitions de Brown (Brown *et al.*, 1992) à partir du corpus de Yelp [1]. (San Vicente *et al.*, 2015) créent également des partitions de Clark (Clark, 2003) et des partitions Word2Vec (Mikolov *et al.*, 2013). Ces méthodes de partitionnements non supervisées

1. Corpus de commentaires de restaurant : Yelp Dataset Challenge https://www.yelp.com/dataset_challenge

permettent de regrouper des mots en classes dont l'identifiant sert de trait pour décrire les mots dans les systèmes d'apprentissages supervisés.

3 Expérimentations

L'objectif de nos travaux est de créer un système proche de l'état de l'art d'extraction d'expressions-cibles en anglais puis de le transposer au français. Nous nous sommes inspirés du système Lsislif (Hamdan *et al.*, 2015) car il obtient de bons résultats (top 4) et parce que le modèle ainsi que l'ensemble des traits utilisés sont simples à transférer au français.

Afin de pouvoir nous comparer avec les systèmes des éditions précédentes, nous travaillons avec les données de SEMEVAL 2015 pour l'anglais. En revanche, le français est une nouveauté de l'édition 2016, nous utilisons donc les données de SEMEVAL 2016 mais nous ne disposons pas encore d'autres résultats par rapport auxquels nous positionner. Le corpus anglais est constitué de 254 commentaires avec 1279 annotations d'expressions-cibles. Le corpus français est constitué de 335 commentaires avec 1770 annotations d'expressions-cibles. À notre connaissance, il n'existe pas encore d'autre système d'extraction d'expressions-cibles pour le français.

3.1 Création d'un système pour l'anglais

Notre système réalise les tâches habituelles de tokenisation des phrases (avec *TweetTokenizer* de NLTK), d'extraction de traits, de l'entraînement d'un modèle et de son évaluation sur un corpus de test. Nous utilisons l'outil Wapiti (Lavergne *et al.*, 2010) pour entraîner des modèles CRF. Les fichiers sont annotés pour l'entraînement du modèle CRF selon la notation IOB.

Les organisateurs de SEMEVAL ont fourni un script de référence permettant de mesurer les performances d'un système qui prend en entrée un fichier annoté par notre système et un fichier *gold*. Les expérimentations sur l'anglais sont évaluées sur le corpus de test qui avait été utilisé durant l'édition 2015.

Nous utilisons les traits suivants pour représenter les termes des commentaires :

- le mot-forme du terme ;
- l'étiquette morphosyntaxique (*POS tag*), l'entité nommée et le syntagme non-récursif (*chunk*) identifiés par Senna [2] ;
- le lemme identifié par TreeTagger [3] ;
- la forme du mot compressée (toutes les majuscules/minuscules/chiffres/autres sont remplacées par A/a/0/_, p. ex. on représente *Judging* par Aa) ;
- le type de mot (majuscule, ponctuation, numérique, ...) ;
- tous les préfixes du mot de 1 à 4 lettres ;
- tous les suffixes du mot de 1 à 4 lettres ;
- si le mot est fonctionnel en utilisant la liste NLTK.

En plus de ces traits, nous avons également utilisé le lexique de l'opinion de Bing Liu (Hu & Liu, 2004), le lexique de subjectivité MPQA (Wilson *et al.*, 2005) ainsi qu'une version du lexique MPQA

2. http://ronan.collobert.com/senna
3. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

enrichie avec des expressions multi-mots. Nous utilisons la polarité des mots comme trait pour notre système. Les lexiques de l'opinion ont deux fonctions. Premièrement, ils permettent de filtrer les mots de l'opinion car ceux-ci ne peuvent pas être (sauf exception) des expressions-cibles. Deuxièmement, les lexiques de l'opinion permettent de détecter les expressions explicites de l'opinion ce qui est utile car une expression-cible n'est valide que si la phrase exprime une opinion à son encontre.

Les résultats des expérimentations de notre système *LINA* sont présentés dans la table 2. Nous avons également reporté les scores de référence indiqués par les organisateurs.

TABLE 2 – Résultats des systèmes sur le corpus anglais

Système	F1	Précision	Rappel
EliXa	0.700 5	0.689 3	0.712 2
NLANGP	0.671 2	0.705 3	0.640 2
LINA	0.634 4	0.683 0	0.592 3
IHS-RD	0.631 3	0.675 8	0.592 3
Lsislif	0.622 3	0.713 6	0.551 7
baseline-semeval	0.480 7		

Le système EliXa est clairement devant les autres, mais les autres systèmes (NLANGP, IHS-RD, Lsislif, et le notre *LINA*) ont des performances relativement similaires. La principale différence entre le système EliXa et les autres est le modèle utilisé. Le système EliXa utilise un modèle perceptron alors que les autres utilisent un modèle CRF. L'un des avantages du modèle perceptron est qu'il peut manipuler des valeurs continues pour décrire les traits alors que le modèle CRF est limité aux valeurs symboliques.

3.2 Transposition du système au français

Comme expliqué précédemment, nous avons transposé le système de l'anglais au français. Nous avons donc globalement utilisé les mêmes traits que pour l'anglais, en adaptant les systèmes à la langue lorsque cela était nécessaire (étiquetage morphosyntaxique, lemmatisation, mot fonctionnel) ainsi que les lexiques de l'opinion. Nous avons utilisé le lexique de l'opinion LIDILEM (Augustyn *et al.*, 2006) et le lexique de l'opinion Blogoscopie (Vernier & Monceaux, 2010).

En l'absence d'autres systèmes auxquels se comparer, on peut observer dans la table 3 que les performances semblent aussi bonnes que pour l'anglais. Nous avons utilisé une validation croisée en 10 plis car seul le corpus d'entraînement est disponible.

TABLE 3 – Résultats du système sur le corpus français

Système	F1	Précision	Rappel
LINA_fr	0.702 1	0.751 6	0.659 5

Une analyse des erreurs rencontrées par notre système en anglais et en français est présentée dans la section 4.2, mais il existe deux types d'erreurs que nous avons observés uniquement en français. Le premier concerne les coordinations et autres expressions d'énumérations, le deuxième porte sur les compléments du nom.

Premièrement, dans les commentaires, la coordination est souvent exprimée par les conjonctions

et et *ou*, mais aussi par la virgule (dans le cas d'une énumération). Notre système n'a pas appris correctement à détecter les coordinations et énumérations ce qui le conduit souvent à n'extraire qu'une partie des expressions-cibles. Par exemple dans la phrase ***Ambiance** et **deco** super.* seul *Ambiance* est identifié par notre système, et dans *Des **hamburgers**, des **pizzas** succulentes, des **brushettas**, et une **escalope milanaise** exceptionnelle.*, seuls *pizza* et *brushettas* sont identifiés par notre système. Pour résoudre ce type d'erreur nous pourrions utiliser un CRF d'ordre supérieur afin de propager les annotations plus loin et ainsi "sauter" l'expression de la conjonction.

Deuxièmement, nous avons observé à plusieurs reprises que notre système identifiait, à tort, le complément du nom comme expression-cible au lieu du groupe nominal principal. Par exemple dans la phrase *La terrasse du restaurant est sublime.*, notre système va extraire *restaurant* comme expression-cible alors que l'expression-cible correcte est la *terrasse*. Des traits exploitant une analyse syntaxique de la phrase pourraient permettre de résoudre ce type de problème.

4 Analyse comparative des systèmes

Nous avons réalisé une étude comparative des systèmes sur les deux langues. Nous présentons d'abord une comparaison des performances des systèmes selon les traits utilisés et leurs impacts. Puis, nous listons des erreurs fréquemment rencontrées en anglais et en français.

4.1 Comparaison des traits

Nous avons réalisé différents tests avec nos systèmes en retirant ou en gardant certaines catégories de traits. L'objectif est d'évaluer l'apport de chaque catégorie de traits à la tâche, en anglais comme en français. Les tableaux 4 et 5 présentent respectivement les résultats de nos tests en anglais et en français. Le système *all* correspond au système complet avec tous les traits, *all - morph* au système privé de l'étiquette morphosyntaxique, *all - form* au système privé des traits sur les mots (mot-forme, préfixes/suffixes, casses, etc.), *all - lex* au système privé des lexiques. Les systèmes *morph*, et *form* correspondent respectivement aux systèmes dotés des étiquettes morphosyntaxiques uniquement, et des traits sur les mots uniquement.

TABLE 4 – Étude des traits en anglais

Traits	F1	Précision	Rappel
all	0.634 4	0.683 0	0.592 3
all - form	0.620 5	0.705 9	0.553 5
form	0.617 3	0.697 7	0.553 5
all - lex	0.617 1	0.667 4	0.573 8
all - morph	0.605 7	0.682 9	0.544 3
morph	0.443 4	0.603 2	0.350 6

TABLE 5 – Étude des traits en français

Traits	F1	Précision	Rappel
all	0.702 1	0.751 6	0.659 5
all - lex	0.697 4	0.743 6	0.657 8
all - form	0.695 6	0.759 9	0.642 3
all - morph	0.694 4	0.750 5	0.646 8
form	0.693 1	0.750 6	0.644 6
morph	0.483 6	0.692 6	0.373 2

On observe que les performances du système en anglais comme en français baissent peu lorsque l'on retire un seul type de trait. Une étude par ablation des traits du système Lsislif est présentée par Hamdan 2015. Cependant, les traits sont retirés un par un alors que nous retirons des groupes de traits selon leur catégorie (morphologique, morphosyntaxique ou lexical). Leurs résultats ne montrent pas de traits clairement plus importants que d'autres.

Dans nos expérimentations, le résultat le plus intéressant concerne le système *all - form*. Ce système obtient de bon résultats en anglais et en français alors qu'il n'utilise que les étiquettes morphosyntaxiques et les lexiques de l'opinion. L'absence des traits sur les mots signifie qu'il ne peut pas apprendre le vocabulaire du domaine étudié, dans notre cas sur la restauration. C'est la structure de la phrase et la proximité avec un mot de l'opinion qui permet de détecter les expressions-cibles. Ce résultat suggère qu'il est possible d'entraîner un modèle d'extraction d'expressions-cibles qui n'est pas dépendant d'un domaine et donc de supprimer le travail d'annotation souvent nécessaire lorsque l'on change de domaine.

4.2 Étude des erreurs

Nous avons étudié quelques erreurs rencontrées par nos systèmes en anglais et en français. Le type d'erreur le plus fréquemment rencontré en anglais et en français concerne les mots qui apparaissent dans tout le corpus et sont, selon le contexte, des expressions-cibles ou non. Le mot *restaurant* en particulier, est utilisé à la fois pour décrire un lieu mais sans être une cible de l'opinion et en tant que cible de l'opinion. Pour résoudre ce type d'erreur, il ne suffit pas de détecter l'expression de l'opinion dans une phrase, il faut que le système soit capable de rattacher précisément une expression de l'opinion à sa cible. Ce type d'erreur milite pour une prise en compte d'informations syntaxiques de plus haut niveau, par exemple l'arbre syntaxique des phrases (comme (Toh & Su, 2015)). L'autre principal type d'erreur partagé sur l'anglais et le français porte sur les entités nommées. En particulier lorsque celle-ci sont formées de mots fonctionnels (p. ex *The Four Seasons*, *Le Baroudeur*). L'usage de systèmes de détection d'entité nommées plus performants devrait aider à réduire ce type d'erreur.

Le dernier type d'erreur rencontré par les systèmes en anglais et français porte sur l'expression d'opinion implicite dans la phrase. Nous appelons opinion implicite l'usage d'expressions qui font appel à des connaissances générales. Par exemple dans la phrase *But $500 for a **dinner for two** that didn't include Wine ?*, l'auteur n'utilise pas de mot marqueur de l'opinion. Ici, le point d'interrogation est un marqueur faible de l'expression d'une opinion et un annotateur peut estimer qu'un repas à 500$ sans vin est un peu cher. Cependant, il est souvent possible de résoudre ce type d'erreur en élargissant le contexte observé. Au lieu de considérer les commentaires comme une suite de phrases sans liens, il faut regarder l'ensemble des phrases et propager l'information d'une phrase à l'autre. Ainsi, la phrase précédente de l'exemple était *There is "Expensive-but-worth-it" and there is "Expensive-and-WTF" ?* et la phrase suivante *Are you freaking kidding me ?*, phrases dans lesquelles l'opinion y est clairement exprimée. Dans d'autres commentaires, on peut observer une opinion clairement négative en début ou fin de commentaire, et les autres phrases composant le commentaire sont un descriptif de tout ce qui n'allait pas. Au vu de ces erreurs, il nous semble que la propagation d'information de l'opinion d'une phrase à l'autre est une piste à développer.

5 Conclusion

Nous avons présenté un système d'extraction d'expressions-cibles [4] pour l'anglais proche de l'état de l'art, ainsi que sa transposition au français. Le système produit par transposition commet des erreurs propres à la langue française mais les résultats obtenus sur le corpus d'entraînement sont similaires aux résultats état de l'art en l'anglais. De plus, nous avons étudié l'efficacité des traits en anglais et en

4. Le code source est disponible sous licence libre : https ://github.com/daimrod/lina-opinion-target-extractor

français qui tend à montrer qu'il est possible de réaliser un système d'extraction d'expressions-cibles indépendant au domaine. Pour finir, l'analyse comparative des erreurs commises par nos systèmes en anglais et français ouvre plusieurs pistes d'améliorations pour chacun des types d'erreurs identifiés.

6 Remerciement

Ce travail a bénéficié d'une aide de l'Etat attribuée au labex COMIN Labs et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01.

Références

AUGUSTYN M., BEN HAMOU S., BLOQUET G., GOOSSENS V., LOISEAU M. & RINCK F. (2006). Lexique des affects : constitution de ressources pédagogiques numériques. In *Colloque International des étudiants-chercheurs en didactique des langues et linguistique.*, p.≈-, Grenoble, France.

BROWN P. F., PIETRA V. J. D., DE SOUZA P. V., LAI J. C. & MERCER R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, **18**(4), 467–479.

CHERNYSHEVICH M. & STANKEVITCH V. (2015). Ihs-rd-belarus : Identification and normalization of disorder concepts in clinical notes. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 380–384, Denver, Colorado : Association for Computational Linguistics.

CLARK A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, p. 59–66, Stroudsburg, PA, USA : Association for Computational Linguistics.

HAMDAN H. (2015). *Sentiment Analysis in Social Media*. PhD thesis, Université d'Aix-Marseille.

HAMDAN H., BELLOT P. & BECHET F. (2015). Lsislif : Crf and logistic regression for opinion target extraction and sentiment polarity analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 753–758, Denver, Colorado : Association for Computational Linguistics.

HU M. & LIU B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, p. 168–177.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, **abs/1310.4546**.

PONTIKI M., GALANIS D., PAPAGEORGIOU H., MANANDHAR S. & ANDROUTSOPOULOS I. (2015). Semeval-2015 task 12 : Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 486–495, Denver, Colorado : Association for Computational Linguistics.

SAN VICENTE I. N., SARALEGI X. & AGERRI R. (2015). Elixa : A modular and flexible absa platform. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 748–752, Denver, Colorado : Association for Computational Linguistics.

TOH Z. & SU J. (2015). Nlangp : Supervised machine learning system for aspect category classification and opinion target extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 496–501, Denver, Colorado : Association for Computational Linguistics.

VERNIER M. & MONCEAUX L. (2010). Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques. *Traitement Automatique des Langues*, **51**(1), 125–149.

WILSON T., WIEBE J. & HOFFMANN P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*.