Approche mixte utilisant des outils et ressources pour l'anglais pour l'identification de fragments textuels subjectifs français

Michel Généreux et Thierry Poibeau
Laboratoire d'Informatique de Paris-Nord
(CNRS UMR 7030 et Université Paris 13)
99, av. J.-B. Clément – 93430 Villetaneuse
{genereux,poibeau}@lipn.fr

Résumé – Abstract

Cet article présente une méthode hybride pour l'analyse de la subjectivité dans un texte en misant d'une part sur des outils et des ressources disponibles pour l'anglais, et d'autre part sur une approche mixte combinant statistique et linguistique. Les annotations d'un corpus en anglais projetées sur un corpus parallèle en français forment la base d'apprentissage pour un classifieur de phrases subjectives (approche statistique) ainsi qu'un identificateur de patrons syntaxiques pertinents au corpus subjectif (approche linguistique). La phase linguistique permet d'une part de consolider les résultats obtenus lors de la phase statistique, et d'autre part de raffiner l'analyse à des fragments de phrase plus circonscrits.

This article presents a hybrid method for the analysis of subjectivity in a text focusing primarily on the tools and resources available for English and on a hybrid approach combining statistics and linguistics. Annotations from a corpus in English projected on a parallel corpus in French form the learning basis for a subjectivity classifier (statistical approach) and an identifier of syntactic patterns relevant to subjective texts (languistic approach). On one hand, the linguistic phase allows for the consolidation of the results from the statistical phase, while on the other hand refines the analysis of sentences to shorter fragments.

Mots-clefs – Keywords

Analyse de subjectivité, corpus parallèle, patron syntaxique
Subjectivity analysis, parallel corpus, syntactic pattern

1 Introduction

Il y a depuis quelques années un intérêt croissant pour l'extraction automatique d'éléments en rapport avec les sentiments et les émotions dans les textes, et pour fournir des outils susceptibles d'être intégrés dans un traitement plus global des langues et de leur aspect subjectif. La plupart des recherches à ce jour ont porté sur l'anglais, ce qui s'explique principalement par la disponibilité des ressources pour l'analyse de la subjectivité, telles que les lexiques et les corpus annotés manuellement. Dans cet article, nous misons sur un corpus parallèle anglais-français ainsi qu'un outil permettant de classifier chaque phrase du corpus anglais en subjectif ou objectif. En projetant ces annotations sur le corpus français, nous obtenons la ressource nécessaire pour entraîner un système pour classifier automatiquement ces phrases selon leur niveau de subjectivité. Cette ressource nous servira aussi à identifier les patrons syntaxiques les plus saillants dans les phrases annotées comme subjectives. Notons donc qu'en dehors d'un pont (ici un corpus parallèle) entre la langue source (ici l'anglais) et la langue cible (ici le français), notre approche ne nécessite aucune annotation manuelle, ni pour pour la création des ressources, ni pour le développement des classifieurs. Ainsi, compte tenu d'un pont entre l'anglais et la langue cible, les méthodes peuvent être appliquées à d'autres langues. Notons la somme considérable de travail mise en œuvre pour la création de ces ressources en anglais. Nous tenterons donc d'évaluer à travers cette campagne la valeur de cette approche mixte pour identifier les phrases subjectives, voire les fragments subjectifs, d'un texte en français.

Après avoir fourni un bref état de l'art, nous présentons comment nous avons obtenu notre ressource principale (un corpus de phrases en français annotées selon leur niveau de subjectivité), suivi des expériences menées pour le développement du classifieur statistique et de l'identifiant de fragments subjectifs. L'architecture du système est présentée sur la figure 1.

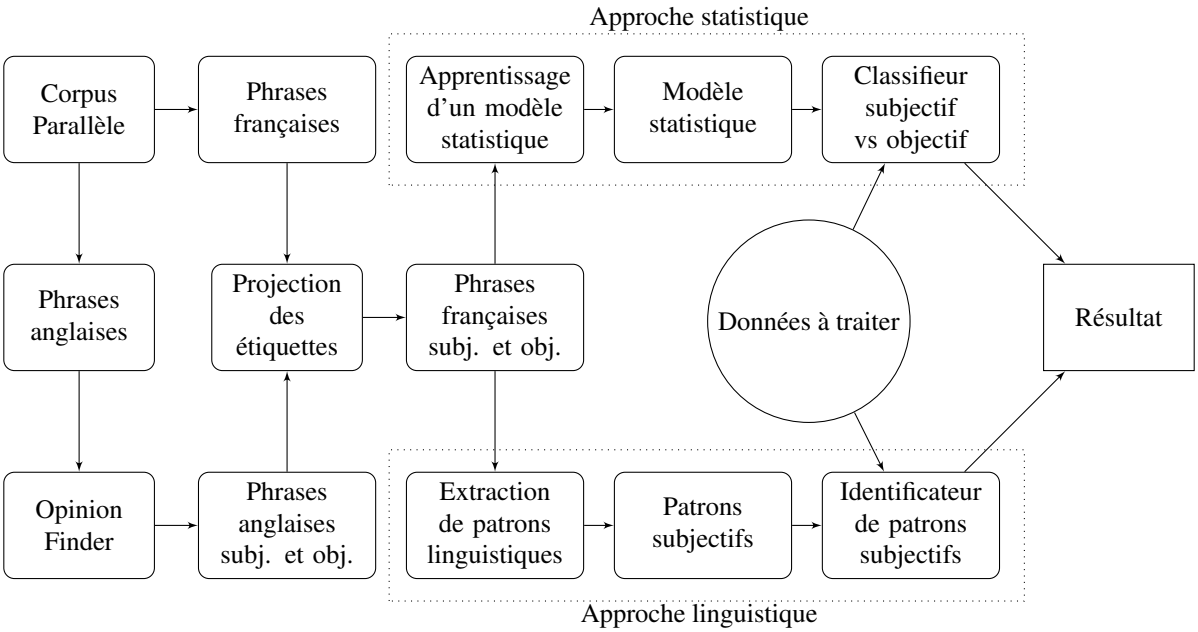

Figure 1: Une approche mixte pour la détection de subjectivité

2 État de l'art

Notons d'abord que la tâche à laquelle nous nous attaquons (segmentation de textes selon leur niveau de subjectivité) est nouvelle et représente donc un défi considérable. Nous pouvons en effet prendre pour exemple la campagne Text Analysis Conference (TAC 2009), qui a décidé de supprimer la tâche de création de résumés d'opinions, présente lors de TAC 2008, les organisateurs ayant convenu de la difficulté à extraire des éléments subjectifs d'un texte et de les organiser convenablement pour la production d'un résumé. Il y a tout de même plusieurs travaux valables dans ce domaine qui peuvent être mentionnés ici.

Dans le domaine des opinions, les travaux précédents se sont surtout attardés à leur détection ainsi qu'à la gradation de leur niveau affectif, et ce selon trois niveaux principaux de sous-tâches. La première sous-tâche consiste à distinguer les textes *subjectifs* des textes *objectifs* (Yu & Hatzivassiloglou, 2003). La seconde sous-tâche s'attarde à classer les textes subjectifs en *positifs* ou *négatifs* (Turney, 2002). Le troisième niveau de raffinement essaie de déterminer jusqu'à quel point les textes sont positifs ou négatifs (Wilson *et al.*, 2004). L'impulsion donnée par des campagnes telles que *TREC Blog Opinion Task* depuis 2006 est incontestable (Zhang *et al.*, 2007; Dey & Haque, 2008). Signalons les efforts récents pour réintroduire des approches plus linguistiques et discursives (prise en compte de la modalité, de l'énonciateur) dans ce domaine (Asher *et al.*, 2008).

Les méthodes d'analyse automatique de subjectivité ont été utilisées dans une grande variété d'applications de traitement de texte, telles que le suivi de l'humeur sur des forums en ligne (Lloyd *et al.*, 2005; Balog *et al.*, 2006), le classement d'opinions ou commentaires (Turney, 2002; Pang *et al.*, 2002) ainsi que leur extraction (Hu & Liu, 2004), l'analyses sémantique des textes (Wiebe & Mihalcea, 2006; Esuli & Sebastiani, 2006) et le résumé de textes d'opinion (Bossard *et al.*, 2008). Le travail se rapprochant le plus du nôtre est (Mihalcea *et al.*, 2007), où un lexique bilingue et un corpus parallèle traduit manuellement sont utilisés pour générer un classifieur de phrases selon leur niveau de subjectivité pour le roumain.

Bien que beaucoup de travaux récents en analyse de la subjectivité mettent l'accent sur le sentiment (un type de subjectivité, positif ou négatif), notre travail porte sur la reconnaissance de la subjectivité en général. Comme le soulignent (Banea *et al.*, 2008), les chercheurs en analyse de sentiment ont démontré qu'une approche en deux

étapes est souvent bénéfique, dans laquelle on distingue d'abord l'objectif du subjectif, pour ensuite classifier les éléments subjectifs en fonction de la polarité (Yu & Hatzivassiloglou, 2003; Pang & Lee, 2004; Wilson *et al.*, 2005; Kim & Hovy, 2006). En fait, le problème de la distinction subjective versus objective s'est souvent avéré plus difficile que l'étape ultérieure visant à classifier selon la polarité (positive vs négative). Les améliorations dans la première auront donc un effet nécessairement bénéfique sur la seconde, ce qui est par ailleurs montré dans certains travaux (Takamura *et al.*, 2006).

3 Création d'un corpus de phrases françaises subjectives et objectives

À partir d'un corpus parallèle anglais-français[1] et d'un outil permettant de classer automatiquement des phrases en anglais selon qu'elles soient objectives ou subjectives (OpinionFinder (Riloff *et al.*, 2003)), nous projetons les étiquettes obtenues pour les phrases en anglais sur le corpus français. Le corpus comporte 1 130 104 paires de phrases parallèles, et après un nettoyage sommaire pour éliminer des phrases trop courtes, qui parasitent l'analyse (e.g. *the House adjourned at ...*) ou dont le ratio des longueurs respectives s'éloigne trop de un et fait suspecter une erreur d'alignement ou de traduction, le corpus est réduit à 63 251 phrases (paires). Les phrases anglaises sont soumises à OpinionFinder pour l'étiquetage *subjectif* ou *objectif* de chacune d'entre elles. Plus précisément, OpinionFinder utilise deux classifieurs basés sur des indicateurs subjectifs obtenus d'un grand lexique.

Le premier classifieur étiquète chaque phrase comme *subjectif* ou *objectif*. Ce classifieur utilise une stratégie qui donne l'exactitude la plus élevée. Évalué sur 9 732 phrases (4 352 objectives et 5 380 subjectives) du corpus MPQA[2], ce classifieur obtient une exactitude de 74% et une précision de 78,4%, un rappel de 73,2% et une F-mesure de 75,7% pour l'étiquette *subjectif*. L'exactitude de référence est de 55,3%.

Le deuxième classifieur optimise la précision au détriment du rappel. Une phrase est classifiée comme subjective ou objective que si on peut le faire avec un certain degré de confiance, sinon la phrase reçoit l'étiquette "inconnu". Évaluée sur les mêmes 9 732 phrases du corpus MPQA, cette stratégie obtient 91,7% de précision et 30,9% de rappel pour l'étiquette *subjective*. La précision est de 83,0% et le rappel 32,8% pour l'étiquette *objective*.

Puisque dans cette campagne nous nous intéressons plus particulièrement aux fragments de texte subjectif, notre stratégie a été de favoriser la précision et d'assigner cette étiquette que si les deux classifieurs produisaient une étiquette subjective. Cette stratégie a scindé les 63 251 phrases en 27 121 phrases subjectives et 36 130 phrases objectives. Ces étiquettes ont été projetées sur chacune des phrases correspondantes du corpus français.

4 Approche statistique

Notre première approche prend comme unité de traitement la phrase complète. À partir de 10 000 phrases de ce corpus français étiquetées subjectives et 10 000 étiquetées objectives, nous avons entraîné un classifieur de type SVM (Joachims, 1998) avec l'implémentation Weka[3] et un noyau linéaire. Nous avons choisi comme traits tous les lemmes des adjectifs, noms, verbes et adverbes tels que donnés par TreeTagger[4]. Puisque seulement 863 de ces lemmes se sont révélés avoir une valeur de gain d'information non-nulle, nous avons conservé les 800 traits ayant un gain d'information le plus élevé. Quelques-uns de ces traits sont présentés dans le tableau 2, avec une indication si le trait considéré se retrouve aussi dans le grand lexique (anglais) utilisé par Opinion Finder. Évalué sur 1 000 phrases objectives et 1 000 subjectives, le classifieur obtient une exactitude de 83,3% (voir tableau 1).

Précision	**Rappel**	**F-Mesure**	**Classe**
0.789	0.909	0.845	objectif
0.893	0.757	0.819	subjectif

Table 1: Évaluation du classifieur de subjectivité sur 1 000 phrases

[1]Tiré du corpus Hansard du parlement canadien et disponible à `http://www.cse.unt.edu/~rada/wpt/data/English-French.training.tar.gz`. Alignement produit par Ulrich Germann.

[2]Multi-Perspective Question Answering, disponible à `www.cs.pitt.edu/mpqa/`.

[3]`http://www.cs.waikato.ac.nz/ml/weka/`.

[4]`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.

Trait	Présence dans 1 000 phrases subjectives	Présence dans 1 000 phrases objectives	Membre du Lexique d'OF?
vouloir	133	4	oui
croire	29	0	oui
devoir	83	8	oui
savoir	72	4	oui
être	388	151	non
bien	61	14	oui
dire	101	14	non
faire	120	44	non
très	45	4	oui
penser	30	1	oui
rapport	13	21	non
espérer	27	1	oui
suivre ou être	36	2	non
motion	15	21	non
pouvoir	95	24	oui
fait	35	3	oui

Table 2: Seize traits avec le gain d'information le plus élevé

5 Approche linguistique

Une analyse statistique à base de traits ne permet pas une analyse fine en deçà de la phrase en elle-même. Pour être en mesure d'étiqueter des fragments de phrases, nous avons adopté une approche basée sur la détection de patrons linguistiques[5]. Ces patrons sont issus de travaux visant à extraire des expressions subjectives pour l'anglais (Riloff & Wiebe, 2003), dont nous reproduisons quelques exemples dans le tableau 3.

FORME SYNTAXIQUE	PATRON EXEMPLE
<sujet> verbe-passif	<sujet> was satisfied
<sujet> verbe-actif	<sujet> complained
<sujet> verbe-actif <objet direct>	<sujet> dealt blow
<sujet> verbe infinitif	<sujet> appear to be
<sujet> auxiliaire nom	<sujet> has position
verbe-actif <objet direct>	endorsed <objet direct>
infinitif <objet direct>	to condemn <objet direct>
verbe infinitif <objet direct>	get to know <objet direct>
nom auxiliaire <objet direct>	fact is <objet direct>
nom préposition <syntagme nominal>	opinion on <syntagme nominal>
verbe-actif préposition <syntagme nominal>	agrees with <syntagme nominal>
verbe-passif préposition <syntagme nominal>	was worried about <syntagme nominal>
infinitif préposition <syntagme nominal>	to resort to <syntagme nominal>

Table 3: Patrons syntaxiques en anglais pour la subjectivité

Les patrons pour l'anglais font intervenir une analyse linguistique très détaillée, comme par exemple la fonction grammaticale ou la détection des formes actives ou passives. Ne disposant pas des ressources nécessaires pour une analyse semblable dans le cas du français, nous nous sommes limités aux cinq formes syntaxiques illustrées dans le tableau 4. Pour la reconnaissance des patrons syntaxiques pour le français, nous avons utilisé le sécateur (*chunker*) de TreeTagger. 4 903 patrons ont été extraits des phrases subjectives et 2 814 des phrases objectives. L'idée principale est ici de trouver un ensemble de patrons syntaxiques qui sont pertinents pour les phrases subjectives, i.e. qui ont une distribution statistiquement plus élevée que dans les phrases objectives. Nous avons donc sélectionné tous les patrons apparaissant au moins dix fois au total dans tout le corpus, avec au moins 80% d'apparition dans les phrases subjectives, ce qui produit une liste de 79 patrons dits *subjectifs*, dont nous reproduisons les dix plus fréquents, dans des contextes réels, dans le tableau 5. Ainsi, l'approche linguistique consiste à extraire tous ces patrons des textes à traiter.

[5]Nous appellons cette approche linguistique, bien qu'elle utilise en partie des donnés statistiques.

FORME SYNTAXIQUE	PATRON EXEMPLE
<syntagme nominal> verbe <syntagme nominal>	<syntagme nominal> appuyer <syntagme nominal>
<syntagme nominal> verbe verbe	<syntagme nominal> devoir être
<syntagme nominal> verbe préposition	<syntagme nominal> être en
verbe verbe <syntagme nominal>	avoir prendre <syntagme nominal>
verbe préposition <syntagme nominal>	dire à <syntagme nominal>

Table 4: Patrons syntaxiques en français pour la subjectivité

6 Résultats

Nous avons participé à la tâche numéro deux, qui consistait à :

> Segmenter un texte en passages objectifs, qui donnent des faits, ou le thème du texte, et en passages subjectifs qui délivrent une opinion, un sentiment, sur ces faits, concernant ce thème. Un passage peut aller d'un mot (par exemple un modifieur) à plusieurs phrases.

Notre système doit être évalué sur deux corpus : un ensemble d'articles issus des journaux *Le Monde* (corpus 1, 25 176 articles, 327 000 phrases) et les débats du Parlement européen (corpus 2, 19 370 articles, 180 635 phrases). Les données de référence de la tâche 2 ont été établies par un vote majoritaire entre les passages notés "subjectif" par les participants. Un passage réunissant autant de votes "subjectif" qu'"objectif" a été noté "indéterminé". Pour harmoniser les résultats, la taille de référence d'un passage équivaut au texte compris entre deux ponctuations. Pour chaque fichier d'exécution, un passage est considéré comme *subjectif* si au moins un mot de ce passage était marqué *subjectif*. Les F-scores calculés sont relatifs à l'"accord" entre les participants sur les passages subjectifs. Les organisateurs de DEFT09 ont aussi souhaité calculer le score des participants à la tâche 2 dans le cas où leur système fût appliqué à la tâche 1, détection du caractère objectif/subjectif global d'un texte, en comptabilisant au niveau de chaque document les mots marqués subjectif et les mots marqués objectif. Un document avec une majorité de mots marqués subjectif a été marqué comme subjectif.

6.1 Exécutions

Nous avons produit trois exécutions différentes, chacune mettant plus ou moins à contribution les étiquettes des deux approches pour la subjectivité. La première (exécution 1) ne considère que les étiquettes fournies par l'approche statistique, donc chaque phrase est soit subjective, soit objective. La deuxième exécution (exécution 2) considère les étiquettes fournies par les deux approches : une phrase est classée subjective que si elle s'est vue attribuer une étiquette subjective par l'approche statistique et possède au moins un patron syntaxique subjectif issu de l'approche mixte. Cette stratégie vise à augmenter la précision au détriment du taux de rappel. Finalement, la troisième exécution (exécution 3) attribue un indice de confiance selon le principe suivant :

- chaque patron (approche linguistique) extrait d'une phrase étiquetée subjective par l'approche statistique se voit attribué un indice de confiance de 1;
- chaque patron (approche linguistique) extrait d'une phrase étiquetée objective par l'approche statistique se voit attribué un indice de confiance de 0.2;
- chaque fragment de phrase étiquetée subjective par l'approche statistique mais ne faisant pas partie d'un patron se voit attribuée un indice de confiance de 0.8.

Nous avons aussi produit quatre exécutions hors-concours (hc) dans le but d'affiner notre analyse des résultats. Ces exécutions ont été réalisées une fois les données de références obtenues pour la tâche 2. Ces quatres exécutions portent directement sur les passages tels que définis précédemment et ont pour but d'évaluer la performance de quatre classifieurs utilisant chacun une stratégie distincte :

- un passage est subjectif s'il est étiqueté subjectif par l'approche statistique (exécution 4);
- un passage est subjectif s'il renferme au moins un patron selon l'approche linguistique (exécution 5);

FORME SYNTAXIQUE	Fréquence totale	% dans les phrases subjectives
<syntagme nominal> être en	181	82%
Le message : oui, **Mikhaïl Khodorkovski est en** prison depuis octobre 2003, et son procès pour évasion fiscale et malversations se poursuit à Moscou, mais la société qu'il a créée, elle, continue de travailler.		
<syntagme nominal> devoir être	150	83%
D'autres magistrats doivent être entendus.		
<syntagme nominal> avoir de	68	81%
En octobre, **Samir Azzouz a de** nouveau été arrêté, sur la base d'écoutes et de diverses observations.		
dire à <syntagme nominal>	54	85%
Or, ce que j'ai **dit à son sujet** est une analyse à long terme qui n'encourage guère les entreprises à investir.		
joindre à <syntagme nominal>	40	83%
Peu avant 8 heures, une délégation de postiers se **joint à la troupe**.		
opposer à <syntagme nominal>	38	97%
Pour l'heure, M. Mer s'**oppose à toute nouvelle baisse** du barème.		
arriver à <syntagme nominal>	36	81%
La majorité de ceux qui **arrivent à Ceuta** sont marocains.		
profiter de <syntagme nominal>	34	85%
Je **profite de l'occasion** pour exprimer à nouveau notre disposition à résoudre autour d'une table de négociations le différend prolongé entre les Etats-Unis et Cuba, a-t-il déclaré, sur des principes d'égalité, de réciprocité, de non-ingérence et de respect mutuel.		
penser à <syntagme nominal>	34	88%
De nombreux pays - je **pense à la Suède** ou au Canada - ont entrepris un réexamen systématique des actions conduites par l'Etat.		
<syntagme nominal> appuyer <syntagme nominal>	31	97%
Tandis que **l'IGAD appuie le TFG,** avec le soutien de l'Ethiopie, la Ligue arabe, emmenée par le Soudan, a mis sur pied un mécanisme concurrent qui a permis, au terme de trois jours de pourparlers, de signer, dans la nuit de dimanche à lundi, un protocole en douze points entre des représentants des Tribunaux islamiques et du gouvernement de transition.		

Table 5: Dix patrons syntaxiques français en contexte

- un passage est subjectif s'il est étiqueté subjectif par l'approche statistique ET renferme au moins un patron selon l'approche linguistique (exécution 6);
- un passage est subjectif s'il est étiqueté subjectif par l'approche statistique OU renferme au moins un patron selon l'approche linguistique (exécution 7).

Les résultats de notre système pour les sept exécutions sont présentés dans le tableau 6.

7 Discussion

Comme notre approche mixte repose essentiellement sur la disponibilité d'un bon corpus de phrases étiquetées objectives ou subjectives, examinons la validité de ce corpus. Notre corpus découle d'une projections des étiquettes obtenues automatiquement par un système pour l'anglais (OpinionFinder), paramétré pour obtenir une grande précision (autour de 80%) avec une bonne performance (environ 75%) sur les étiquettes subjectives, ces mesures découlant d'évaluations impliquant des jugements humains. Après projection sur les phrases en français, ces mesures devraient refléter une dégradation plus ou moins grande due à des erreurs d'alignement ou de traduction, que nous avons par ailleurs essayer de limiter. Néammoins, un classifieur SVM construit à partir de ces phrases obtient un très bon niveau de performance (83.3%, voir tableau 1), une mesure d'évaluation qui peut être mise en relation directe avec des évaluations humaines, puisqu'une majorité de traits saillants du classifieur se retrouve dans le lexique de référence utilisé par OpinionFinder (voir tableau 2). Notons au passage, comme l'ont d'ailleurs fait (Riloff & Wiebe, 2003) pour l'anglais, que le lemme du nom *fait*, terme objectif par excellence, est paradoxalement un bon indicateur de subjectivité!

Bien que limité à des patrons syntaxiques assez simples, l'approche linguistique révèle un certain nombre de

Tâche	Corpus	Exécution	Précision	Rappel	F-Mesure	Exactitude
1	1:Le Monde	1:Phrase subj.	0.573	0.613	0.592	N/A
1	1:Le Monde	2:Phrase subj. avec >= 1 patron	0.520	0.500	0.510	N/A
1	1:Le Monde	3:Indice de confiance pondéré	0.573	0.614	0.593	N/A
2	1:Le Monde	1:Phrase subj.	0.701	0.871	0.777	N/A
2	2:Parlement	1:Phrase subj.	0.806	0.791	0.799	N/A
		Moyenne	**0.754**	**0.831**	**0.788**	N/A
2	1:Le Monde	2:Phrase subj. avec >= 1 patron	0.929	0.579	0.714	N/A
2	2:Parlement	2:Phrase subj. avec >= 1 patron	0.816	0.580	0.678	N/A
		Moyenne	**0.873**	**0.580**	**0.696**	N/A
2	1:Le Monde	3:Indice de confiance pondéré	0.699	0.869	0.775	N/A
2	2:Parlement	3:Indice de confiance pondéré	0.805	0.789	0.797	N/A
		Moyenne	**0.752**	**0.829**	**0.786**	N/A
2:hc	1:Le Monde	4:Phrase subj.	0.595	0.723	0.653	0.615
2:hc	2:Parlement	4:Phrase subj.	0.608	0.780	0.683	0.638
		Moyenne	**0.602**	**0.752**	**0.668**	**0.627**
2:hc	1:Le Monde	5:Patron	0.482	0.090	0.152	0.497
2:hc	2:Parlement	5:Patron	0.563	0.060	0.108	0.507
		Moyenne	**0.523**	**0.075**	**0.130**	**0.502**
2:hc	1:Le Monde	6:Phrase subj. et >= 1 patron	0.590	0.077	0.136	0.512
2:hc	2:Parlement	6:Phrase subj. et >= 1 patron	0.593	0.053	0.098	0.508
		Moyenne	**0.592**	**0.065**	**0.117**	**0.510**
2:hc	1:Le Monde	7:Phrase subj. ou >= 1 patron	0.579	0.737	0.648	0.600
2:hc	2:Parlement	7:Phrase subj. ou >= 1 patron	0.605	0.787	0.684	0.637
		Moyenne	**0.592**	**0.762**	**0.666**	**0.619**

Table 6: Résultats globaux (hc = hors-concours)

structures linguistiques intéressantes. Par exemple, les patrons *opposer à <syntagme nominal>* et *<syntagme nominal> appuyer <syntagme nominal>* sont assez intuitivement subjectifs et apparaissent donc presque toujours (97%) dans des phrases subjectives. D'autres, en revanche, sont moins directement assimilables à des expressions subjectives, comme par exemple *arriver à <syntagme nominal>*, et n'apparaissent d'ailleurs que dans 81% de phrases subjectives. Une évaluation externe et directe de la pertinence de ces patrons n'est possible que par comparaison avec les résultats des autres participants tels que présentés dans le tableau 6.

Rappelons que les scores indiqués pour la tâche 1 sont obtenus en assignant à chaque document une étiquette équivalente au compte majoritaire (subjectif ou objectif) de mots marqués subjectifs et objectifs. De l'aveu même des organisateurs de DEFT09 au moment d'écrire cet article, ces résultats sont un peu moins bons que ceux des participants à la tâche 1, mais étant donné la plus grande difficulté de la tâche 2, ils sont quand même très encourageants.

Pour la tâche 2, la classification utilisant la méthode statistique (exécution 1) présente la meilleure performance, bien que le taux de précision de l'approche linguistique (exécution 2) suggère que les patrons linguistiques identifient avec une bonne fiabilité les passages subjectifs. Le faible taux de rappel s'explique en partie par la faible quantité de patrons retenue (79). L'exécution 3 ne peut malheureusement pas être analysée pour ce qu'elle est, puisque l'évaluation qui nous a été fournie par DEFT ne tenait pas compte de l'indice de confiance, principal caractéristique de cet exécution.

Les quatre exécutions hors-concours (4, 5, 6 et 7) illustrent la faiblesse de l'approche mixte à identifier certains passages subjectifs, à tout le moins à se mettre d'accord avec les autres participants. Bien que la majorité de ces patrons apparaissent pertinents et que le niveau de précision reste au-delà de ce qu'on obtiendrait par chance (50%), le faible niveau de rappel fait chuter la performance sous un niveau acceptable. L'approche linguistique basée sur des patrons de base simples, apparaît donc beaucoup trop conservatrice par rapport à l'approche statistique dans l'identification de passages subjectifs. Dans les deux approches, la différence entre la segmentation des passages (le fragment délimité par la ponctuation) et la structure de base pour l'entraînement des deux approches (la phrase complète) a pu jouer un rôle négatif dans la composition des traits ou des patrons linguistiques. La nature du corpus (journaux versus débats parlementaires) ne semble pas avoir été un facteur déterminant.

8 Conclusion

Nous avons présenté une approche mixte pour traiter de la subjectivité dans les textes. Cette approche nous a permis de montrer d'une part jusqu'à quel point le transfert de ressources et l'utilisation d'outils disponibles pour une langue source permet de construire des outils à base de statistique et de linguistique pour la détection de passages subjectifs. Nos expériences ont montré que ce transfert semble aboutir à des résultats plus performants pour des outils statistiques, bien que l'approche linguistique fournisse un éclairage nouveau sur la composition d'expressions porteuses de subjectivité, ouvrant une voie de recherche tournée vers une meilleure compréhension et un raffinement des ressources et outils linguistiques mis en œuvre pour l'élaboration de patrons linguistiques plus performants.

Références

Asher N., Benamara F. et Mathieu Y. (2008). Distilling opinion in discourse: A preliminary study. In *Coling 2008: Companion volume: Posters*, p. 7–10, Manchester, UK: Coling 2008 Organizing Committee.

Balog K., Mishne G., et de Rijke M. (2006). Why are they excited? identifying and explaining spikes in blog mood levels. In *EACL-2006*.

Banea C., Mihalcea R., Wiebe J. et Hassan S. (2008). Multilingual subjectivity analysis using machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), Honolulu, Hawaii, October 2008*.

Bossard A., Généreux M. et Poibeau T. (2008). Description of the lipn systems at tac2008: Summarizing information and opinions. In *Text Analysis Conference 2008, Workshop on Summarization Tracks, November 17-19 2008, National Institute of Standards and Technology, Gaithersburg, Maryland USA*.

Dey L. et Haque M. (2008). Opinion mining from noisy text data. In *AND '08: Proceedings of the second workshop on Analytics for noisy unstructured text data*, p. 83–90, New York, NY, USA: ACM.

Esuli A. et Sebastiani F. (2006). Determining term subjectivity and term orientation for opinion mining. In *EACL 2006*.

Hu M. et Liu B. (2004). Mining and summarizing customer reviews. In *ACM SIGKDD*.

Joachims T. (1998). Text categorization with support vector machines: Learning with many relevant features. p. 137–142.

Kim S.-M. et Hovy E. (2006). Identifying and analyzing judgment opinions. In *HLT/NAACL 2006*.

Lloyd L., Kechagias D. et Skiena S. (2005). Lydia: A system for large-scale news analysis. In *SPIRE 2005*.

Mihalcea R., Banea C. et Hassan S. (2007). Learning multilingual subjective language via cross-lingual projections. In *ACL 2007*.

Pang B. et Lee L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL 2004*.

Pang B., Lee L. et Vaithyanathan S. (2002). Thumbs up? sentiment classification usingmachine learning techniques. In *EMNLP 2002*.

Riloff E. et Wiebe J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing*, Sapporo, JP.

Riloff E., Wiebe J. et Wilson T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In W. Daelemans & M. Osborne, Eds., *Proceedings of CONLL-03, 7th Conference on Natural Language Learning*, p. 25–32, Edmonton, CA.

Takamura H., Inui T. et Okumura M. (2006). Latent variable models for semantic orientations of phrases. In *EACL 2006*.

Turney P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL 2002*.

Wiebe J. et Mihalcea R. (2006). Word sense and subjectivity. In *COLING-ACL 2006*.

Wilson T., Wiebe J. et Hoffmann P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP 2005*.

Wilson T., Wiebe J. et Hwa R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of AAAI-04, 21st Conference of the American Association for Artificial Intelligence*, p. 761–769, San Jose, US: AAAI Press / The MIT Press.

Yu H. et Hatzivassiloglou V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *EMNLP 2003*.

Zhang W., Yu C. et Meng W. (2007). Opinion retrieval from blogs. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, p. 831–840, New York, NY, USA: ACM.