DEFT2009

Actes du cinquième DÉfi Fouille de Textes

Proceedings of the Fifth DEFT Workshop

22 juin 2009

Paris, France

DEFT2009

Actes du cinquième DÉfi Fouille de Textes

22 juin 2009
Paris, France

Préface

Depuis 2005, les campagnes nationales d'évaluation DEFT (DÉfi Fouille de Textes) proposent des thématiques de recherche exploratoires axées sur la fouille de texte. Les thématiques abordées jusqu'à présent se sont focalisées sur l'identification de locuteurs (2005), la segmentation thématique de textes (2006), la détermination de l'expression d'une opinion (2007) et sur l'identification et la catégorisation de documents (2008).

L'édition 2007 du défi a été plus spécifiquement consacrée à la fouille d'opinion. L'objectif fixé concernait l'attribution automatique de valeurs d'opinion (sur une échelle de 2 à 3 valeurs) à des textes présentant un avis argumenté, qu'il soit positif ou négatif. Cette édition rassembla dix équipes venant d'horizons divers : académiques, services R&D de sociétés privées et jeunes chercheurs. Les résultats obtenus varièrent fortement selon les techniques employées et la difficulté des corpus.

Pour cette cinquième édition du DÉfi Fouille de Textes, nous avons fait le choix de proposer une nouvelle tâche en fouille d'opinion. Cette édition est particulière à plus d'un titre. En premier lieu, nous proposons pour la première fois une édition multilingue du défi (français, anglais et italien) reposant sur deux types de corpus. Un premier corpus se compose d'articles de journaux issus des quotidiens *Il Sole 24 Ore, Le Monde* et *The Financial Times*.

Un second corpus comprend les interventions parlementaires issues du Parlement européen. En second lieu, nous avons varié les tâches autour de la thématique de la fouille d'opinion en offrant trois tâches :

- La détection du caractère objectif ou subjectif global d'un texte depuis un corpus d'articles de journaux ;
- La détection des passages subjectifs d'un texte, sur deux corpus : articles de journaux et débats parlementaires ;
- Enfin, la détermination du parti politique d'appartenance de chaque intervenant dans le corpus parlementaire.

L'organisation adoptée pour cet atelier de clôture a retenu un découpage en sessions, fondée sur les différentes tâches composant cette campagne, que suivent les chapitres de ces actes :

- Le premier chapitre revient sur des travaux réalisés en fouille d'opinion lors d'une édition précédente de DEFT, en 2007. Une intervention émane d'une société privée travaillant dans le domaine de la fouille d'opinion, l'autre communication présente les travaux originaux d'une équipe universitaire ayant obtenu de très bons résultats lors de la campagne DEFT'07 ;
- Le second chapitre se rapporte à l'organisation de cette nouvelle édition. Nous y présentons les différentes tâches retenues ainsi que les corpus que nous avons rassemblés. Nous détaillons les tests que nous avons réalisés avec des évaluateurs humains sur chacune des tâches. Enfin, nous introduisons les résultats obtenus par les participants au défi ;
- Le troisième chapitre s'intitule catégorisation globale et se rapporte aux premières et troisièmes tâches (respectivement, identification subjective/objective globale d'un article et identification du parti politique d'appartenance d'un parlementaire). Il rassemble les articles de trois équipes ayant travaillé exclusivement sur la première tâche et d'une équipe qui s'est essayée aux deux tâches ;
- Le dernier chapitre, intitulé subjectivité locale, concerne plus particulièrement la seconde tâche du défi (identification des passages subjectifs d'un texte). À ce titre, il propose les approches suivies par les deux équipes qui se sont investies sur cette tâche. Précisons toutefois que le premier article de ce chapitre se rapporte également à la participation de l'équipe à la première tâche ; compte-tenu de la spécificité de cette tâche locale, il nous a semblé pertinent de rassembler dans un même chapitre les interventions des participants à la tâche.

Le comité d'organisation de DEFT'09

Comités

Comité de programme

Patrick Paroubek (LIMSI–CNRS, ILES), *président*

Catherine Berrut (LIG)

Fabrice Clérot (France Telecom)

Guillaume Cleuziou (LIFO)

Béatrice Daille (LINA)

Marc El-Bèze (LIA)

Patrick Gallinari (LIP6)

Thierry Hamon (LIPN)

Fidélia Ibekwe-SanJuan (ELICO)

Pascal Poncelet (LIRMM)

Jean-Michel Renders (XRCE)

Christophe Roche (LISTIC)

Mathieu Roche (LIRMM)

Pascale Sébillot (IRISA)

François Yvon (LIMSI–CNRS, TLP)

Comité d'organisation

Martine Hurault-Plantet (LIMSI–CNRS, Orsay), *co-responsable*

Cyril Grouin (LIMSI–CNRS, Orsay), *co-responsable*

Béatrice Arnulphy (LIMSI–CNRS, Orsay)

Jean-Baptiste Berthelin (LIMSI–CNRS, Orsay)

Sarra El Ayari (LIMSI–CNRS, Orsay)

Anne Garcia-Fernandez (LIMSI–CNRS, Orsay)

Arnaud Grappy (LIMSI–CNRS, Orsay)

Isabelle Robba (LIMSI–CNRS, Orsay)

Pierre Zweigenbaum (LIMSI–CNRS, Orsay)

Table des matières

Session I – Invités

Exploration de corpus pour l'analyse de sentiments

Sigrid Maurel[(1)] et Luca Dini[(1)]

[(1)] CELI France, SAS
12-14, rue Claude Genin
38000 Grenoble
{maurel, dini}@celi-france.com
http://www.celi-france.com

Résumé – Abstract

Dans cet article nous présentons l'amélioration de notre système d'extraction de sentiments et opinions SYBILLE. Par rapport à la première version nous avons changé la méthode statistique (initialement basée sur l'apprentissage automatique) vers une exploration ontologique de corpus. Cette nouvelle méthode qui s'insère avant la méthode symbolique permet de survoler les textes et d'en avoir très rapidement un premier aperçu. Elle extrait les concepts des domaines présents dans les textes et en fournissant une ontologie elle facilite le développement de la grammaire de la méthode symbolique.

In this article we present the improvement of our sentiment and opinion mining system SYBILLE. Compared to the first version we changed the statistic method (formerly based on machine learning) to an ontologic corpus discovery. This new method which comes before the symbolic method allows to skim the texts and to get very quickly a first glance. It extracts the concepts of the present domains of the texts and by giving an ontology facilitates the development of the grammar of the symbolic method.

Mots-clefs – Keywords

extraction de sentiments et opinions, exploration de corpus

sentiment and opinion mining, corpus discovery

1 Introduction

1.1 Motivation

Cet article s'intéresse à la classification de textes d'opinion en langue française. Dans ce cas précis, la classification a pour objectif l'analyse de sentiments exprimés dans différents types de textes comme par exemple dans des forums de discussion sur Internet où les internautes échangent des avis et s'entraident. Les textes issus de forums sur Internet constituent des sources d'informations spontanées et récentes, incontournables pour acquérir, au jour le jour, des connaissances sur les consommateurs, pour anticiper leurs besoins et leurs attentes afin de tenter d'améliorer la relation client/fournisseur. En analysant ces textes d'opinion le fournisseur d'un produit ou d'un service peut mieux réagir aux desiderata de ses clients, le client peut de son côté s'inspirer des sentiments et opinions d'autres clients sur le produit auquel il s'intéresse et profiter ainsi d'une aide à la décision (acquérir ou ne pas acquérir le produit, choisir plutôt le produit A ou le produit B, etc.).

Comme le montrent de nombreux travaux de socio- et psycho-linguistique (Sproull & Kiesler, 1991), la communication médiée par ordinateur favorise l'expression des émotions, sentiments et opinions souvent contrôlés ou réprimés dans des cadres de communication plus traditionnels visant à étudier le point de vue des consommateurs (interviews face à face, enquêtes fermées, enquêtes ouvertes, etc.). De là, naît l'intérêt des analystes pour ces sources d'informations.

Les corpus utilisés pour le développement des systèmes de classification sont composés de textes (ou *threads*, fils de discussion) provenant de forums sur Internet qui parlent entre autres de tourisme, de jeux vidéo et d'imprimantes. Un texte (ou message) dans un forum contient un jugement argumenté de l'auteur du message, positif, négatif ou parfois mitigé, sur un sujet donné. Mais il contient aussi des parties exemptes de sentiments, comme c'est le cas par exemple dans la description du jeu vidéo sur lequel porte la critique. L'objectif de l'analyse est donc d'identifier avec précision les parties pertinentes pour la classification automatique du texte dans son entier.

Une des difficultés de la classification en *positif* et *négatif* réside dans la nécessité d'une bonne analyse syntaxique du texte, analyse qui peut se révéler particulièrement difficile dans des cas de coordination entre plusieurs parties d'une phrase, d'anaphore ou de coréférence (la reprise d'un argument présent plus tôt dans le document). Une autre difficulté du langage naturel pour l'analyse automatique de sentiments réside dans les contextes intentionnels, pour lesquels l'expression d'opinion n'est pas un vrai sentiment. C'est le cas dans une phrase comme:

> « Je croyais que la France était un beau pays. »

(Dini & Mazzini, 2002) ont montré le lien qui existe entre les structures syntaxiques et sémantiques d'une phrase et l'expression de l'opinion qu'elle véhicule. Ainsi l'analyse de la phrase par *paquets de mots* donne des résultats peu satisfaisants alors qu'une analyse syntaxique du texte peut aider à trouver les expressions qui contiennent des opinions. Les deux phrases suivantes contiennent les mêmes *paquets de mots* sans pour autant exprimer les mêmes sentiments. En effet, la première phrase contient un sentiment positif alors que la deuxième est négative:

> « Je l'ai apprécié pas seulement à cause de ... »
> « Je l'ai pas apprécié seulement à cause de ... »

Dans le cadre de sa participation à la campagne d'évaluation DEFT'07 (c.f. section 6 pour plus de détails), CELI France a mis au point trois méthodes pour classer les textes des différents corpus. La première est une méthode symbolique qui inclut un système d'extraction d'information adapté aux corpus. Elle est basée sur des règles d'un analyseur syntaxico-sémantique. Cet analyseur contient un lexique de mots qui véhiculent des sentiments sur lesquels réagissent les règles de la grammaire. La deuxième est une méthode statistique basée sur des techniques d'apprentissage automatique. Enfin, la dernière, SYBILLE, est une méthode hybride qui combine les techniques des deux précédentes pour aboutir à des résultats très précis.

Depuis cette première évaluation CELI France a amélioré son système SYBILLE. Nous avons changé la méthode statistique initialement basée sur un apprentissage automatique (Maurel *et al.*, 2009) par une exploration ontologique de corpus qui donne très vite une première idée sur les concepts véhiculés dans les textes. De ce fait elle nous permet d'orienter de manière très précise le développement de la méthode symbolique qui passe à la deuxième place lors du processus. La méthode hybride reste inchangée et combine les résultats des deux premières méthodes.

L'analyse des textes se fait au niveau de la phrase, les sentiments d'un document sont extraits phrase par phrase, et c'est seulement ensuite qu'une valeur globale est attribuée au message entier. Ceci permet d'extraire une information contextuelle qui est donc très précise.

Les sections suivantes présentent brièvement l'état de l'art et les corpus utilisés pour s'attarder ensuite sur les trois méthodes développées et en fournir une première évaluation. Une section sera dédiée à l'interface graphique SYBILLE pour présenter les possibilités que celle-ci offre aux utilisateurs, avant de conclure cet article.

1.2 État de l'art

L'analyse de sentiments se concentre aujourd'hui sur l'attribution d'une polarité à des expressions subjectives (les mots et les phrases qui expriment des opinions, des émotions, des sentiments, etc.) afin de décider de l'orientation d'un document (Turney, 2002), (Wilson *et al.*, 2004) ou de la valeur positive/négative/neutre d'une opinion dans un document (Hatzivassiloglou & McKeown, 1997), (Yu & Hatzivassiloglou, 2003), (Kim & Hovy, 2004).

Des travaux allant au-delà ont mis l'accent sur la force d'une opinion exprimée où chaque proposition dans une phrase peut avoir un fond neutre, faible, moyen ou élevé (Wilson *et al.*, 2004). Des catégories grammaticales ont été utilisées pour l'analyse de sentiments dans (Bethard *et al.*, 2004) où des syntagmes adjectivaux comme *trop riche* ont été utilisés afin d'extraire des opinions véhiculant des sentiments. (Bethard *et al.*, 2004) utilisent une évaluation basée sur la somme des scores des adjectifs et des adverbes classés manuellement, tandis que

(Chklovski, 2006) utilise des méthodes fondées sur un modèle pour représenter des expressions adverbiales de degré telles que *parfois*, *beaucoup*, *assez* ou *très fort*.

L'approche que nous avons adoptée pour la classification de textes d'opinion est caractérisée par une utilisation mixte d'une technologie symbolique fondée sur des règles et d'une technologie statistique reposant sur l'extraction d'une ontologie du domaine, approche dans laquelle la méthode symbolique a un poids plus important (Dini, 2002), (Dini & Mazzini, 2002), (Maurel *et al.*, 2007), (Maurel *et al.*, 2008), (Bosca & Dini, 2009). La technologie symbolique fait d'abord une analyse du texte phrase par phrase et en extrait ensuite les relations qui véhiculent des sentiments, tandis que la technologie statistique traite les textes en une seule phase et fournit une liste de concepts et termes spécifiques du texte.

Il convient de remarquer que, contrairement à d'autres approches actuelles, la technologie de l'analyse de sentiments développée à CELI France (SYBILLE) ne se limite pas à une analyse lexicale (c'est-à-dire identification et pondération de mots positifs et négatifs), mais s'étend à une analyse syntaxique et sémantique. L'analyse syntaxique est effectuée par le biais d'une analyse robuste de surface telle que celles décrites par (Aït-Mokhtar & Chanod, 1997), (Basili *et al.*, 1999), (Aït-Mokhtar *et al.*, 2001), donnant ainsi un résultat très proche de celui produit par des grammaires de dépendance.

2 Les corpus

Les données de type forums de discussion sur Internet s'articulent comme un flux d'interactions, comme par exemple: demande-réponse, argument-contre argument, commentaire-désaccord, etc. Ce flux est distribué sur une dimension temporelle qui nécessite un traitement chronologique du fil de discussion. Contrairement aux corpus utilisés par (Wilson *et al.*, 2004), il n'est pas nécessaire ici d'identifier la personne à qui est associé un sentiment, car dans 95 % des cas, les discours analysés sont des discours à la première personne. Un exemple de flux d'interactions est donné en figure 1.

Les corpus utilisés sont assez différents les uns des autres, que ce soit par la taille des corpus eux-mêmes que par la taille de chaque *thread* (fil de discussion). Nous avons utilisé les corpus de DEFT'07 auxquels nous avons ajouté des corpus collectés sur Internet. Ces corpus nous ont permis d'augmenter la diversité des sujets et de répondre aux exigences de nos clients, selon les domaines demandés. Nous avons donc des textes de domaines très différents, entre autres du domaine de la restauration rapide, du nucléaire, de l'alimentation infantile, etc.[1] Certains corpus sont structurés, d'autres contiennent beaucoup de messages en style *texto*, et le nombre de fautes d'orthographe présentes dans les messages varie aussi beaucoup.

Le comité d'organisation de DEFT'07 a pris soin de nettoyer ses corpus (Grouin *et al.*, 2007). Ainsi, les fins de ligne ont été normalisées, les caractères encodés en ISO-Latin, et les textes ont été annotés manuellement.[2] Les corpus de DEFT'07 contiennent des critiques de films, de livres et de spectacles, des tests de jeux vidéo, des relectures d'articles scientifiques (de différentes conférences sur l'intelligence artificielle) et des notes de débats parlementaires (sur la loi de l'énergie).

Les textes de nos corpus portent essentiellement sur le tourisme (en France et ailleurs dans le monde), les jeux vidéo (critiques et problèmes) et les imprimantes (conseils d'achats). Ils comprennent d'un côté des aides à la solution de problèmes, mais aussi des avis sur des lieux visités et des produits achetés. Chaque *thread* contient les messages des auteurs participant aux forums sur un sujet donné.

Les fautes d'orthographe[3] dans les textes des corpus posent parfois des problèmes d'analyse. Heureusement, les règles syntaxiques de la grammaire (voir la section 4 sur la méthode symbolique) sont dans la plupart des cas assez tolérantes pour permettre l'accord entre un nom et un adjectif, ou un nom et un verbe même si le *e* ou le *s* manque. Mais malheureusement, il y a aussi des messages tellement mal écrits dans les corpus (par exemple en style *texto*, c'est-à-dire avec beaucoup d'abréviations) que l'analyse peut échouer.[4]

[1] Ces derniers ne seront pas abordés plus profondément dans cet article, mais les ressources sont disponibles dans notre système SYBILLE.

[2] En ce qui concerne nos propres corpus, ils sont encodés en UTF-8 et nous n'avons effectué aucun nettoyage. Tous les corpus sont disponibles au format XML.

[3] Nous avons fait le choix de garder les textes tels quels, donc de ne pas appliquer un correcteur automatique d'orthographe ou un lexique d'abréviations. Ce choix s'explique par la volonté de garder toutes les caractéristiques stylistiques présentes dans les textes. Nous considérons qu'une uniformisation des entrées à ce moment-là du processus nous ferait perdre des informations utiles.

[4] D'après ce que nous avons pu observer, ces messages sont heureusement en minorité dans les corpus et ne modifient pas les résultats de façon significative.

```
    Avis sur les châteaux de la loire en France

angie-443*5, posté le 08-10-2006 à 16:18:50:
   J'ai besoin de vos conseil s.v.p. Je vais passer une ou deux
   journée dans la valée de la loire. Y-a-t-il un château en
   loire avec un jardin semblable à celui de versailles (en
   beauté et en superficie)? J'aime aussi l'aspect extérieurs
   des châteaux, plus que l'intérieur. Ce qui me plait d'une
   ville est tout d,abord ses rues piétonnes, animées et
   pittoresques, ses charmantes places et ses promenades.

[...]

BaLadeur, posté le 13-10-2006 à 11:23:43:
   Je partage l'avis d'Aston sur de nombreux points. Villandry
   est quelconque mais son jardin transformé en potager géant
   vaut le détour. Chenonceau est certainement le plus
   photogénique donc le plus connu et il le mérite largement.
   Si tu recherches la monumentalité comme à Versailles, la
   magnificence en plus, il faut absoluement voir Chambort.
   Enfin s'il faut ne visiter qu'une ville ce sera Tours.

[...]

zeus77, posté le 21-10-2006 à 21:59:33:
   A Amboise j'aime beaucoup le manoir du Clos-Lucé qui fut
   la dernière maison de Léonard de Vinci. Le parc est très
   agréable. Enfin un chateau où l'on pourrait vivre! Quel
   changement par rapport aux chateaux royaux. Un chateau que
   j'aime bien aussi c'est celui Du Moulin à Lassay sur Croisne
   entre Contres et Romorantin.

[...]
```

Figure 1: Exemple d'un flux d'interactions de messages, du domaine du *tourisme*. L'orthographe et la ponctuation n'ont pas été modifiées.

3 Méthode statistique distributionnelle

La méthode statistique distributionnelle (inspirée par (Bosca & Dini, 2009)) exploite le corpus afin d'identifier un échantillon de termes qui sont fortement distinctifs du domaine analysé. Les termes ainsi extraits sont ensuite mis en relation par moyens d'analyses statistiques d'occurences de mots à l'intérieur du corpus. L'issue résultante consiste en une représentation structurée (bien que pas une ontologie formelle) des concepts clés du domaine. Ce processus de découverte fonctionne en deux phases (détaillées ci-après), et est particulièrement efficace face à de grands corpus. Il permet d'avoir une idée du contenu et des mots-clés ou des concepts très rapidement.

Le processus qui analyse les textes fournit une moyenne d'exploration du corpus qui permet d'obtenir une sensation des sujets et des concepts discutés et quelles sont les relations entre ces concepts. Pour pouvoir configurer la grammaire de la méthode symbolique (c.f. la section 4) il est utile de savoir ne serait-ce que à peu près ce que l'on cherche dans les textes.

3.1 LOR

La méthode pour extraire une ontologie des textes a besoin de deux corpus. Le premier est le *corpus d'études* qui contient les textes d'un forum ou sous-forum avec les termes spécifiques qui nous intéressent particulièrement. Le deuxième est un corpus générique, un *corpus de référence* qui contient par exemple les textes de tous les sous-forums du forum en question, ou un ensemble de textes d'une source générale comme par exemple une encyclopédie. Dans le cas du *corpus d'études* de nos expériences il s'agit d'un corpus dynamique qui est généré à partir d'une requête de mots-clés sur une base de données contenant les textes.

terme	score de pertinence	occurrences
laitage	8.755508	912
compote	8.644531	2246
biberon	8.35303	9403
blédina	7.907746	354
allaiter	7.7495713	1992
féculent	7.651336	1227
allaitement	7.439835	1338
bib	7.280261	136
candia	7.0153956	1475
diversification	6.496531	1119

Table 1: LOR des termes pertinents du domaine de l'alimentation infantile.

Notre stratégie d'extraction de termes est basée sur la comparaison de fréquence entre le corpus du domaine (*corpus d'études*) et le corpus général (*corpus de référence*). Notre approche exploite comme mesure de termes spécifiques une version modifiée du bien connu *Log Odds Ratio* (LOR, c.f. (Everitt, 1992), (Baroni & Bisi, 2004)). La fonction de mesure de termes spécifiques adoptée dans nos expériences est une combinaison pondérée du LOR et de la mesure de fréquence de termes (TF). Elle peut être formalisée comme suit:

$$TermSpec = k * \frac{TermDF * GC_Docs}{TermGF * DC_Docs} + TermDF * (1 - k)$$

où $TermDF$ représente la fréquence d'un terme donné dans le corpus du domaine, $TermGF$ sa fréquence dans le corpus général, DC_Docs le nombre de documents compris dans le corpus du domaine tandis que GC_Docs est le nombre de documents dans le corpus général. Nous avons expérimenté l'extraction de terminologie avec trois valeurs différentes de k (0, 0.5 et 1) ayant pour résultat donc trois fonctions de mesure différentes: une mesure de TF pure (avec $k = 0$), une mesure de LOR pure (avec $k = 1$) et une mesure équitablement ponderée de LOR/TF (avec $k = 0.5$); les paragraphes suivants décrivent en détail les différentes issues résultant de l'adoption de ces fonctions de mesure différentes.

Le tableau 1 donne un exemple du domaine de l'alimentation infantile. Il montre les dix premiers mots intéressants ou pertinents du corpus analysé. Le score attribué est une valeur de pertinence, ici le terme *laitage* est considéré plus pertinent que le terme *compote*, alors qu'il apparaît moins souvent.

3.2 RI

Les termes ainsi extraits (par LOR) du corpus du domaine sont enrichis avec une terminologie sémantiquement reliée par moyens d'un modèle distributionnel basé sur corpus. Une telle terminologie est basée sur l'hypothèse que le sens d'un terme donné émerge implicitement des contextes différents dans lesquels il apparaît (ici nous entendons par contexte l'unité de texte comme un paragraphe, un document ou une fenêtre textuelle). Puis, la deuxième phase du processus est une approche basée sur la co-ocurrence des mots, le sens d'un mot étant défini par son contexte. Cette méthode calcule donc un vecteur de sens pour chaque mot et plus les vecteurs de deux mots sont proches l'un de l'autre (plus l'angle entre eux est petit) plus leurs sens sont similaires.

L'indexage aléatoire *Random Indexing* (RI) exploite un modèle algébrique afin de représenter la sémantique des termes dans un espace à N dimensions (un vecteur de N coordonnées). L'approche RI crée une matrice *termes par contextes* où chaque ligne représente le degré d'appartenance d'un terme donné aux contextes différents. L'algorithme RI assigne une signature aléatoire à tous les contextes (un vecteur très épars de N coordonnées, avec peu d'éléments non null, choisis aléatoirement) et génère ensuite le modèle de l'espace du vecteur en performant une analyse statistique des documents dans le corpus du domaine et en accumulant sur les lignes des termes toutes les signatures des contextes où les termes apparaissent.

Selon cette approche si deux termes différents ont un sens similaire ils devraient apparaître dans des contextes similaires (à l'intérieur d'un même document ou entourés des mêmes mots), en résultant caractérisés par des coordonnées proches dans l'espace sémantique ainsi généré. Dans nos études de cas nous avons appliqué la technique RI pour générer des clusters de termes en sélectionnant dans l'espace sémantique les termes avec la

terme	score de pertinence	contexte
vache	0.81468266	lait de vache
tire	0.7715641	tire-lait
soja	0.7615201	lait de soja
poudre	0.75552726	lait en poudre
lactose	0.75137496	lait sans lactose
montée	0.7460638	montée de lait
chèvre	0.6994075	lait de chèvre
intolérance	0.6894124	intolérance au lait
biberon	0.64999706	biberon de lait
régurgiter	0.63957196	régurgiter le lait

Table 2: RI des termes en contexte avec le mot *lait*.

distance minimale du mot analysé en exploitant la mesure de distance cosine.

Le tableau 2 donne un autre exemple du domaine de l'alimentation infantile. Il montre les dix premiers mots en contexte avec le mot *lait*.

3.3 Comparaison de LOR et RI sur corpus

Le premier pas de notre expérience découverte est de comprendre si nous pouvons produire une *photo instantanée* générale des contenus du corpus. Afin d'effectuer une telle tâche, les résultats de LOR appliqués sur la base de documents apparaissent quelque peu décevants. En effet, et la liste de termes comme LOR pure et la liste de LOR/TF pondérée (avec un poids de 0.5) semblent être plutôt orientées à mettre en évidence des termes imprévus que des termes descriptifs pertinents.

Nous notons bien sûr l'apparition de quelques termes qui sont probables à caractériser le domaine en question ou les opinions que les auteurs des textes peuvent en avoir, mais la tendance est occultée par les termes qui sont inattendus et probablement arrivés par des discussions hors-sujet. Afin de minimiser l'impact d'hors-sujet et de bruit venant d'analyses peu structurées de pages web, nous restreignons l'algorithme LOR uniquement à des phrases qui contiennent un mot-clé préalablement choisi.

Une fois que nous avons isolé un ensemble de concepts qui constitue le pivot de notre étude, nous pouvons enquêter sur les comportements différents des deux algorithmes. En même temps nous pouvons évaluer l'effet de phénomènes linguistiques comme l'ambiguïté sémantique et la partie syntaxique de notre méthodologie proposée.

4 Méthode symbolique

Une fois que le processus statistique est terminé et les connaissances du corpus acquises par l'exploration, le développement de la grammaire symbolique peut se baser dessus pour améliorer les règles symboliques.

Comme nous l'avons dit plus haut, la méthode symbolique se base sur une analyse syntaxique du texte faite par un analyseur fonctionnel et relationnel (c.f. les travaux sur l'analyse syntaxique et sémantique de (Basili *et al.*, 1999), (Aït-Mokhtar *et al.*, 2001), (Dini, 2002), (Dini & Mazzini, 2002), (Dini & Segond, 2007)). Cet analyseur traite, phrase par phrase, un texte donné en entrée et en extrait, pour chaque phrase, les relations syntaxiques présentes. Il s'agit de relations syntaxiques fonctionnelles de base, telles que le modifieur d'un nom, d'un verbe, sujet et objet d'une phrase, ainsi que de relations plus complexes telles que la coréférence entre deux syntagmes au sein d'une même phrase.

L'utilisateur a la possibilité d'élaborer une grammaire à sa guise et d'ajouter de nouvelles règles afin d'extraire les relations auxquelles il s'intéresse. Pour ce faire, il peut modifier les règles d'extraction de relations (par exemple ajouter des règles pour de nouvelles relations), augmenter/diminuer les traits sur les mots dans le lexique qui agissent sur les règles, enlever certaines parties du traitement, etc.

La polarité positive ou négative attribuée au message entier[5] dépend du rapport entre la quantité de relations

[5]L'attribution d'un sentiment global au message entier est utilisée dans des contextes spécifiques, comme par exemple pour l'évaluation

d'opinions positives et négatives. Une majorité de relations d'opinions positives détermine une polarité positive du message, tandis qu'une majorité de relations d'opinions négatives provoque une polarité négative.

4.1 Grammaire

La grammaire utilisée a été initialement développée afin d'extraire les relations de sentiments exprimés dans une phrase dans le cadre d'un projet sur le tourisme en France. Elle a été ensuite modifiée et améliorée en vue de la participation à DEFT'07 (c.f. section 6, (Maurel *et al.*, 2007)). Dans un deuxième temps, la grammaire a été divisée en deux parties: une première partie de base (la grammaire *générique*) s'appliquant à tous les textes qui contiennent des sentiments, et une deuxième partie pour chaque domaine différent, selon le sujet du corpus: tourisme, jeux vidéo, imprimantes, etc. Les différences se situent essentiellement dans les lexiques appliqués, chaque domaine ayant ses propres mots et expressions.

Ainsi les mots se rattachant à la vitesse (*lent*, *rapide*, etc.) ont des polarités différentes selon qu'ils qualifient une imprimante ou un voyage. De même, comme le montrent les phrases ci-dessous, l'adjectif *effrayant* est plutôt perçu comme positif dans une description romanesque alors qu'il est perçu comme négatif dans le domaine des assurances ou du tourisme:

> « Dans *Ghost*, les habitants du village sont vraiment effrayants! »
> « C'est effrayant de voir comment la côte est de plus en plus bétonnée. »

En général, une relation de sentiment a deux arguments: le premier est l'expression linguistique qui véhicule le sentiment en question, le deuxième est la cause ou l'objet du sentiment (si la cause est exprimée dans la phrase). Ceci donne pour la phrase

> « J'aime beaucoup Grenoble. »

la relation `SENTIMENT_POSITIF(aimer,Grenoble)`. L'attribut `POSITIF` de la relation, c'est-à-dire la valeur de sa classe, indique qu'il s'agit d'un sentiment positif dont l'objet est *Grenoble*. Dans le cas d'une phrase comme

> « Je déteste!!!!! »

la relation n'aura qu'un seul argument: `SENTIMENT_NEGATIF(détester)`, dans la mesure où l'objet du sentiment n'est pas exprimé dans la phrase.

L'objectif de la grammaire est d'extraire le plus d'informations possible dans le *thread*, en particulier les sentiments positifs et négatifs, les lieux et produits. Pour ceci, les *threads* sont analysés phrase par phrase. Chaque phrase peut contenir zéro, une ou plusieurs relations de sentiment. Il est tout à fait possible d'avoir des relations de sentiments positifs et négatifs dans une même phrase:

> « En qualité d'impression, la Epson est meilleure, en texte comme en photo, malheureusement c'est aussi la plus chère. »
> ⇒ `SENTIMENT_POSITIF(meilleur,Epson)`
> ⇒ `SENTIMENT_NEGATIF(cher,ce)`

Les parties de la grammaire qui varient selon le corpus se distinguent essentiellement par le lexique de mots qui reçoivent les traits `positif` et `négatif` correspondant aux valeurs des classes des textes. Par exemple, le lexique de la grammaire du *tourisme* contient les mots *joli* et *beau*:

> « Ce monument est vraiment *beau*. »

Pourtant, dans un corpus qui porte sur le cinéma, les livres ou les jeux vidéo, ces mêmes mots n'expriment pas toujours des sentiments. Ils ont donc été supprimés du lexique de la grammaire des *jeuxvidéo* parce qu'ils produisent trop de relations éronnées:

DEFT'07 (c.f. section 6). Sinon nous n'attribuons pas de sentiment global mais gardons les sentiments attribuées à chaque phrase.

> « Cela dépendra moins de vous que de l'imbécillité contagieuse des ennemis qui attendent sagement derrière un petit muret, leur *beau* visage buriné dépassant allègrement. »

Comme on le voit dans la phrase précédente, dans ce contexte, les mots de type *joli* ou *beau* sont utilisés pour décrire une action ou un personnage, mais pas un sentiment. La difficulté réside dans le fait de pouvoir distinguer les parties subjectives des parties objectives d'un texte. La description d'une action peut contenir des phrases avec des sentiments, donc subjectives, qui se réfèrent au déroulement de l'histoire. Cependant ces phrases devront être considérées comme étant objectives pour l'évaluation.

4.2 Lexique de sentiments

L'analyse du texte se base sur les mots du lexique qui ont reçu des traits spécifiques marquant le sentiment positif ou négatif. Il s'agit pour la plupart de verbes (*aimer*, *apprécier*, *détester*, ...) et d'adjectifs (*magnifique*, *superbe*, *insupportable*, ...), mais aussi de quelques noms communs (*plaisir*, ...) et d'adverbes (*malheureusement*, ...). Par exemple, quand une relation de modifieur du nom est extraite (*paysage magnifique*) et que le modifieur (*magnifique*) porte le trait `sents`, la relation de sentiment (⇒ `SENTIMENT_POSITIF(magnifique,paysage)`) est extraite ensuite entre le nom et son modifieur. Après cette phase d'analyse, il y a évidemment des règles plus complexes pour extraire les relations des phrases plus compliquées.

Le lexique a été défini par un linguiste au fur et à mesure de l'avancé de chaque projet. A chaque fois qu'un mot intéressant est apparu dans les textes qui n'était pas encore dans le lexique il a été ajouté à ce dernier, selon le domaine du texte. Pour chaque domaine il y a le même lexique de base et ensuite un lexique spécifique qui contient les mots du domaine en question.

L'attribut de la relation (`positif` ou `négatif`) d'un sentiment sera inversé quand une négation est présente dans la phrase, comme par exemple:

> « J'aime pas du tout les randonnées en montagne! »
> ⇒ `SENTIMENT_NEGATIF(aimer,randonnée)`
> « Ce n'est pas un mauvais restaurant. »
> ⇒ `SENTIMENT_POSITIF(mauvais,restaurant)`

Quand cela est possible, les pronoms *qui* et *que* se rapportant à une entité présente ailleurs dans la même phrase, seront remplacés par cette même entité:

> « Grenoble est une ville qui vaut vraiment le détour hiver comme été. »
> ⇒ `SENTIMENT_POSITIF(valoir,ville)`

Certains noms communs ainsi que des verbes de type interrogatif ont reçu un trait (`no-sents`) pour empêcher l'extraction de relations. Dans *Je cherche un bon hôtel.*, *Bon voyage!* ou *Bonne journée!* il ne s'agit pas de sentiments proprement dit exprimés par l'auteur du texte, mais plutôt de souhaits comme on peut les trouver surtout au début ou à la fin de messages. C'est pour cette raison que nous essayons d'éviter d'extraire ces relations.

Les noms de lieu et de produit ont également des traits spéciaux pour pouvoir extraire d'autres relations qui seront potentiellement intéressantes dans le futur. Voici un extrait du lexique où les mots reçoivent des traits en plus de ceux qu'ils portent déjà (la valeur 1 ajoute ce trait au mot, la valeur 0 l'enlève).

Chaque mot qui peut véhiculer un sentiment reçoit le trait `sents`, puis le trait `positif` ou `négatif` selon sa polarité. D'après la taxonomie d'(Ogorek, 2005) (c.f. la section suivante 4.3) sont ajoutées des valeurs de sentiment plus fines comme `à l'aise`, `détendu`, etc. Les mots qui ne doivent pas entrer en relation de sentiment reçoivent le trait `no-sents`. Les traits `genre` et `plateforme` servent à extraire d'autres relations intéressantes dans le domaine des *jeuxvidéo*.

```
Lexique:

agréable = {sents=1, positif=1, à l'aise=1}
sympathique = {sents=1, positif=1, détendu=1}
aimer = {sents=1, positif=1, enchanté=1}
conseiller = {sents=1, positif=1, conseil=1}
```

```
plaisir = {sents=1, positif=1, enchanté=1}
décevant = {sents=1, negatif=1, triste=1}
cher = {sents=1, negatif=1, cher=1}
regretter = {sents=1, negatif=1, triste=1}
malheureusement = {sents=1, negatif=1, triste=1}
appétit = {no-sents=1}
vacance = {no-sents=1}
chercher = {no-sents=1}
aventure = {genre=1}
PC = {plateforme=1}
```

La taille du lexique varie selon le domaine d'application. Le lexique de la grammaire de base des sentiments contient environ 250 mots (noms, verbes, adjectifs, etc.) avec des traits de sentiment (`positif` et `négatif`). À ce lexique de base, s'ajoutent environ 150 mots dans le domaine du *tourisme*, et environ 250 mots dans le domaine des *jeuxvidéo*.

4.3 Annotation manuelle de textes

La configuration de la grammaire générique a été faite sur la base d'un travail d'annotation manuelle (à l'aide du logiciel Protégé 3.2[6] avec le plugin Knowtator[7]) de *threads* venant du domaine du tourisme. Ce corpus du *tourisme* contient une centaine de *threads* annotés (avec comme sujet différentes régions et destinations en France). Chaque *thread* est composé de messages des utilisateurs du forum; la longueur varie entre dix et 55 messages par document. Un message peut ne contenir qu'une phrase ou plusieurs paragraphes. L'annotation de ce corpus avec Protégé et Knowtator a été faite dans la lignée des travaux de (Riloff *et al.*, 2005), (Riloff *et al.*, 2006), (Wiebe & Mihalcea, 2006).

L'annotation inclut les informations de cause/objet, d'intensité et de l'émetteur du sentiment. Dans

> « J'aime énormément Grenoble. »

aimer véhicule le sentiment, *Grenoble* est l'objet du sentiment et *je* est l'émetteur du sentiment. L'adverbe *énormément* exprime l'intensité, le sentiment ici est plus intense que dans la phrase

> « J'aime bien Grenoble. »

L'annotation pour le *tourisme* ne contient pas seulement les deux valeurs *positif* et *négatif* pour classer les sentiments, mais est détaillée beaucoup plus finement (c.f. par exemple les travaux de (Mathieu, 2000), (Mathieu, 2006)). Le schéma d'annotation choisi est même plus fin et on voit donc que la classification des sentiments que l'on propose permet un grand nombre de modalités et va au-delà de la simple opposition positif-négatif.

En effet, nous avons repris la taxonomie d'(Ogorek, 2005) qui propose 33 sentiments différents (17 positifs et 16 négatifs) auxquels nous avons ajouté les pseudo-sentiments comme *bon-marché*, *conseil*, *cher* et *avertissement*, car dans le domaine du *tourisme* il y a beaucoup de messages concernant les prix des prestations dont les auteurs des messages sont contents (ou pas).

Les sentiments de la taxonomie d'Ogorek sont classés en groupes[8] comme AMOUR-DÉSIR (*amour, envie, tendresse, désir*), JOIE (*enchanté, excité, heureux, joyeux*), TRISTESSE- DÉTRESSE (*découragé, bouleversé, démoralisé, triste*), COLÈRE-DÉGOÛT-MÉPRIS (*colère, mépris, désapprobation*), etc.

5 SYBILLE, la méthode hybride

La méthode hybride est une combinaison des deux méthodes précédentes (c.f. sections 3 et 4). La méthode statistique distributionnelle sert dans un premier temps à faciliter le développement de la méthode symbolique

[6] `http://protege.stanford.edu/`
[7] `http://bionlp.sourceforge.net/Knowtator/index.shtml`
[8] Sauf les pseudo-sentiments concernant les prix et conseils introduits par notre équipe comme *gratuit*, etc.

selon le domaine des textes. Pour chaque domaine d'application une ontologie de concepts propre est extraite. La création de lexique pour la grammaire symbolique est ainsi facilitée et accélérée.

La méthode statistique distributionnelle permet de faire une première fouille dans les textes pour obtenir les concepts du domaine des textes analysés. Ensuite, l'utilisateur qui a configuré la grammaire de la méthode symbolique peut modifier et améliorer celle-ci pour obtenir de meilleurs résultats. Le travail prend alors la forme d'un cycle où les résultats s'améliorent constamment.

L'analyse du *thread* se fait au niveau des phrases et permet d'améliorer le résultat en ajoutant ou supprimant par exemple des mots au lexique. Ceci a l'avantage de montrer exactement quelles phrases du document expriment un sentiment, les phrases objectives n'étant pas pris en compte.

C'est une approche qui permet de pouvoir extraire rapidement les concepts intéressants du domaine d'application et d'améliorer en même temps le développement de la grammaire de la méthode symbolique. Ceci permet d'intégrer les spécificités du cahier des charges, c'est-à-dire les particularités de chaque corpus (à l'aide de lexiques différents selon le domaine d'application).

La méthode hybride a été évaluée dans sa première version (c'est-à-dire avec la méthode statistique basée sur l'apprentissage automatique, avant l'intégration de l'exploration de corpus), notamment au moment du concours DEFT'07 (c.f. la section 6), avec la mesure du F-score[9]. Elle a été utilisée pour trois des quatre corpus DEFT'07 et a donné les meilleurs résultats pour les corpus *jeuxvidéo* avec un F-score de 0,71, contre 0,54 (méthode symbolique) et 0,70 (méthode statistique) et *relectures* avec un F-score de 0,54, contre 0,48 (méthode symbolique) et 0,51 (méthode statistique).[10] Pour le corpus *débats politiques* seule l'ancienne méthode statistique a été utilisée.

Une évaluation avec la nouvelle méthode statistique est prévue dès que le système a été mis au point.

6 Première évaluation

La section suivante décrit la première évaluation du système SYBILLE, faite en 2007. Depuis le changement des méthodes nous n'avons malheureusement pas encore eu l'occasion de refaire une nouvelle évaluation empirique avec des objectifs bien définis et des résultats satisfaisants. Ce sera l'objet d'une future publication.

DEFT (le DÉfi Fouille de Texte) est une campagne d'évaluation dont le thème était en 2007[11] la classification de textes d'opinion, présents dans différents types de textes. Plusieurs groupes de recherche (laboratoires universitaires ou entreprises privées) ont pu tester leurs systèmes de classification sur les mêmes textes. Dans la phase initiale, chaque groupe inscrit a reçu les deux tiers de chacun des quatre corpus différents qui avaient comme sujet des critiques de films et de livres, des tests de jeux vidéo, des relectures d'articles scientifiques et des notes de débats parlementaires. Pour les trois premiers corpus, une note à trois valeurs (positif, moyen ou négatif) a été attribuée à chaque texte par le comité des organisateurs, une note à deux valeurs seulement (positif ou négatif) pour le dernier corpus. Après un certain temps pendant lequel chaque groupe a mis au point son ou ses systèmes de classification, un troisième tiers de chaque corpus a été envoyé pour faire les tests dont les résultats ont dû être soumis quelques jours plus tard.

La grammaire de l'analyseur a été paramétrée pour répondre aux besoins des différents corpus DEFT'07, du point de vue lexical mais aussi pour résister aux fautes d'orthographes répétitives. Le point le plus important à modifier a été la classification du message entier qui peut contenir plusieurs sentiments avec une seule valeur globale, et en particulier l'introduction de la notion de sentiment moyen. Dans notre approche standard, au niveau des phrases, les sentiments sont positifs ou négatifs. Il n'est pas nécessaire d'utiliser des sentiments moyens dans le domaine du tourisme, dans la mesure où la taxonomie utilisée (c.f. section 4.3) permet de nuancer suffisamment.

Les sentiments moyens pour DEFT'07 n'ont pas été extraits à l'aide de mots dans le lexique avec un trait moyen, mais d'après des structures de phrase. Par exemple à une phrase qui contient un sentiment positif et un sentiment négatif coordonnés par *mais* est attribué un sentiment moyen à la place:

> « Ce jeu est *amusant* au début **mais** *ennuyant* la deuxième semaine. »

[9] Le F-score utilisé dans nos expériences est calculé de la manière suivante: $F_{score}(\beta) = \frac{(\beta^2+1)*Précision*Rappel}{\beta^2*Précision+Rappel}$ avec $\beta = 1$.

[10] Pour le corpus *aVoiraLire* le meilleur résultat a été obtenu par la méthode statistique avec un F-score de 0,52, contre 0,51 (méthode hybride) et 0,42 (méthode symbolique). Ce corpus n'est probablement pas assez uniforme (il parle de livres, films actuels au cinéma, disques, films plus anciens enregistrés, ...) pour pouvoir faire une liste de termes plus performante.

[11] `http://deft07.limsi.fr/`

Quelques mots clés (surtout des adverbes comme *malgré, pourtant, ...*) sont utilisés pour aider à classifier un texte qui contient des phrases avec des sentiments positifs et négatifs (c.f. les travaux de (Sándor, 2005)). Le texte entier est alors classé comme moyen.

7 L'interface graphique SYBILLE

Pour conclure cet article voici quelques figures qui présentent notre interface graphique SYBILLE, une interface qui aide l'analyste et le client à naviguer parmi les messages analysés pour en prendre connaissance. L'interface que nous utilisons est une dérivation du navigateur *Longwell* du MIT[12].

Les figures 2 et 3 suivantes montrent l'interface graphique du système SYBILLE, dans le domaine des *imprimantes*. Sur la figure 2 en haut à droite il y a un champ dans lequel l'utilisateur peut faire une recherche **(1)** de messages qui contiennent un mot de son choix; sinon, il peut ne choisir que les messages positifs ou négatifs **(2)**. Une autre façon de faire une recherche serait de se limiter aux messages qui n'évoquent qu'une marque précise **(3)**, ou en dessous un domaine d'application plus spécialisé **(4)**, ou encore un mot précis d'un domaine. On offre aussi l'option de sélectionner un forum donné parmi tous ceux qui ont été analysés. Les options de recherche peuvent être combinées à volonté pour limiter le nombre de réponses souhaitées.

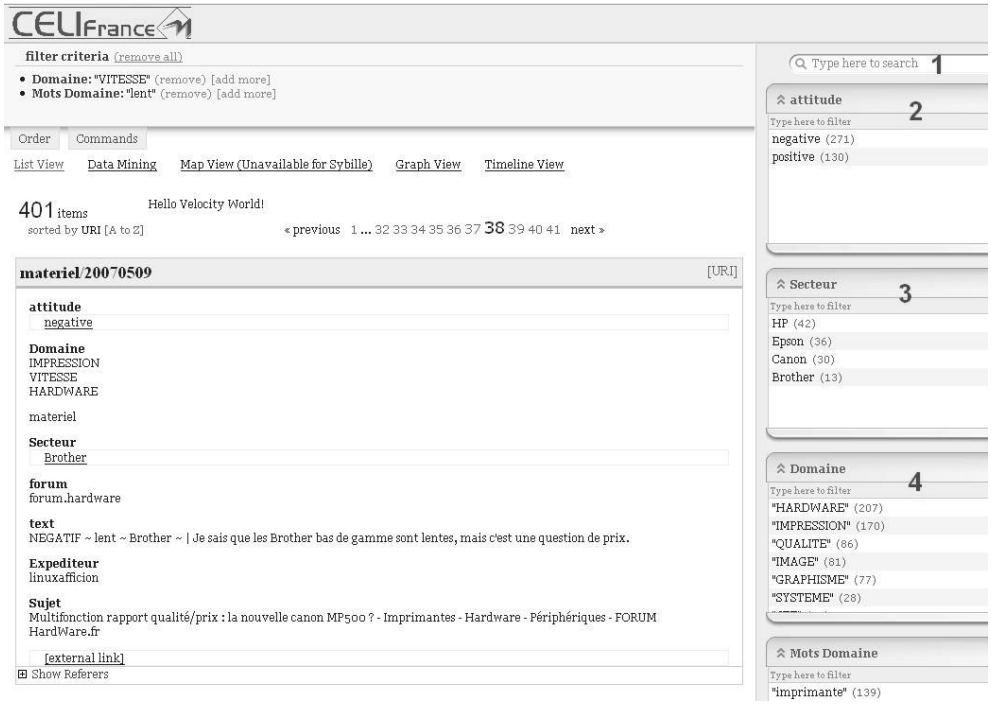

Figure 2: L'interface graphique SYBILLE, ici pour le domaine des imprimantes. Plusieurs moyens différents permettent de naviguer dans les résultats des messages analysés.

La figure 3 montre en détail la relation de sentiment qui est indiquée avec ses arguments **(5)**, la phrase qui contient le sentiment et un lien (*external link* **(6)**) vers le *thread* entier qui permet de visualiser le contexte.

[12] `http://simile.mit.edu/wiki/Longwell`

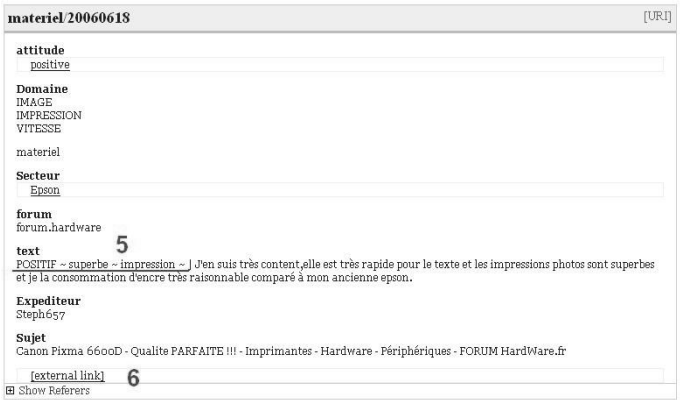

Figure 3: Exemple détaillé d'une relation de sentiment positif, toujours dans le domaine des imprimantes.

8 Conclusion

Nous avons présenté dans cet article comment l'utilisation de grammaires syntaxiques, un outil du traitement automatique du langage naturel, peut améliorer la qualité d'un système d'extraction de sentiments. Nous avons décrit une méthode symbolique avec une grammaire adaptée au domaine des textes, et une méthode statistique distributionnelle pour la création d'une ontologie des concepts du domaine des textes. Cette dernière nous offre la possibilité d'une exploration rapide d'un corpus d'un domaine donné qui souligne en même temps les attitudes différentes des différents groupes d'opinions par rapport à l'objet de l'étude.

La première évaluation de notre système de classification SYBILLE en 2007 a montré que la combinaison de la méthode symbolique et l'ancienne méthode statistique donne des résultats plus précis que chacune des méthodes employée séparément. L'intérêt de la méthode hybride repose sur la prise en compte des contextes d'application de ses résultats. Il est bien connu que la méthode purement symbolique a souvent pour le client un coût d'entrée plutôt élevé. Cette considération est liée au temps de configuration, de repérage ou de création de lexiques spécifiques, de taxonomies etc.

L'utilisation d'une méthode hybride permet, au contraire, de minimiser les coûts de configuration, en réduisant une partie du travail à l'annotation de textes, une tâche qui dans la plupart des cas peut être réalisée par le client lui-même. Les algorithmes d'apprentissage automatique sont alors en mesure de donner des premiers jugements au niveau du texte entier.

Ce qui est le plus important, c'est qu'avec ce type de système on peut ajouter, selon la méthode exposée dans cet article, une couche *symbolique* au fur et à mesure, de plus en plus importante dès que les exigences d'une application deviennent plus précises. On peut par exemple superposer une couche d'identification de jugement, qui permet d'avoir une visibilité sur les jugements sans devoir lire le texte dans son entier. On peut identifier certains patrons sémantiques qui sont d'importance capitale pour une application donnée et qui doivent avoir la priorité sur les résultats statistiques (par exemple le souci de sécurité exprimé par les internautes sur un certain modèle de voiture).

Les exemples pourraient être multipliés. Ce qui apparaît avant tout intéressant, c'est que la démarche hybride est importante non seulement pour des raisons scientifiques de performance (le meilleur résultat entre les technologies que nous avons adoptées) mais, aussi et surtout pour des raisons de développement et d'acceptation par le marché.

Références

Aït-Mokhtar S. et Chanod J.-P. (1997). Subject and object dependency extraction using finite-state transducers. In P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo & Y. Wilks, Eds., *Automatic information extraction and building of lexical semantic resources for NLP applications*, p. 71–77. Association for Computational Linguistics.

Aït-Mokhtar S., Chanod J.-P. et Roux C. (2001). A multi-input dependency parser. In *Actes d' IWPT'01*.

Baroni M. et Bisi S. (2004). Using cooccurrence statistics and the web to discover synonyms in a technical language. In *Actes de LREC'04*, p. 1725–1728.

Basili R., Pazienza M. T. et Zanzotto F. M. (1999). Lexicalizing a shallow parser. In *Actes de TALN'99*.

Bethard S., Yu H., Thornton A., Hatzivassiloglou V. et Jurafsky D. (2004). Automatic extraction of opinion propositions and their holders. In *Actes d' AAAI'04*.

Bosca A. et Dini L. (2009). Ontology based law discovery. In S. Montemagni & D. Tiscornia, Eds., *Semantic processing of legal texts*. Springer, à paraître.

Chklovski T. (2006). Deriving quantitative overviews of free text assessments on the web. In *Actes d' IUI'06*, p. 155–162.

Dini L. (2002). Compréhension multilingue et extraction de l'information. In F. Segond, Ed., *Multilinguisme et traitement de l'information (Traité des sciences et techniques de l'information)*. Editions Hermes Science.

Dini L. et Mazzini G. (2002). Opinion classification through information extraction. In A. Zanasi, C. A. Brebbia, N. F. F. Ebecken & P. Melli, Eds., *Data Mining III*, p. 299–310. WIT Press.

Dini L. et Segond F. (2007). La linguistique informatique au service des sentiments. In *Revue de l'électricité et de l'électronique*, p. 66–77. Editions SEE.

Everitt B. (1992). *The Analysis of Contingency Tables*. Chapman and Hall, 2nd edition.

Grouin C., Berthelin J.-B., El Ayari S., Heitz T., Hurault-Plantet M., Jardino M., Khalis Z. et Lastes M. (2007). Présentation de DEFT'07 (DÉfi Fouille de Textes). In *Actes de DEFT'07*, p. 1–8.

Hatzivassiloglou V. et McKeown K. R. (1997). Predicting the semantic orientation of adjectives. In *Actes d' ACL'97*, p. 174–181.

Kim S.-M. et Hovy E. (2004). Determining the sentiment of opinions. In *Actes de COLING'04*, p. 1267–1373.

Mathieu Y. Y. (2000). *Les verbes de sentiment. De l'analyse linguistique au traitement automatique*. CNRS Editions.

Mathieu Y. Y. (2006). A computational semantic lexicon of french verbs of emotion. In J. G. Shanahan, Y. Qu & J. Wiebe, Eds., *Computing attitude and affect in text*: *Theorie and applications*, p. 109–124. Springer.

Maurel S., Curtoni P. et Dini L. (2007). Classification d'opinions par méthodes symbolique, statistique et hybride. In *Actes de DEFT'07*, p. 111–117.

Maurel S., Curtoni P. et Dini L. (2008). L'analyse des sentiments dans les forums. In *Actes de FODOP'08*, p. 9–22.

Maurel S., Curtoni P. et Dini L. (2009). Extraction de sentiments et d'opinions basée sur des règles. In *Fouille des données d'opinions*. RNTI, à paraître.

Ogorek J. R. (2005). Normative picture categorization: Defining affective space in response to pictorial stimuli. In *Actes de REU'05*.

Riloff E., Patwardhan S. et Wiebe J. (2006). Feature subsumption for opinion analysis. In *Actes d' EMNLP'06*, p. 440–448.

Riloff E., Wiebe J. et Phillips W. (2005). Exploiting subjectivity classification to improve information extraction. In *Actes d' AAAI'05*.

Sándor A. (2005). A framework for detecting contextual concepts in texts. In *Actes du Electra Workshop*.

Sproull L. et Kiesler S. (1991). *Connections*: *New ways of working in the networked organization*. Cambridge: MIT Press.

Turney P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Actes d' ACL'02*.

Wiebe J. et Mihalcea R. (2006). Word sense and subjectivity. In *Actes d' ACL'06*, p. 1065–1072.

Wilson T., Wiebe J. et Hwa R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *Actes d' AAAI'04*.

Yu H. et Hatzivassiloglou V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Actes d' EMNLP'03*, p. 129–136.

Fusion probabiliste appliquée à la détection et classification d'opinions

Juan-Manuel Torres-Moreno[1,2], Marc El-Bèze[1], Fréderic Béchet[1] et Nathalie Camelin[1]

[1]Laboratoire Informatique d'Avignon – Université d'Avignon et des Pays de Vaucluse
BP 1228, 84911 Avignon Cédex 09 France
[2]École Polytechnique de Montréal – Département de génie informatique
H3C3P8 Montréal (Québec) Canada

Résumé – Abstract

Nous présentons des modèles d'apprentissage probabilistes appliqués à la tâche de classification telle que définie dans le cadre du défi DEFT'07 : la classification d'un texte suivant l'opinion qu'il exprime. Pour classer les textes, nous avons utilisé plusieurs classifieurs et une fusion. Une comparaison entre les résultats en validation et en tests montrent une coïncidence remarquable et mettent en évidence la robustesse et performances de l'algorithme de fusion. Les résultats que nous obtenons, en termes de précision, rappel et F-score sur les sous corpus de test nous ont permis de remporter le défi.
We present probabilistic learning models applied to sentiment classification task as defined in the DEFT'08 challenge. In this task, the texts must be classified following theirs opinions. We have used a mix of several classifiers. A comparison between the results and validation tests shows a remarkable coincidence and highlight the robustness and performance of our mixture algorithm. Our results, (precision, recall and F-score) on the test corpus, enabled us to win the challenge.

Mots-clefs – Keywords

Méthodes probabilistes, Apprentissage automatique, Classification de textes par leur contenu, défi DEFT.
Probabilistic methods, Machine learning, Text Classification, DEFT challenge.

1 Introduction

En juillet 2007 dans le cadre de la plate-forme AFIA 2007[1], a été organisé la troisième édition de DEFT (DÉfi Fouille de Textes) (Azé & Roche, 2005; Azé *et al.*, 2006). Cela a été la deuxième participation dans DEFT de l'équipe Traitement Automatique de la Langue Naturelle (TALNE) du Laboratoire Informatique d'Avignon (LIA)[2]. Lors de la première competition en 2005 (El-Bèze *et al.*, 2005), notre équipe avait remporté le défi. À l'époque, le problème était de classer les segments des allocutions de Jacques Chirac et François Mitterrand préalablement mélangées[3]. Le défi DEFT en 2007[4] a été motivé par le besoin de mettre en place des techniques de fouille des textes permettant de classer de textes suivant l'opinion qu'ils expriment. Concrètement, il s'agisait de classer les textes de quatre corpus en langue française selon les opinions qui y sont formulées. La classification d'un corpus en classes pré-déterminées, et son corollaire le profilage de textes, est une problématique importante du domaine de la fouille de textes. Le but d'une classification est d'attribuer une classe à un objet textuel donné, en fonction d'un profil qui sera explicité ou non suivant la méthode de classification utilisée. Les applications sont variées. Elles vont du filtrage de grands corpus (afin de faciliter la recherche d'information ou la veille scientifique et économique) à la classification par le genre de texte pour adapter les traitements linguistiques aux particularités d'un corpus. La tâche proposée par DEFT'07 visait le domaine applicatif de la prise de décision. Attribuer une classe à un texte, c'est aussi lui attribuer une valeur qui peut servir de critère dans un processus de décision. Et en effet, la classification

[1]Association Française pour l'Intelligence Artificielle, `http://afia.lri.fr`
[2]`http://www.lia.univ-avignon.fr`
[3]Pour plus de détails concernant DEFT'05, voir le site `http://www.lri.fr/ia/fdt/DEFT05`
[4]`http://deft07.limsi.fr/`

d'un texte suivant l'opinion qu'il exprime a des implications notamment en étude de marchés. Certaines entreprises veulent désormais pouvoir analyser automatiquement si l'image que leur renvoie la presse est plutôt positive ou plutôt négative. Des centaines de produits sont évalués sur Internet par des professionnels ou des internautes sur des sites dédiés : quel jugement conclusif peut tirer de cette masse d'informations un consommateur, ou bien encore l'entreprise qui fabrique ce produit ? En dehors du marketing, une autre application possible concerne les articles d'une encyclopédie collaborative sur Internet telle que Wikipédia : un article propose-t-il un jugement favorable ou défavorable, ou est-il plutôt neutre suivant en cela un principe fondateur de cette encyclopédie libre ? À priori, un travail de détection et de classification d'opinion paraît très simple. Or, de nombreuses raisons font que le problème est complexe. Facteur aggravant : on ne dispose que de corpus de taille moyenne, déséquilibrés par rapport à leurs classes. Dans cet article nous décrivons les méthodes employées dans le cadre de DEFT'07 qui nous ont permis de remporter le défi. Nous décrivons en section 2 le corpus et la méthode d'évaluation proposée. En section 4 nous présentons les outils de classification de texte utilisées. La représentation de textes ainsi qu'une agglutination et normalisation graphique sont detaillées en section 3. Nos outils de classification sont décrits en section 4. Des expériences et résultats sont rapportés et discutés en section 5, avant de conclure et d'envisager quelques perspectives.

2 Description des corpus

Les organisateurs du défi DEFT'07 ont mis à la disposition des participants quatre corpus héterogènes :

aVoiraLire. Critiques de films, livres, spectacles et bandes dessinées. Ce corpus comporte 3 460 critiques et les notes qui leur sont associées. Etant donné que beaucoup d'organes de diffusion de critiques de films ou de livres[5] attribuent, en plus du commentaire, une note sous la forme d'une icône. Les organisateurs du défi ont retenu une échelle de 3 niveaux de notes. Ceci donne lieu à 3 classes bien discriminées : 0 (mauvais), 1 (moyen), et 2 (bien).

jeuxvideo. Le corpus de tests de jeux vidéo comprend 4 231 critiques. Chaque critique comporte une analyse des différents aspects du jeu – graphisme, jouabilité, durée, son, scénario, etc. – et une synthèse globale du jugement. Comme pour le corpus précédent, a été retenue une échelle de 3 niveaux de notes, qui donne les 3 classes 0 (mauvais), 1 (moyen), et 2 (bien).

relectures. Relectures d'articles de conférences. Ce corpus comporte 1 484 relectures d'articles scientifiques qui alimentent les décisions de comités de programme de conférences et renvoient des conseils et critiques aux auteurs. L'échelle retenue comporte 3 niveaux de jugement. La classe 0 est attribuée aux relectures qui proposent un rejet de l'article, la classe 1 est attribuée aux relectures qui retrouvent l'acceptation sous condition de modifications majeures ou en séance de posters, et la classe 2 regroupe les acceptations d'articles avec ou sous des modifications mineures. Ce corpus (comme le suivant) a subi un processus préalable d'anonymisation de noms des personnes.

débats. Le corpus des débats parlementaires est composé de 28 832 interventions de députés portant sur des projets de lois examinés par l'Assemblée Nationale. À chaque intervention, est associé le vote de l'intervenant sur la loi discutée. 0 (en faveur) ou 1 (contre).

Les corpus ont été scindés par les organisateurs en deux parties : une partie (environ 60%) des données a été fournie aux participants comme données d'apprentissage afin de mettre au point leurs méthodes, et une autre partie (environ 40%) a été réservée pour les tests proprement dits. Sous peine de disqualification, aucune donnée, en dehors de celles fournies par le comité d'organisation ne pouvait être utilisée. Ceci exclut notamment l'accès aux sites web ou à n'importe quelle autre source d'information. Nous présentons au tableau 1, des statistiques brutes (nombre de textes et nombre de mots) des différents corpus. Des exemples portant sur la structure et les détails des corpus, peuvent être consultés dans le site du défi[6].

2.1 Évaluation stricte

Le but du défi a consisté à classer chaque texte, issu des quatre corpus, selon l'avis qui y est exprimé. Positif, négatif ou neutre dans le cas où il y a trois classes, pour ou contre dans le cas binaire (corpus de débats parlemen-

[5] Par exemple voir le site `http://www.avoir-alire.com`

[6] `http://deft07.limsi.fr/corpus-desc.php`

Corpus	Textes (A)	Mots (A)	Textes (T)	Mots (T)
aVoiraLire	2 074	490 805	1 386	319 788
jeuxvideo	2 537	1 866 828	1 694	1 223 220
relectures	881	132 083	603	90 979
débats	17 299	2 181 549	11 533	1 383 786

Table 1: Statistiques brutes sur les quatre corpus d'apprentissage (A) et de test (T).

taires). Intuitivement, la tâche de classer les avis d'opinion des articles scientifiques est la plus difficile des quatre car le corpus afférent contient beaucoup moins d'informations que les trois autres, mais d'autres caractéristiques particulières à chaque corpus ont aussi leur importance. Les algorithmes seront évalués sur des corpus de test (T) avec des caractéristiques semblables à celui d'apprentissage (A) (cf. tableau 1), en calculant le *Fscore* des documents bien classés, moyenné sur tous les corpus :

$$Fscore(\beta) = \frac{(\beta^2 + 1) \times \langle Précision \rangle \times \langle Rappel \rangle}{\beta^2 \times \langle Précision \rangle + \langle Rappel \rangle} \tag{1}$$

où la précision moyenne et le rappel moyen sont calculés comme :

$$\langle Précision \rangle = \frac{\sum_{i=1}^{n} Précision_i}{n} \; ; \; \langle Rappel \rangle = \frac{\sum_{i=1}^{n} Rappel_i}{n} \tag{2}$$

Etant donné pour chaque classe i :

$$Précision_i = \frac{\{\text{Nb de documents correctement attribués à la classe } i\}}{\{\text{Nb de documents attribués à la classe } i\}} \tag{3}$$

$$Rappel_i = \frac{\{\text{Nb de documents correctement attribués à la classe } i\}}{\{\text{Nb de documents appartenant à la classe } i\}} \tag{4}$$

D'après les règles du défi, un document est attribué à la classe d'opinion i si : i/ seule la classe i a été attribuée à ce document, sans indice de confiance spécifié ; ii/ la classe i a été attribuée à ce document avec un meilleur indice de confiance que les autres classes (s'il existe un indice de confiance).

2.2 Indice de confiance pondéré

Un système de classification automatique peut attribuer à un document une distribution de probabilité sur les différentes classes au lieu de lui attribuer une seule classe. L'indice de confiance est cette probabilité pour un document d'appartenir à une classe d'opinion donnée. Le F-score pondéré par l'indice de confiance a été utilisé, à titre indicatif, pour des comparaisons complémentaires entre les méthodes mises en place par les équipes. Dans le F-score pondéré, la précision et le rappel pour chaque classe ont été pondérés par l'indice de confiance.

$$Précision_i = \frac{\sum_{\text{AttribuéCorrect}_i=1}^{\text{NbAttribuéCorrect}_i} \text{Indice\_confiance}_{\text{AttribuéCorrect}_i}}{\sum_{\text{Attribué}_i=1}^{\text{NbAttribué}_i} \text{Indice\_confiance}_{\text{Attribué}_i}} \tag{5}$$

$$Rappel_i = \frac{\sum_{\text{AttribuéCorrect}_i=1}^{\text{NbAttribuéCorrect}_i} \text{Indice\_confiance}_{\text{AttribuéCorrect}_i}}{\{\text{Nb de documents correctement attribués à la classe } i\}} \tag{6}$$

avec :

- $\text{NbreAttribuéCorrect}_i$: nombre de documents appartenant effectivement à la classe i et auxquels le système a attribué un indice de confiance non nul pour cette classe.
- NbreAttribué_i : nombre de documents attribués auxquels le système a attribué un indice de confiance non nul pour la classe i.

Dans le cadre de DEFT'07, le calcul du F-score retenu par les organisateurs est ensuite calculé à l'aide des formules (1) et (2), du F-score classique modifié (cette réécriture suppose évidemment que β soit égal à 1 de façon à ne privilégier ni précision ni rappel).

3 Représentations de documents

Un même texte peut être représenté par les différents paramètres qu'il est possible d'en extraire. Les représentations les plus courantes sont les mots, les étiquettes morpho-syntaxiques –*Part Of Speech*, POS– ou les lemmes. La tâche qui nous occupe consiste à retrouver l'*opinion* exprimée dans les textes. En nous inspirant de l'approche typique de l'analyse des opinions (Hatzivassiloglou & McKeown, 1997), nous utilisons un paramètre de représentation supplémentaire, une étiquette nommée *seed*. Un *seed* est un mot susceptible d'exprimer une polarité positive ou négative (Wilson *et al.*, 2005). Notre protocole de construction du lexique de *seeds* consiste en deux étapes. Premièrement, une liste de mots polarisés a été créée manuellement. Elle contient par exemple : *aberrant, compliments, discourtois, embêtement, ...*. Afin de généraliser la liste de mots polarisés obtenue, chaque mot a été remplacé par son lemme. Nous obtenons ainsi un premier lexique de 565 *seeds*. Deuxièmement, l'algorithme *BoosTexter* a appris sur les textes représentés en mots. Les mots sélectionnés par ce modèle ont été filtrés manuellement, lemmatisés et ajoutés au lexique. Au final, nous obtenons un lexique d'environ 2 000 *seeds*. En effet, une phrase représentée en *seeds* ne contient alors que les lemmes faisant partie de ce lexique.

3.1 Agglutination et normalisation graphique

Ce qui est mis en œuvre dans cette phase pourrait être vu comme une simple étape préalable au cours de laquelle sont appliquées des règles de réécritures pour regrouper les mots[7] en unités de base. Un autre ensemble de règles appropriées est mis à contribution pour normaliser les graphies. Pour rester indépendant de la langue et de la tâche, nous n'avons pas souhaité demander à des experts de produire ces deux ensembles de règles. Le recours à une étape de prétraitement comme la lemmatisation est motivé par le taux de flexion élevé de la langue française. Néanmoins, dans le problème qui nous occupe, il s'avère utile de ne pas voir disparaître nombre d'informations comme par exemple certains conditionnels ou subjonctifs. Dans une relecture d'article, la présence de propositions comme " *Il aurait été préférable* " ou " *il eût été préférable* " laisse supposer que l'arbitre n'est pas totalement en faveur de l'acceptation du texte qu'il a relu. Pour ne en être privés, nous avons bridé la lemmatisation pour un petit nombre de cas susceptibles de servir de points d'appui lors de la prise de décision. Pour au moins deux systèmes, les textes lemmatisés ont été soumis à une étape que l'on pourrait qualifier de normalisation graphique. Quelque 30 000 règles écrites pour l'occasion ont permis de réunifier les variantes graphiques (essentiellement des noms propres) et de corriger un grand nombre de coquilles. Il est à noter que certaines de ces fautes d'orthographe ont pu être introduites par l'étape de réaccentuation automatique que nous avons appliquée au préalable sur les quatre corpus. En cas d'ambiguïté, ces récritures sont faites en s'appuyant sur les contextes gauches ou droits (parfois les deux). Par exemple : *Thé-Old-Republic* $\Rightarrow$ *the-Old-Republic*. Ces règles de réécriture avaient aussi pour but de combler certaines lacunes de notre lemmatiseur. Il n'est pas inutile de ramener à leur racine des flexions même peu fréquentes de verbes qui ne se trouvaient pas dans notre dictionnaire (comme *frustrer, gâcher*, ou *gonfler*). Enfin, quelques règles (peu nombreuses) avaient pour mission d'unifier sous une même graphie des variantes sémantiques (par exemple : *tirer-balle-tempe* et *tirer-balle-tête*).

Les différents exemples donnés ci-dessus font apparaître des regroupements sous la forme d'expressions plus ou moins figées[8]. Celles-ci ont été constituées par application de règles régulières portant sur des couples de mots. Pour leur plus grande partie, les 30 000 règles que nous avons utilisées proviennent d'un simple calcul de collocation effectué selon la méthode du rapport de vraisemblance (Mani & Maybury, 1999). Une autre part non négligeable est issue de listes d'expressions disponibles sur la toile[9]. Nous y avons ajouté également des proverbes (comme *tirer-son-épingle-jeu, mettre-feu-poudre*) extraits de listes se trouvant sur des sites web[10]. Mais nous sommes conscients que même si nous avons tenté de contrôler au maximum ces ajouts, des expressions comme "*les pieds sur terre* " ou "*un pied à terre* " ont pu être fondues à tort dans une même graphie **pied-terre**. Enfin d'autres expressions proches des slogans martelés lors de campagnes électorales de 2007, (comme *travailler-plus-pour-gagner-plus* ou *ordre-juste*) nous ont été fournies, à l'époque, par une actualité plus brûlante. Pour DEFT'08 nous avons changé la forme de cette agglutination/normalisation (Béchet *et al.*, 2008). L'objectif était de faire émerger, de façon automatique, ces règles à partir des textes.[11]

[7] Il serait plus correct de dire leurs lemmes car nous utilisons les formes lemmatisées par *LIA_TAGG*

[8] Pour l'identification de plusieurs noms propres (noms de jeux vidéo et vedettes du show-bizz) les étudiants et les enfants de l'un des co-auteurs de cet article ont été mis à contribution. Qu'ils en soient ici remerciés.

[9] Comme celle qui se trouve à l'adresse `http://www.linternaute.com/expression/recherche`

[10] Comme `http://www.proverbes.free.fr/rechprov.php`

[11] Nous avons choisi de prendre appui sur le contexte, les classes, et une mesure numérique. Deux termes consécutifs ne sont "collés" que si le pouvoir discriminant (par exemple, le critère de pureté de Gini) de l'agglutination qui en résulte est supérieur à celui de chacun de ces composants, et si la fréquence d'apparition est supérieure à un certain seuil. Le principe est le même pour les règles de réécriture dont la

4 Outils de classification

Les outils de classification de texte peuvent se différencier par la méthode de classification utilisée et par les éléments choisis afin de représenter l'information textuelle (mot, étiquette POS, lemmes, stemmes, sac de mots, sac de n-grammes, longueur de phrase, etc.). Parce qu'il n'y a pas de méthode générique ayant donné la preuve de sa supériorité (dans toutes les tâches de classification d'information textuelle), nous avons décidé d'utiliser une combinaison de différents classifieurs et de différents éléments de texte. Cette approche nous permet, en outre, d'en déduire facilement les mesures de confiance sur les hypothèses produites lors de l'étiquetage. Neuf systèmes de décision ont été implantés en utilisant les classifieurs présentés ci-bas et les différentes représentations présentées dans la section 3. Ainsi, il s'agit d'obtenir des *avis différents* sur l'étiquetage d'un texte. En outre, le but n'est pas d'optimiser le résultat de chaque classifieur indépendamment mais de les utiliser comme des outils dans leur paramétrage par défaut et d'approcher l'optimum pour la fusion de leurs résultats. Parce que ces outils sont basés sur des algorithmes de classification différents avec des formats d'entrée différents, ils n'utilisent pas les mêmes éléments d'information afin de caractériser un concept. Une combinaison de plusieurs classifieurs utilisant différentes sources d'information en entrée peut permettre d'obtenir des résultats plus fiables, évaluée par des mesures de confiance basées sur les scores donnés par les classifieurs. Nous ferons ensuite une présentation brève des classifieurs utilisés.

4.1 LIA_SCT

LIA_SCT (Béchet *et al.*, 2000) est un classifieur basé sur les arbres de décisions sémantiques (*SCT-Semantic Classification Tree* (Kuhn & De Mori, 1995)). Il suit le principe d'un arbre de décision: à chaque nœud de l'arbre une question est posée qui subdivise l'ensemble de classification dans les nœuds fils jusqu'à la répartition finale de tous les éléments dans les feuilles de l'arbre. La nouveauté des SCT réside dans la construction des questions qui se fait à partir d'un ensemble d'expressions régulières basées sur une séquence de composants. Leur ordre dans le vecteur d'entrée a donc une importance. De plus, chaque composant peut se définir suivant différents niveaux d'abstraction (mots et POS par exemple) et d'autres paramètres plus globaux peuvent également intégrer le vecteur (nombre de mots du document par exemple). Lorsque l'arbre est construit, il prend des décisions sur la base de règles de classification statistique apprises sur ces expressions régulières. Lorsqu'un texte est classé dans une feuille, il est alors associé aux hypothèses conceptuelles de cette feuille selon leur probabilité. Dans LIA_SCT les textes sont représentés en lemmes.

4.2 BoosTexter

BoosTexter (Schapire & Singer, 2000) est un classifieur à large marge basé sur l'algorithme de boosting : *Adaboost* (Freund & Schapire, 1996). Le but de cet algorithme est d'améliorer la précision des règles de classification en combinant plusieurs hypothèses dites *faibles* ou peu précises. Une hypothèse faible est obtenue à chaque itération de l'algorithme de boosting qui travaille en re-pondérant de façon répétitive les exemples dans le jeu d'entraînement et en ré-exécutant l'algorithme d'apprentissage précisément sur ces données re-pondérées. Cela permet au système d'apprentissage faible de se concentrer sur les exemples les plus compliqués (ou problématiques). L'algorithme de *boosting* obtient ainsi un ensemble d'hypothèses faibles qui sont ensuite combinées en une seule règle de classification qui est un vote pondéré des hypothèses faibles et qui permet d'obtenir un score final pour chaque constituant de la liste des concepts. Les composants du vecteur d'entrée sont passés selon la technique du sac de mots (l'ordre des mots est irrelevant) et les éléments choisis par les classifieurs simples sont alors des n-grammes sur ces composants. Quatre de nos systèmes utilisent le classifieur *BoosTexter* :

- Système LIA_BOOST_BASELINE : la représentation d'un document se fait en mots. BoosTexter est appliqué en mode 3-grammes ;
- Système LIA_BOOST_BASESEED : chaque document est représenté en seeds, chaque seed est pondéré par son nombre d'occurrences, en mode uni-gramme ;

vocation est soit de corriger d'éventuelles coquilles, soit de généraliser une expression (par exemple remplacer les noms des mois par une entité abstraite `MOIS`). Nous avons proposé en DEFT'08 une modélisation plus élaborée qui apporte une réponse à la question : comment, au moyen des opérateurs de concaténation et d'alternance, inférer des automates probabilistes à partir d'un corpus étiqueté ? À l'issue d'une cinquantaine d'itérations nous avons produit automatiquement entre 15 000 et 20 000 règles de réécriture et entre 25 000 et 70 000 règles d'agglutination. Ces nombres permettent d'imaginer le temps et l'expertise nécessaires si nous avions dû produire manuellement ces règles.

- Système LIA_BOOST_SEED : chaque document est représenté par les mots et également par les *seeds* toujours pondérés par leur nombre d'occurrences, en mode uni-grammes ;
- Système LIA_BOOST_CHUNK : L'outil *LIA-TAGG*[12] est utilisé pour découper le document en un ensemble de syntagmes lemmatisés. Chaque syntagme contenant un *seed* ainsi que le syntagme précédent et suivant sont retenus comme représentation. Les autres syntagmes sont rejetés de la représentation du document. *BoosTexter* est appliqué en mode 3-grammes sur cette représentation.

4.3 SVM Nath_Torch

SVMTorch (Collobert et al.*, 2002)* est un classifieur basé sur les machines à support vectoriel (*Support Vector Machines –SVM–*) proposées par Vapnik (Vapnik, 1982; Vapnik, 1995). Les *SVM* permettent de construire un classifieur à valeurs réelles qui découpe le problème de classification en deux sous-problèmes : transformation non-linéaire des entrées et choix d'une séparation linéaire *optimale*. Les données sont d'abord projetées dans un espace de grande dimension où elles sont linéairement séparables selon une transformation basée sur un noyau linéaire, polynomial ou gaussien. Puis dans cet espace transformé, les classes sont séparées par des classifieurs linéaires qui déterminent un hyperplan séparant correctement toutes les données et maximisant la *marge*, la distance du point le plus proche à l'hyperplan. Elles offrent, en particulier, une bonne approximation du principe de minimisation du risque structurel (*i.e*: trouver une hypothèse h pour laquelle la probabilité que h soit fausse sur un exemple non-vu et extrait aléatoirement du corpus de test soit minimale). Dans nos expériences, la technique la plus simple du sac de mots est utilisée: un document est représenté comme un vecteur dont chaque composante correspond à une entrée du lexique de l'application et chaque composante a pour valeur le nombre d'occurrences de l'entrée lexicale correspondant dans le texte. Le système (LIA_NATH_TORCH) est obtenu avec *SVMTorch*. Le vecteur d'entrée est représenté par le lexique des *seeds*.

4.4 Timble

Timble (Daelemans et al.*, 2004)* est un classifieur implémentant plusieurs techniques de *Memory-Based Learning –MBL–*. Ces techniques, descendantes directes de l'approche classique des k-plus-proches-voisins (*K Nearest Neighbor k-NN*) appliquée à la classification, ont prouvé leur efficacité dans un large nombre de tâches de traitement du langage naturel. Le paramétrage par défaut de *TiMBL* est un algorithme *MLB* qui construit une base de données d'*instances de base* lors de la phase d'entraînement. Comme pour *SVM-Torch*, une instance est un vecteur de taille fixe dont les composantes sont les entrées du lexique ayant pour valeur le nombre d'occurrences dans le document. À cela s'ajoute une composante indiquant quelle est la classe à associer à ce vecteur de paires { caractéristique-valeur }. Lorsque la base de données est construite, une nouvelle instance est classée par comparaison avec toutes les instances existantes dans la base, en calculant la distance de celle-ci par rapport à chaque instance en mémoire. Par défaut, *TiMBL* résout l'algorithme *1-NN* avec la métrique *Overlap Metric* qui compte simplement le nombre de composantes ayant une valeur différente dans chacun des 2 vecteurs comparés. Cette métrique est améliorée par l'*Information Gain –IG–* introduit par (Quinlan, 1986; Quinlan, 1993) qui permet de mesurer la pertinence de chaque composante du vecteur. Le système LIA_TIMBLE est formé de l'outil *TiMBL* appliqué sur les *seeds*.

4.5 Modélisation probabiliste uni-lemme et familles de mots

Nous avons voulu simplifier au maximum un classifieur et savoir si les modèles n-grammes avec $n > 1$ apportent vraimment des éléments discriminants. Nous avons décidé d'implanter un classifieur incorporant des techniques élementaires sur les n-lemmes. Ces techniques, descendantes directes de l'approche probabiliste (Mani & Maybury, 1999) appliquées à la classification de texte, ont prouvé leur efficacité dans le défi précédent (El-Bèze *et al.*, 2005).

Les textes ont été filtrés légèrement (afin de garder notamment des petites tournures comme la voix passive, les formes interrogatives ou exclamatives), un processus d'agregation de mots composés, puis regroupés dans des mots de la même famille (via un dictionnaire d'environ 300 000 formes). Ce processus comporte un regroupement et lemmatisation particulièrs. Ainsi, des mots tels que : *chantaient*, *chant*, *chantons*, et même *chanteurs* et *chanteuses*

[12] `http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html`

seront ramenés au lemme **chanter**, ce qui diffère d'une lemmatisation classique. Nous avons limité notre modèle à $n = 1$, soit des uni-lemmes, ce qui nous évite de calculer beaucoup de coefficients de lissage. Nous avons transformé donc chaque document en un sac d'uni-lemmes. Puis nous avons calculé la classe d'appartenance d'un document comme :

$$P_t(w) \approx \prod_i \lambda_1 P_t(w_i) + \lambda_0 U_0 \tag{7}$$

Nous avons appliqué ce modèle d'uni-lemmes à tous les corpus, sans faire d'autres traitements particulières.

4.6 Modélisation sélon la théorie de l'information

Nous avons envisagé ici de recourir à une modélisation somme toute classique en théorie de l'information, tout en cherchant à y intégrer quelques unes des spécificités du problème. La formulation que nous avons retenue initialement se rapproche de celle que nous avions employée lors d'un précédent DEFT (El-Bèze *et al.*, 2005).

$$\widetilde{t} = Arg_t \max P(t) \times P(w|t) = Arg_t \max P(t) \times P_t(w) \tag{8}$$

L'étiquette t pouvant prendre ses valeurs dans un ensemble de cardinal réduit à 2 ou 3 éléments [0-1] ou [0-2], a priori le problème pourrait paraître simple, et la quantité des données fournies suffisante pour bien apprendre les modèles. Même si le vocabulaire propre aux différents corpus n'est pas si grand (entre 9ă000 mots différents pour le plus petit corpus et 50ă000 pour le plus grand), il reste que certaines entrées sont assez peu représentées. Aussi dans la lignée de ce qui se fait habituellement pour calculer la valeur du second terme de l'équation 8 nous avons opté pour un lissage de modèles n-lemmes (n allant de 0 à 3).

$$P_t(w) \approx \prod_i \lambda_3 P_t(w_i|w_{i-2}w_{i-1}) + \lambda_2 P_t(w_i|w_{i-1}) + \lambda_1 P_t(w_i) + \lambda_0 U_0 \tag{9}$$

L'originalité de la modélisation que nous nous sommes proposés d'employer dans le cadre de DEFT'07 réside essentiellement dans les aspects discriminants du modèle. Par manque de place, il ne nous est pas possible de détailler ici les différentes caractéristiques de cette nouvelle approche. Cela sera fait lors d'une publication ultérieure. Mais nous pouvons en dire au moins quelques mots. Lors de l'apprentissage, les comptes des n-lemmes sont rééchelonnés en proportion de leur pouvoir discriminant. Ce dernier est estimé selon un point de vue complémentaire au critère d'impureté de Gini selon la formule suivante.

$$G(w,h) \approx \sum_i P_t^2(t|w,h) \tag{10}$$

Les entrées w et leurs contextes gauches h qui ne sont apparus qu'avec une étiquette donnée t et pas une autre, ont un pouvoir discriminant égal à 1. Ce critère a été lissé avec un sous-critère G' permettant de favoriser (certes dans une moindre mesure que G) les couples (w,h) qui n'apparaissent que dans 2 étiquettes sur 3. Notons tout d'abord que l'emploi de tels critères discriminants est une façon de pallier le fait que l'apprentissage par recherche d'un maximum de vraisemblance ne correspond pas vraiment aux données du problème. Deuxièmement, il est aisé de comprendre combien un regroupement massif des entrées lexicales par le biais des collocations (cf. section 3) peut avoir un effet déterminant sur le nombre des événements à coefficient discriminant élevé. Ces deux remarques visent à souligner que sur ce point particulier le fameux croissement entre methode symbolique et numérique a son mot à dire. En dernier lieu, nous avons aussi adapté le calcul du premier terme $P(t)$ de l'équation 8 en combinant la fréquence relative de l'étiquette t avec la probabilité de cette même étiquette sachant la longueur du texte traité. Pour cela, nous avons eu recours à la loi Normale.

5 Résultats et discussion

5.1 Validation croisée

Afin de tester nos méthodes et de règler leurs paramètres, nous avons scindé l'ensemble d'apprentissage (A) de chaque corpus en cinq sous-ensembles approximativement de la même taille (en nombre de textes à traiter). La méthode suivie pour l'apprentissage et le réglage des paramètres de classifieurs est celle de la validation croisée en 5 sous-ensembles (*5-fold cross validation*). Le principe général de la validation croisée est le suivant:

- Diviser toutes les données D disponibles en k groupes $D = G_1, \ldots, G_k$;
- *erreur* = 0;
- Pour i allant de 1 à k
  - *E_test* = G_i ; *E_train* = $D - G_i$;
  - apprentissage du modèle M sur *E_train*;
  - *erreur* $+=$ évaluation de M sur *E_test*;

À l'issue de k itérations, *Erreur* contient l'évaluation de la méthode de classification sur l'ensemble des données disponibles. En minimisant cette quantité lors du développement et de la mise au point des différents classifieurs, l'avantage nous est donné d'avoir testé ces méthodes sur l'ensemble des données disponibles, tout en ayant limité le risque de sur-apprentissage. Pour chaque tâche du défi, nous avons segmenté le corpus d'apprentissage en 5 sous-ensembles. Nous allons présenter nos résultats en deux items : d'abord ceux obtenus sur les ensembles de devéloppement (D) et de validation (V) où nous avons paramétré nos systèmes, et ensuite les résultats sur les données de test (T) en appliquant les algorithmes.

5.2 Évaluation sur les corpus de devéloppement (D) et de validation (V)

Le decoupage des corpus en cinq sous-ensembles de devéloppement (D) est le fruit d'un tirage aléatoire. Ce découpage permet, selon nous, d'éviter de régler les algorithmes sur un seul ensemble d'apprentissage (et un autre seul de test), ce qui pourrait conduire à deux travers, le biais expérimental et/ou le phenomène de surapprentissage. Nous présentons aux tableaux 2 (**aVoiraLire**), 3 (**jeuxvideo**), 4 (**relectures**) et 5 (**débats**) des statistiques des sous-ensembles de devéloppement (T) et de validation (V) en fonction de leurs classes pour chacun des corpus.

Corpus **aVoiraLire**							
	Total	Classe 0		Classe 1		Classe 2	
Ensembles (D)	Textes	Textes	%	Textes	%	Textes	%
1	1 660	231	13,92	486	29,27	943	56,81
2	1 659	249	15,01	494	29,78	916	55,21
3	1 659	244	14,71	498	30,02	917	55,27
4	1 659	248	14,95	490	29,54	921	55,51
5	1 659	264	15,91	492	29,66	903	54,43
Ensembles (V)	textes	Textes	%	Textes	%	Textes	%
1	414	45	10,84	123	29,64	247	59,52
2	415	61	14,70	123	30,12	229	55,18
3	415	65	15,66	117	28,19	233	56,14
4	415	60	14,46	121	29,16	234	56,39
5	415	78	18,84	129	31,16	207	50,00

Table 2: Statistiques par classe sur les ensembles de devéloppement (D) et de validation (V), **aVoiraLire**.

Sur la figure 1, nous montrons le F-score du système de fusion sur les quatre corpus (V). L'apprentissage a été réalisé sur les ensembles de devéloppement et le F-score a été calculé sur les cinq ensembles de validation (V). On peut constater que le corpus de relectures d'articles scientifiques est le plus difficile à traiter. En effet, ce corpus comporte le plus petit nombre de textes (environ 704 en devéloppement et 177 en validation). Il est aussi très dur à classer étant donnée des particularités propres à ce corpus que nos avons détecté : les arbitres corrigent souvent le texte des articles à la volée (directement dans leurs commentaires), ce qui est une introduction de bruit. Nous y reviendrons lors de la discussion de nos résultats.

5.3 Évaluation sur les corpus de test

Nous avons défini l'ensemble d'apprentissage $\{A_j\} = \{D_j\} \cup \{V_j\}$; $j =$**{aVoiraLire**, **jeuxvideo**, **relectures**, **débats**}. Le tableau 7 montre les statistiques par classe pour les quatre corpus de test (T) et d'apprentissage (A). On peut constater que la distribution des données en apprentissage et en test est très homogène, ce qui en principe, facilite la tâche de n'importe quel classifieur.

Corpus **jeuxvideo**							
	Total	Classe 0		Classe 1		Classe 2	
Ensembles (D)	textes	Textes	%	Textes	%	Textes	%
1	2 032	412	20,28	917	45,13	703	34,59
2	2 029	395	19,47	905	44,60	729	35,93
3	2 029	350	17,25	951	46,87	728	35,88
4	2 029	467	23,02	946	46,62	616	30,36
5	2 029	364	17,94	945	46,57	720	35,48
Ensembles (V)	textes	Textes	%	Textes	%	Textes	%
1	505	133	26,18	221	43,50	154	30,31
2	508	30	5,91	220	43,31	258	50,79
3	508	147	28,94	215	42,32	146	28,74
4	508	102	20,08	261	51,38	145	28,54
5	508	85	16,83	249	49,31	171	33,86

Table 3: Statistiques par classe sur les ensembles de devéloppement (D) et de validation (V), **jeuxvideo**.

Corpus **relectures**							
	Total	Classe 0		Classe 1		Classe 2	
Ensembles (D)	textes	Textes	%	Textes	%	Textes	%
1	708	151	21,33	262	37,01	295	41,67
2	704	179	25,43	208	29,55	317	45,03
3	704	184	26,14	200	28,41	320	45,46
4	704	206	29,26	214	30,40	284	40,34
5	704	188	26,70	228	32,39	288	40,91
Ensembles (V)	textes	Textes	%	Textes	%	Textes	%
1	173	39	22,03	50	28,25	88	49,72
2	177	21	11,86	64	36,16	92	51,98
3	177	43	24,29	78	44,07	56	31,64
4	177	48	27,12	70	39,55	59	33,33
5	177	76	43,93	16	9,25	81	46,82

Table 4: Statistiques par classe sur les ensembles de devéloppement (D) et de validation (V), **relectures**.

La figure 2 montre les performances en F-score de chacun de nos classifieurs, ainsi que leurs moyennes sur les quatre ensembles de test. On constate que les classifieurs LIA_TIMBLE et LIA_SCT ont les perfomances les plus basses.

La figure 3 illustre les performances en F-score d'une fusion *incrémentale* des méthodes ajoutées. Cependant, l'ordre affiché n'a strictement aucun impact dans la fusion finale : il a été choisi uniquement pour mieux illustrer les résultats. On peut voir que nos résultats se placent bien au dessus de la moyenne des équipes participantes dans le défi DEFT'07, tous corpus confondus.

Sur la figure 4 nous montrons une comparaison du F-score de l'ensemble de validation (V) vs. celui de test (T), sur les quatre corpus. On peut constater la remarquable coïncidence entre les deux, ce qui signifie que notre stratégie d'apprentissage et de validation sur cinq sous-ensembles et de fusion de plusieurs classifieurs a bien fonctionné.

5.4 Discussion

Nous avons constaté que l'utilisation de collocations et réécriture (cf. section 3.1) permet d'augmenter les perfomances des méthodes. Par exemple, avec la méthode probabiliste à base d'uni-lemmes sur le corpus de validation nous sommes passés de 1 285 à 1 310 bien classés (F=57,41 → 58,89) dans le corpus **aVoiraLire**, de 1801 à 1916 (F=70,77 → 75,15) en **jeuxvideo**, de 445 à 455 (F=48,36 → 49,53) en **relectures** et de 10 364 à 11 893 (F=62,21 → 67,12) en **débats**. Dans le corpus de test les gains sont aussi non négligeables. Nous sommes passés en **aVoirAlire** de 863 à 860 (F=57,40 → 56.32); de 7530 à 7635 (F=65,82 → 66,88) en **débats** ; de 1169 à 1205 (F=69,51 → 71,48) en **jeuxvideo** et de 317 à 313 (F=51,81 → 52,04) en **relectures**. Ceci confirme l'hypothèse

Corpus **débats**					
	Total	Classe 0		Classe 1	
Ensembles (D)	textes	Textes	%	Textes	%
1	13 840	7 893	57,03	5 947	42,97
2	13 839	8 525	61,60	5 314	38,40
3	13 839	8 710	62,94	5 129	37,06
4	13 839	7 587	54,82	4 890	35,33
5	13 839	7 913	57,18	5 926	42,82
Ensembles (V)	textes	Textes	%	Textes	%
1	3 459	2 487	71,88	973	28,12
2	3 460	1 841	53,21	1 619	46,79
3	3 460	1 690	48,84	1 770	51,16
4	3 460	1 875	58,39	1 585	56,54
5	3 460	2 507	72,48	952	27,52

Table 5: Statistiques par classe sur les ensembles de devéloppement (D) et de validation (V), **débats**.

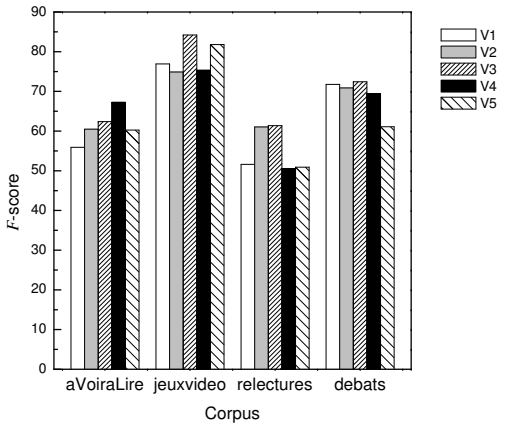

Figure 1: F-score obtenu par l'algorithme de fusion sur les cinq ensembles de validation (V). Nous affichons des résultats regroupés par corpus.

que la réécriture aide à mieux capturer la polarité des avis.

Nous avons realisé une analyse *post-mortem* de nos résultats. Nous présentons ci-bas, quelques exemples de notices qui ont été mal classés par nos systèmes. Nous avons delibérement gardé les notices dans leur état : majuscules mal placés et même avec les fautes d'ortographe ou de grammaire. En particulier, nous avons décidé de montrer majoritairement, des avis d'opinion venant du corpus de relectures d'articles scientifiques, corpus qui avait posé plus de difficultés aux algorithmes (F-score plus faible) que les autres. Par exemple, considérez la notice 3:36 (**relectures**) :

> ***3:36 relectures***
> *L'idée d'appliquer les méthodes de classification pour définir des classes homogènes de pages web est assez originale par contre, la méthodologie appliquée est classique. Je recommande donc un « weak accept » pour cet article.*

Nos systèmes l'ont classé 1 (accepté avec des modifications majeures), et après une lecture directe, on pourrait effectivement en déduire que la classe est 1 alors que la référence est 2 (accepté).

Notice 3:2 (**relectures**). L'article a été accepté mais notre système le classe comme rejeté. Il comporte beaucoup d'expressions négatives comme : " *parties de l'article me paraissent déséquilibrées* ", " *Le travail me paraît inachevé*

Corpus	Précision	Rappel	F-score	Correctes	Total
aVoiraLire (V)	0,6419	0,5678	0,6026	1 385	2 074
jeuxvideo (V)	**0,8005**	**0,7730**	**0,7865**	2 005	2 537
relectures (V)	0,5586	0,5452	0,5518	510	881
débats (V)	0,7265	0,7079	0,7171	12 761	17 299

Table 6: Précision, Rappel et F-score obtenus par notre méthode de fusion, sur les corpus de validation (V).

Corpus de test	Total textes	Classe 0 Textes	Classe 0 %	Classe 1 Textes	Classe 1 %	Classe 2 Textes	Classe 2 %
aVoiraLire (T)	1 386	207	14,94	411	29,65	768	55,41
jeuxvideo (T)	1 694	332	19,60	779	45,99	583	34,42
relectures (T)	603	157	26,04	190	31,51	256	42,45
débats (T)	11 533	6 572	56,98	4 961	43,02	⊘	⊘
Corpus d'apprentissage	Total textes	Classe 0 Textes	Classe 0 %	Classe 1 Textes	Classe 1 %	Classe 2 Textes	Classe 2 %
aVoiraLire (A)	2 074	309	14,90	615	29,65	1150	55,45
jeuxvideo (A)	2 537	497	19,59	1166	45,96	874	34,45
relectures (A)	881	227	25,77	278	31,55	376	42,68
débats (A)	17 299	10 400	60,12	6 899	39,88	⊘	⊘

Table 7: Statistiques par classe sur les quatre corpus d'apprentissage (A) et de test (T).

", " *la nouvelle méthode proposée pose des problèmes complexes ... qui ne sont pas traités dans ce papier* ", cependant il a été accepté.

> ***3:2 relectures***
> ***Les différentes parties de l'article me paraissent déséquilibrées****. Les auteurs présentent d'abord un état de l'art dans le domaine de la visualisation des connaissances dans les systèmes de gestion de connaissances. Ils décrivent ensuite le serveur <anonyme /> et sa représentation des connaissances sous forme d'arbre en section 3 et une partie de la section 4. L'approche proposée par les auteurs (représentation par graphes n'est présentée qu'en 4.2 sur moins d'une page).* ***Les problèmes posés par cette méthode sont survolés par les auteurs****, ils font référence aux différents papiers traitant de ces problèmes et* ***n'exposent pas du tout*** *les heuristiques choisies dans leurs approches.* ***Le travail me paraît inachevé****, et* ***la nouvelle méthode proposée pose des problèmes complexes*** *au niveau de la construction de ce graphe* ***qui ne sont pas traités dans ce papier****.*

Notice 3:6 (**relectures**). L'article a été accepté, alors que notre système le classe comme rejeté. L'article arbitré est peut-être trop court, mais la relecture qui le concerne, elle l'est aussi :

> ***3.6 relectures***
> *Article trop court pour pouvoir être jugé. Je suggère de le mettre en POster si cela est prévu.*

Pour la notice 3:9 (**relectures**) on décèle le même problème : l'article est accepté alors que le système le rejette. Constatons que l'arbitre focalise uniquement sur des remarques de forme :

> ***3:9 relectures***
> *Question : comment est construit le réseau bayesien ? Un peu bref ici... Remarques de forme : page 2, 4ème ligne, " comprend " 5ème ligne : "annotées" ou "annoté" page 3 : revoir la phrase confuse précédant le tableau dernière ligne, répétition de "permet" page 5 : 7ème ligne accorder "diagnostiqué" et "visé" avec "états" ou avec "connaissances"*

Pour le texte de la notice 3:567 (**relectures**), l'article en question est rejeté alors que le système l'accepte. Phrases encourageantes au début finisent par être mitigées. Beaucoup d'expressions positives (" *bien organisé* ", " *facile à suivre* ", " *bibliographie est plutôt complète* ", "*solution proposée et interessante et originale* ", " *La soumission*

Corpus	Précision	Rappel	F-score	Correctes	Total
aVoiraLire (T)	0,6540	0,5590	0,6028	931	1 386
jeuxvideo (T)	**0,8114**	**0,7555**	**0,7824**	1 333	1 694
relectures (T)	0,5689	0,5565	0,5626	353	603
débats (T)	0,7307	0,7096	0,7200	8 403	11 533

Table 8: Précision, Rappel et F-score obtenus par notre méthode de fusion, sur les corpus de test (T).

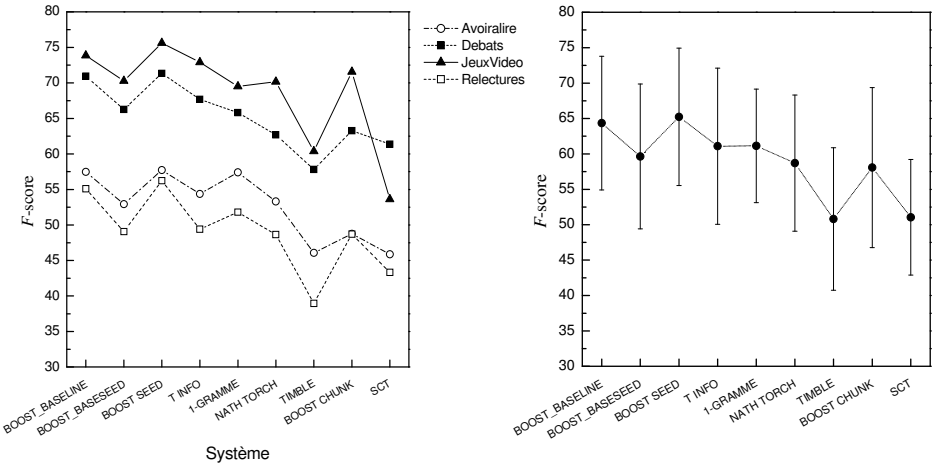

Figure 2: F-score de chacune des méthodes sur les quatre corpus de test, ainsi que leurs moyennes.

d'une nouvelle version ... sera intéressante ") n'arrivent pas a renverser le rejet.

> ***3:567 relectures***
> *Commentaire : L'article est* ***plutôt bien organisé*** *(malgré de trop nombreux chapitres), le cheminement de* ***la logique est facile à suivre****. Cependant il y a de trop nombreuses fautes de français ainsi que d'anglais dans le résumé.* ***La bibliographie est plutôt complète****.* ***La solution proposée et interessante et originale****, cependant des notions semblent mal maîtrisées. Ainsi dans la section 8, la phrase «Cette convergence ne vient pas des algorithmes génétiques de manière intrinsèque, mais de l'astuce algorithmique visant à conserver systématiquement le meilleur individu dans la population » démontre une incompréhension du fonctionnement même d'un algorithme génétique. La soumission d'une nouvelle version modifiée de cet article, présentant également les premiers résultats obtenus avec le prototype à venir* ***sera intéressante pour la communauté****.*
> *Références : Originalité : Importance : Exactitude : Rédaction :*

Pour finir, étudions une notice du corpus de films, livres et spectacles : le texte 1:10 du corpus **aVoiraLire**. Malgré des expressions avec une certaine charge positive, telles que : "*...est un événement* ", "*Agréable surprise* " ou encore "*l'image d'une cohérence artistique retrouvée* ", la notice reste difficile à classer. Évidemment notre système se trompe. Mettons à notre tour le lecteur au défi de trouver la véritable classe[13].

[13] Si vous êtiez tenté de le mettre en classe 2 (bien), sachez que la classe véritable est la 1 (moyen).

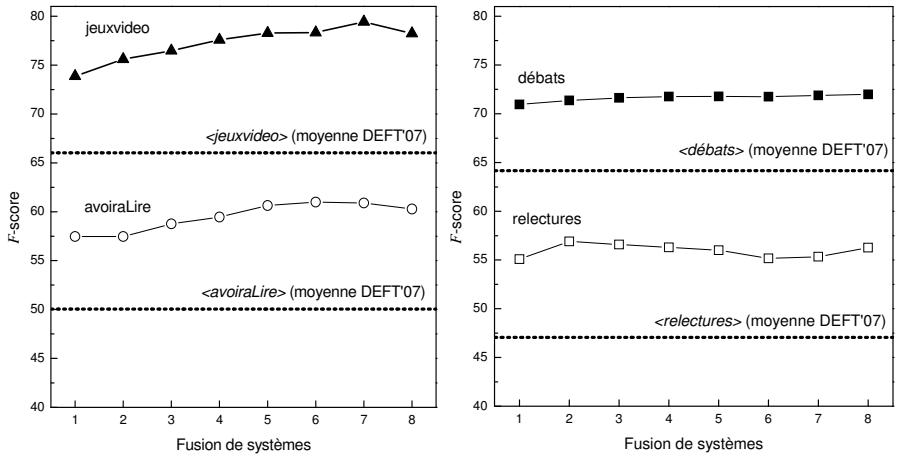

Figure 3: F-score de la fusion suivant nos neuf méthodes ajoutées. 1: BOOST_BASELINE ; 2: (1) ∪ BOOST_BASESEED ∪ BOOST_SEED ; 3: (2) ∪ Théorie information ; 4: (3) ∪ 1-gramme ; 5: (4) ∪ NATH_TORCH ; 6: (5) ∪ TIMBLE ; 7: (6) ∪ BOOST_CHUNK ; 8: (7) ∪ SCT

aVoiraLire 1:10
Depuis trente-six ans, chaque nouvelle production de David Bowie est un événement. Heathen , ne fait pas exception à cette règle. On reconnaît instantanément la patte de son vieux compère Tony Visconti. La voix de Bowie est mise en avant. Agréable surprise, surtout qu'elle n'a rien perdu depuis ses débuts. Là, commence le voyage. Ambiance, mélange dosé des instruments. Dès l'ouverture de l'album avec Sunday , un sentiment étrange nous envahit. Comme si Bowie venait de rentrer d'un voyage expérimental au coeur même de la musique. Retour aux sources. L'ensemble du disque est rythmé par cette pulsation dont le duo a le secret. Le tout saupoudré de quelques pincées d'électronique. Le groupe est réduit au minimum. Outre Bowie en chef d'orchestre et Visconti, David Torn ponctue les compositions de ses guitares aventureuses et Matt Chamberlain apporte de l'âme à la rythmique. Un quatuor à cordes fait une apparition, comme Pete Townshend (The Who) ou Dave Grohl (ex-batteur de Nirvana). Avec trois reprises réarrangées et neuf compositions originales, le 25e album de Bowie est à l'image d'une cohérence artistique retrouvée.

6 Conclusion et perspectives

La classification de documents textuels en fonction des tendances d'opinion qu'ils expriment reste une tâche très difficile, même pour une personne. La lecture directe ne suffit pas toujours pour se forger un avis et privilégier une classification pa rapport à une autre. Nous avons décidé d'utiliser des approches de représentation numériques et probabilistes, afin de rester aussi indépendant que possible des sujets traités. Nos méthodes ont fait leur preuve. Nous avons confirmé l'hypothèse que la réécriture (normalisation graphique) et l'agglutination par (collocations) aident à capturer le sens des avis. Ceci se traduit par un gain important de performances. Probablement la méthode de normalisation et d'agglutination utilisée en DEFT'08 aurait eu un impact favorable pour les performances. Cela reste un sujet à étudier. Nous avons présenté une stratégie de fusion de méthodes assez simple. Celle ci s'est avérée robuste et performante. Une stratégie similaire a été utilisée lors du défi DEFT'08. Dans ce dernier cas, le défi comportait la prise en compte des variations en genre et en thème dans un système de classification automatique[14]

[14]Des corpus électronique issus du journal *Le Monde* ou du site *Wikipedia.* Les tâches ont été les suivantes : *Tâche 1 Catégorie* - Reconnaissance de la catégorie thématique parmi les quatre classes : *ART* (Art), *ECO* (Économie), *SPO* (Sports), *TEL* (Télévision) ; *Tâche 1 Genre* - Reconnaissance du genre parmi les deux classes : *W* (Wikipédia), *LM* (Le Monde) ; *Tâche 2 Catégorie* - Reconnaissance de la catégorie

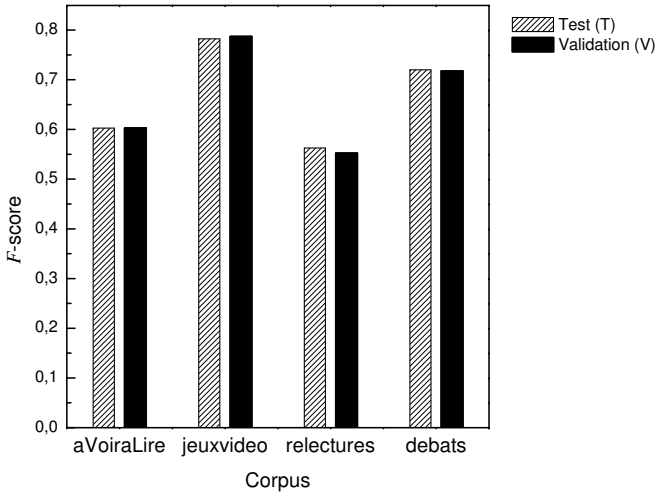

Figure 4: Comparaison du F-score de l'ensemble de validation (V) vs. celui de test (T) pour chacun des corpus, obtenu par notre système de fusion.

Cette stratégie nous a permis de remporter également le défi DEFT'08, en ex-equo avec deux autres équipes. Nos F-scores sont au-dessus des moyennes sur les quatre corpus de test, notamment sur celui de **jeuxvideo**. La stratégie de fusion a montré des résultats supérieurs à n'importe laquelle des méthodes individuelles. La dégradation en précision et rappel reste faible, même si nous n'avons pas écarté de la fusion des méthodes peut-être moins adaptées à cette problématique. La fusion est donc une façon robuste de combiner plusieurs classifieurs. Il faut souligner la remarquable équivalence entre les résultats obtenus lors de l'apprentissage et la prédiction sur les ensembles de test : à un point près de différence. Le module de fusion n'a pas vraiment été optimisé, car un même poids a été attribué au vote de chaque système. En effet, diverses méthodes peuvent être employées pour fusionner des hypothèses de classification : vote simple, vote pondéré, moyenne pondérée des scores de confiance, régression, classifieur de classifieurs, etc. En DEFT'07 et 08 nous avons choisi de privilégier les méthodes simples (Béchet *et al.*, 2008). Ceci peut ouvrir facilement la voie à une possible amélioration. D'abord, on aurait pu utiliser utiliser une méthode exhaustive de recherche de poids. De ce fait, on aurait pu utiliser, par exemple, la moyenne pondérée des scores de confiance, avec un jeu de coefficients choisi pour minimiser l'erreur sur le corpus d'apprentissage. Une autre façon de trouver les poids, est au moyen des techniques d'apprentissage. Des techniques d'optimisation ou probabilistes pourraitent être intégrées afin de regler automatiquement les paramètres de pondération des juges, tel qu'on a fait lors de la campagne DEFT'05 (El-Bèze *et al.*, 2005) avec un perceptron optimale (Torres-Moreno *et al.*, 2002; Gordon & Berchier, 1993). Il faut dire qu'un système modulaire comme le notre, peut accepter l'ajout d'autres méthodes de classification (comme celle que nous avons présenté en DEFT'08 qui classe les textes en utilisant uniquement les signes de ponctuation et les mots vides de signification) dans le but d'accroitre les performances. Il restent donc des voies à explorer. En ce qui concerne l'analyse detaillée des resultats, le corpus de **relectures** des articles scientifiques reste de loin le plus difficile à traiter. Nous avions déjà avancé l'hypothèse qu'en raison du faible nombre de notices, il serait difficile à classer. Il y a d'autres facteurs qui interviennent également. Les relectures souvent comportent, dans le corps du texte, des corrections adressées aux auteurs. Ceci vient bruiter nos classifieurs. Les relectures sont parfois trop courtes, ou bien elles ont été redigées par des arbitres non francophones (encore une autre source de bruit) ou bien elles contienent beaucoup d'anglicismes (*weak acceptation, boosting, support vector,...*). Un autre facteur, peut-être plus subtil : un article peut être lu par plusieurs arbitres (deux, trois voire plus) qui émettent des avis opposés. Dans une situation où les arbitres A et B acceptent l'article et un troisième C le refuse, normalement l'article doit être accepté. Donc, dans le corpus **relectures**, l'avis de C sera asimilé à la classe acceptée, et cela malgré son avis négatif.

thématique parmi les six classes : *FRA* (Politique française), *INT* (International), *LIV* (Littérature), *SCI* (Sciences), *SOC* (Société).

Remerciements

Nous remercions Martine Hurault-Plantet et Cyril Grouin (LIMSI-LIR) ainsi que le comité d'organisation de DEFT'09.

Références

Azé J., Heitz T., Mela A., Mezaour A.-D., Peinl P. et Roche M. (2006). Préparation de DEFT'06 (DÉfi Fouille de Textes). In *Proc. of Atelier DEFT'06*, volume 2.

Azé J. et Roche M. (2005). Présentation de l'atelier DEFT'05. In *Proc. of TALN 2005 - Atelier DEFT'05*, volume 2, p. 99–111.

Béchet F., El-Bèze M. et Torres-Moreno J.-M. (2008). En finir avec la confusion des genres pour mieux séparer les thèmes. In *DEFT'08*, p. 161–170.

Béchet F., Nasr A. et Genet F. (2000). Tagging unknown proper names using decision trees. In *38th Annual Meeting of the Association for Computational Linguistics, Hong-Kong, China*, p. 77–84.

Collobert R., Bengio S. et Mariéthoz J. (2002). Torch: a modular machine learning software library. In *Technical Report IDIAP-RR02-46, IDIAP*.

Daelemans W., Zavrel J., van der Sloot K. et van den Bosch A. (2004). Timbl: Tilburg memory based learner, version 5.1, reference guide. *ILK Research Group Technical Report Series*, p. 04–02.

El-Bèze M., Torres-Moreno J.-M. et Béchet F. (2005). Peut-on rendre automatiquement à César ce qui lui appartient ? Application au jeu du Chirand-Mitterrac. In *TALN 2005 - Atelier DEFT'05*, volume 2, p. 125–134.

Freund Y. et Schapire R. E. (1996). Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, p. 148–156.

Gordon M. et Berchier D. (1993). Minimerror: A perceptron learning rule that finds the optimal weights. In M. Verleysen, Ed., *ESANN*, p. 105–110, Brussels: D facto.

Hatzivassiloglou V. et McKeown K. R. (1997). Predicting the semantic orientation of adjectives. In *European chapter of the ACL*, p. 174–181, Morristown, NJ, USA: ACL.

Kuhn R. et De Mori R. (1995). The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**(5), 449–460.

Mani I. et Maybury M. T. (1999). *Advances in Automatic Text Summarization*. MIT Press.

Quinlan J. (1986). Induction of decision trees. *Machine Learning*, **1**(1), 81–106.

Quinlan J. (1993). *C4. 5: Programs for Machine Learning*. Morgan Kaufmann.

Schapire R. E. et Singer Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning*, **39**, 135–168.

Torres-Moreno J.-M., Aguilar J. et Gordon M. (2002). Finding the number minimum of errors in N-dimensional parity problem with a linear perceptron. *Neural Processing Letters*, **1**, 201–210.

Vapnik V. N. (1982). *Estimation of Dependences Based on Empirical Data*. New York, USA: Springer-Verlag Inc.

Vapnik V. N. (1995). *The nature of statistical learning theory*. New York, USA: Springer-Verlag Inc.

Wilson T., Wiebe J. et Hoffmann P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*, p. 347–354, Vancouver, Canada.

Session II - Présentation et résultats

Actes du cinquième défi fouille de texte, DEFT2009, Paris, France, 22 juin 2009.
Proceedings of the Fifth DEFT Workshop, DEFT2009, Paris, France, 22nd June 2009.
Pages 35-50
DEFT'09 « DÉfi Fouille de Textes », Atelier de clôture, Paris, 22 juin 2009

Présentation de l'édition 2009 du DÉfi Fouille de Textes (DEFT'09)

Cyril Grouin[(1)], Béatrice Arnulphy[(1)], Jean-Baptiste Berthelin[(1)], Sarra El Ayari[(1)], Anne García-Fernandez[(1)], Arnaud Grappy[(1)], Martine Hurault-Plantet[(1)], Patrick Paroubek[(1)], Isabelle Robba[(1)] et Pierre Zweigenbaum[(1)]

[(1)]LIMSI–CNRS
BP 133 – F-91403 Orsay Cedex
prenom.nom@limsi.fr

Résumé – Abstract

La cinquième édition du défi fouille de textes (DEFT) porte sur la fouille d'opinion. Deux corpus multilingues (français, anglais et italien) ont été produits, le premier composé d'articles de journaux, le second de débats parlementaires européens. Trois tâches sont proposées : 1^o Identifier le caractère globalement objectif ou subjectif d'un article de journal, 2^o Identifier les passages subjectifs dans des articles de journaux et dans des interventions parlementaires, et 3^o Identifier le parti politique d'appartenance d'un parlementaire à partir d'interventions. Cet article présente le déroulement de la campagne, de la constitution des corpus aux mesures d'évaluation utilisées en passant par les évaluations humaines.
The fifth edition of the DEFT (DÉfi Fouille de Textes) Text Mining Challenge focuses on opinion mining. Two multilingual corpora (French, English and Italian) were produced, the first composed of newspaper articles, the second of parliament debates. Three tasks are proposed : 1st Identify the overall objectiveness or subjectiveness in a newspaper article, 2nd Identifying subjective passages in newspaper articles and speeches in parliament, and 3rd Identify the political party affiliation of a parliamentarian from interventions. This article describes the campaign, the creation of corpus, the evaluation measures used and the human scores.

Mots-clefs – Keywords

Corpus multilingues, fouille d'opinion, référence par votes majoritaires
Majority vote reference, multilingual corpora, opinion mining

1 Introduction

Pour cette cinquième édition du DÉfi Fouille de Textes, nous avons fait le choix de proposer une nouvelle tâche en fouille d'opinion. Il s'agit d'un thème intéressant à plus d'un titre : des entreprises en vivent, parfois même en complément de sondages d'opinion plus classiques, et le Web fournit des données en abondance, issues de blogs, de réseaux sociaux, de sites d'évaluation de produits, ou encore de journaux en ligne. Les applications concernent l'analyse et le suivi d'une « image » publique ou médiatique, avec des sphères d'application dans le commerce (image d'un produit, d'un service, d'une société), la vie publique (image d'une personnalité médiatique) ou politique (perception d'un projet politique).

Une analyse d'opinion commence par la détection du caractère plus ou moins subjectif d'un texte ou d'un passage, c'est-à-dire par déterminer s'il est porteur d'un « sentiment », d'un jugement, d'une opinion, ou au contraire de données essentiellement factuelles. Les parties de texte qui contiennent une opinion sont ensuite analysées pour donner une valeur à l'opinion exprimée, soit suivant une polarité positive/négative, soit suivant une échelle de valeurs[1]. Enfin, le jugement exprimé sur un sujet particulier peut être influencé par, ou laisser transparaître, des opinions d'un type plus général comme par exemple une opinion politique.

[1]C'était le thème retenu pour l'édition 2007 de DEFT : `http://deft07.limsi.fr`.

Pour cette campagne d'évaluation, nous avons proposé une approche multilingue de l'analyse d'opinion (français, anglais et italien) sur les trois tâches complémentaires suivantes :

– La détection du caractère objectif ou subjectif global d'un texte depuis un corpus d'articles de journaux ;
– La détection des passages subjectifs d'un texte, sur deux corpus : articles de journaux et débats parlementaires ;
– Enfin, la détermination du parti politique d'appartenance de chaque intervenant dans le corpus parlementaire.

Préalablement au lancement du défi, un groupe de juges humains a réalisé ces différentes tâches, sur un petit échantillon du corpus. L'objectif de ces évaluations humaines consistait, d'une part, à tester la faisabilité des tâches du défi, et d'autre part, à disposer d'un ordre de grandeur sur les résultats auxquels il était possible de prétendre pour chacune des tâches (Berthelin *et al.*, 2008). Précisons que le rôle de ces juges humains se limitait à une stricte participation aux tâches qui leur ont été soumises. En aucune manière ils n'ont eu pour charge de produire les données de référence des différents tâches.

Les données de référence utilisées pour chacune des tâches ont été constituées suivant deux méthodes différentes. Alors qu'il est possible de définir automatiquement les données de référence des tâches 1 et 3 (voir sections 4.1 et 6.1), nous ne disposons d'aucune référence de passages subjectifs annotés à l'intérieur d'un document. Pour la tâche 2, nous avons donc pris pour principe que la référence serait constituée par le croisement des résultats des participants à cette tâche. Nous reviendrons plus en détail sur ce choix dans la section 5.1.

Dans cet article, nous nous proposons de présenter dans un premier temps les principaux éléments du déroulement du défi, puis les corpus que nous avons rassemblés. Ensuite, pour chacune des trois tâches, nous montrerons d'abord comment nous avons constitué les données de référence, puis nous présenterons l'évaluation humaine de la tâche et enfin les résultats des participants.

2 Déroulement du Défi

2.1 Calendrier de la campagne

L'ouverture des déclarations d'intention de participation a été réalisée le 1er décembre 2008. Les corpus d'apprentissage ont été distribués à partir du 7 janvier aux équipes s'étant inscrites et ayant retourné signé le contrat d'utilisation des corpus. Comme lors des précédentes éditions, nous avons offert la possibilité aux participants de choisir leur période de test, soit trois jours complets à définir dans un intervalle d'un mois, du 18 mars au 17 avril.

Les résultats obtenus par les participants ont été diffusés, équipe par équipe, le 24 avril. Contrairement aux éditions antérieures, nous n'avons délivré aucun indice de comparaison entre participants (moyenne ou écart-type). Ce choix repose sur le modèle des campagnes TREC où les résultats globaux ne sont présentés que le jour de l'atelier de clôture.

2.2 Participations

Pour la première fois de l'existence du défi, nous avons inscrit cette campagne sous le signe du multilinguisme. À cet effet, nous avons donc proposé de travailler sur trois langues (français, anglais et italien). Cependant pour chaque tâche, nous n'avons formulé qu'une seule obligation aux participants, celle de travailler sur le français, les deux autres langues étant optionnelles. Par ailleurs, chaque équipe a pu librement choisir les tâches pour lesquelles elle souhaitait concourir sans obligation de nombre minimum de tâches.

Du fait du multilinguisme affiché pour cette campagne, plusieurs équipes internationales se sont inscrites au défi. Nous avons donc reçu les inscriptions de six équipes francophones (dont une de Belgique et une du Québec) et deux équipes non francophones (en provenance d'Allemagne et du Royaume-Uni).

Les équipes qui ont poursuivi leur travail jusqu'aux phases de tests et qui ont soumis des résultats sont les suivantes :

– CHART, *Cognition Humaine et ARTificielle* (Paris, France) : D. Legros, A. El Ghali, Y. V. Hoareau
– LINA, *Laboratoire d'Informatique Nantes Atlantique* (Nantes, France) : M. Vernier, B. Daille, N. Hernandez, L. Monceaux, S. Pena-Saldarriaga, F. Poulard
– LIPN, *Laboratoire d'Informatique de Paris Nord* (Villetaneuse, France) : M. Généreux, Th. Poibeau
– UCL, *Université Catholique de Louvain-la-Neuve* (Belgique) : G. Lories, Y. Bestgen
– UdeM, *Université de Montréal* (Canada) : D. Forest, M. Bélanger, D. Létourneau, A. van Hoeydonck
– UKP, *Ubiquitous Knowledge Processing* (Darmstadt, Allemagne) : C. Toprak, I. Gurevych

La tâche 1 (détection du caractère objectif/subjectif global d'un texte) a eu le plus de participants avec les laboratoires du CHART (sur le français, l'anglais et l'italien), du LINA (sur le français), de l'UCL (sur le français et l'anglais), de l'UdeM (sur le français) et de l'UKP (sur le français et l'anglais).

La tâche 2 (détection des passages subjectifs d'un texte) n'a eu que deux participants, le LINA et le LIPN, tous deux sur le français uniquement. Il faut souligner la particulière difficulté de cette tâche ainsi que des possibles controverses sur les principes de constitution des données de référence (voir section 5.1).

Finalement, un seul participant a poursuivi la tâche 3 (détermination du parti politique auquel appartient l'orateur) jusqu'au bout. Trois participants s'y étaient inscrits, mais deux ont renoncé à donner leurs résultats. Il est vrai que les résultats des logiciels sont faibles sur cette tâche, mais néanmoins tout à fait conformes à ce que laissait prévoir notre évaluation humaine (voir section 6.2).

2.3 Mesures d'évaluation des résultats

Les différentes tâches peuvent être considérées comme des tâches de classification, un élément à classer étant alors :

- Pour la tâche 1 : un document (article de journal) avec les classes OBJECTIF et SUBJECTIF ;
- Pour la tâche 2 : un passage de texte d'un document avec la classe SUBJECTIF, les parties de texte non étiquetées appartenant par défaut à la classe OBJECTIF ;
- Pour la tâche 3 : un document (intervention dans les débats parlementaires) avec les classes Verts-ALE, GUE-NGL, PSE, ELDR, PPE-DE.

Chaque fichier de résultat pour une tâche a été évalué en calculant la F-mesure sur toutes les classes de cette tâche avec $\beta = 1$, ce qui ne privilégie ni la précision ni le rappel, mais un équilibre entre les deux.

$$\mathrm{F}_{\text{mesure}}(\beta) = \frac{(\beta^2 + 1) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

La précision et le rappel sur les classes d'une tâche sont ici calculés suivant la macro-moyenne (Nakache & Métais, 2005) dans laquelle chaque classe compte à égalité avec les autres, qu'elle ait un fort ou un faible effectif.

F-mesure pondérée Dans la F-mesure classique, une seule classe peut être attribuée à chaque document. Cependant, un système de classification automatique peut attribuer à un document une distribution de probabilité sur les différentes classes au lieu de lui attribuer une seule classe. L'indice de confiance est cette probabilité pour un document d'appartenir à une catégorie donnée.

La F-mesure pondéré par l'indice de confiance sera utilisé à titre indicatif pour des comparaisons complémentaires entre les méthodes mises en place par les équipes.

Dans la F-mesure pondéré, la précision et le rappel pour chaque classe sont pondérés par l'indice de confiance. Ce qui donne :

$$\text{Précision}_i = \frac{\sum_{\text{attribué correct. } i=1}^{\text{Nombre attribué correct. } i} \text{indice de confiance}_{\text{attribué correct.}_i}}{\sum_{\text{attribué } i=1}^{\text{Nombre attribué } i} \text{indice de confiance}_{\text{attribué}_i}}$$

$$\text{Rappel}_i = \frac{\sum_{\text{attribué correct. } i=1}^{\text{Nombre attribué correct. } i} \text{indice de confiance}_{\text{attribué correct.}_i}}{\text{nombre de documents appartenant à la classe } i}$$

Avec :

- Nombre attribué correct.$_i$: nombre de documents attribué correct.$_i$ appartenant effectivement à la classe i et auxquels le système a attribué un indice de confiance non nul pour cette classe ;
- Nombre attribué$_i$: nombre de documents attribués$_i$ auxquels le système a attribué un indice de confiance non nul pour la classe i.

La F-mesure pondéré est ensuite calculé à l'aide des formules de la F-mesure classique.

Macro-moyenne

$$\text{Précision} = \frac{\sum_{i=1}^{n}\left(\frac{TPi}{(TPi+FPi)}\right)}{n} \quad \text{Rappel} = \frac{\sum_{i=1}^{n}\left(\frac{TPi}{(TPi+FNi)}\right)}{n}$$

Avec :
– TPi = nombre de documents correctement attribués à la classe i ;
– FPi = nombre de documents faussement attribués à la classe i ;
– FNi = nombre de documents appartenant à la classe i et non retrouvés par le système ;
– n = nombre de classes.

3 Présentation des corpus

Nous avons rassemblé deux types de corpus dans chacune des trois langues du défi. Le premier type de corpus concerne des articles de journaux tandis que le second se compose d'interventions parlementaires au Parlement européen.

Le corpus d'articles de journaux est destiné à servir aux tâches 1 (identification du caractère globalement objectif ou subjectif de l'article) et 2 (identification des passages subjectifs d'un article), tandis que le corpus des interventions parlementaires est utilisé dans le cadre des tâches 2 et 3 (détermination du parti politique auquel appartient l'orateur).

3.1 Articles de journaux

Le premier corpus intègre des articles de journaux provenant de trois titres européens :

– Le corpus en français est issu du quotidien *Le Monde*, 42 000 articles sur les années 2003 à 2006 ;
– Le corpus en anglais provient du quotidien économique *The Financial Times*, 13 000 articles de l'année 1993 ;
– Le corpus en italien comprend 2 500 articles du journal économique *Il Sole 24 Ore* sur la période 1992/1993.

Les corpus des journaux *The Financial Times* et *Il Sole 24 Ore* ont été rassemblés dans le cadre du projet MLCC (MultiLingual Corpora for Co-operation). Ce projet visait deux objectifs principaux. En premier lieu, permettre la réalisation de travaux sur des corpus comparables (à partir d'une collection d'articles de journaux en 6 langues[2] d'Europe de l'Ouest). En second lieu, fournir les bases pour des travaux de traduction (corpus parallèles multilingues dans 9 langues[3] européennes provenant de questions écrites et de débats parus au Journal Officiel de la Communauté Européenne).

Tous ces corpus d'articles de journaux sont disponibles auprès de l'agence ELDA qui les commercialise[4] sous les références ELRA-W0015 pour le corpus du *Monde* et ELRA-W0023 pour le corpus MLCC.

3.2 Débats parlementaires européens

Le corpus de débats parlementaires européens a été constitué en récupérant, depuis le site Internet du Parlement européen[5], les archives multilingues des 313 séances parlementaires qui se sont tenues entre 1999 et 2004. Dans ces archives, chacune des séances a été intégralement retranscrite et traduite dans les 11 langues officielles[6] de l'Union européenne. Chaque intervention est par ailleurs enrichie de plusieurs méta-données : le nom du parlementaire, la langue dans laquelle il s'exprime (qui n'est pas nécessairement celle de son pays d'origine) et le nom du groupe politique européen[7] duquel il relève.

[2]Corpus comparable MLCC d'articles de journaux : allemand (*Handelsblatt*), anglais (*The Financial Times*), espagnol (*Expansíon*), français (*Le Monde*), italien (*Il Sole 24 Ore*) et néerlandais (*Het Financieele Dagblad*).

[3]Corpus parallèle MLCC : allemand, anglais, danois, espagnol, français, grec, italien, néerlandais et portugais.

[4]Voir le site Internet `http://catalog.elra.info/` pour plus de précisions.

[5]Le site Internet du Parlement européen `http://www.europarl.europa.eu/` propose un libre accès à ses archives.

[6]Entre 1999 et 2004, l'UE comptait onze langues officielles : allemand, anglais, danois, espagnol, finnois, français, grec, italien, néerlandais, portugais et suédois. Depuis 2005, le nombre total de langues officielles a été porté à vingt-trois.

[7]Il existe neuf groupes politiques européens : EDD (Europe des Démocraties et des Différences), ELDR (parti Européen des Libéraux, Démocrates et Réformateurs), GUE/NGL (groupe confédéral de la Gauche Unitaire Européenne et Gauche Verte Nordique), NI (les non inscrits), PPE-DE (Parti Populaire Européen (démocrates chrétiens) et Démocrates Européens), PSE (Parti Socialiste Européen), TDI (groupe Technique des Députés Indépendants), UEN (Union pour l'Europe des Nations) et enfin, les Verts/ALE (Verts, Alliance Libre Européenne).

4 Tâche 1 : Détection du caractère objectif/subjectif global d'un texte

4.1 Constitution des données de référence

4.1.1 Préparation des données

Chaque corpus de journal enrichit ses articles de méta-données[8]. Afin de constituer automatiquement la référence de cette tâche, nous avons étudié les méta-données disponibles en nous focalisant sur deux aspects principaux : la disponibilité des méta-données de manière transversale au corpus et la possibilité d'effectuer une catégorisation sur la base des méta-données retenues.

Concernant le journal *Le Monde*, nous avons utilisé le secteur de rédaction[9] sous lequel a paru chacun des articles. Le secteur de rédaction est une subdivision du journal qui correspond à la maquette. Nous avons ainsi considéré comme étant objectifs les articles relevant des secteurs de rédaction « France » et « International » (autrement dit, des articles traitant de politique nationale et internationale) tandis que les articles des secteurs « Éditorial – Analyses » et « Débats – Décryptages » ont été qualifiés d'articles subjectifs.

Pour ce qui concerne le corpus d'articles du journal britannique *The Financial Times*, nous avons pris en compte les éléments d'indexation de chaque article en nous intéressant à deux descripteurs : les articles indexés « CMMT Comment & Analysis » ont été qualifiés d'articles subjectifs alors que ceux indexés « NEWS General News » ont été enregistrés parmi les articles objectifs. Les autres descripteurs étant plus spécifiques (*COMP Company News, MGMT Management & Marketing, RES Capital expenditures,* etc), il n'a guère été possible d'en faire usage.

Enfin, nous avons classé les articles du journal italien *Il Sole 24 Ore* en étudiant les descripteurs d'indexation : les articles ont été qualifiés d'article subjectifs s'ils étaient indexés par le descripteur « Opinioni e commenti » et objectifs dans les autres cas. Nous n'avons pas utilisé les autres descripteurs du fait de leur trop grande spécificité (*Attività immobiliari, Inchieste e notizie giudiziarie, Statistiche monetarie e finanziarie,* etc).

Journal	Objectifs	Subjectifs	Répartition
Le Monde	34 761	7 232	83%/17%
The Financial Times	5 708	7 403	44%/56%
Il Sole 24 Ore	1 559	936	62%/38%

TAB. 1 – Nombre d'articles objectifs et subjectifs pour chaque corpus de journal.

4.1.2 Exemples

Le Monde – objectif. SOUPÇONNÉ par la direction de la surveillance du territoire (DST) d'être l'un des informateurs anonymes des juges Renaud Van Ruymbeke et Dominique de Talancé, qui enquêtent sur l'affaire des frégates de Taïwan, Imad Lahoud, informaticien chez EADS, rompt le silence par la voix de son avocat. Dans un communiqué adressé au Monde, Me Olivier Pardo assure qu'« en dépit de la multiplicité des assertions aucune preuve d'une quelconque participation de -son- client à cette affaire n'existe ». « Dans une ronde sans fin, la calomnie s'installe et risque de devenir extrêmement préjudiciable », ajoute Me Pardo, qui précise que son client « n'a jamais rencontré le député Alain Marsaud », contrairement à ce qu'a laissé entendre la DST dans un rapport (Le Monde du 20 juillet). « M. Lahoud demande que cessent ces assertions et l'instrumentalisation de sa personne, conclut l'avocat. Il ne réclame qu'une chose : qu'il puisse continuer à animer son équipe de recherche dans la grande entreprise européenne à laquelle il a l'honneur d'appartenir, dans la sérénité et la quiétude. »

[8] Un document du *Financial Times* comprend les éléments suivants : titre, date de publication, signature, article, indexation par un thésaurus, lieu d'édition et numéro de page.
Un document du *Monde* comprend les éléments suivants : date de publication, secteur de rédaction, titraille (têtière, sur-titre, titre, sous-titre), chapô, nom et localisation géographique du journaliste, article, éléments d'indexation (titre complémentaire, catégories) et mots-clés (de type France, Étranger et Personne) issus d'un thésaurus interne.
Un document du journal *Il Sole 24 Ore* comprend les éléments suivants : date de publication, rubrique, chapô, signature, article, et éléments d'indexation (de type descripteurs, aire géographique et didascalies).

[9] Une étude de la répartition des articles dans les différents secteurs de rédaction du journal entre 1987 et 2006 est disponible sur le document `http://perso.limsi.fr/grouin/rubriques_lemonde_1987-2006.html` d'après un travail de S. Loiseau et C. Grouin.

Le Monde – subjectif. Avec la campagne présidentielle qui se profile, voici revenu le temps des effets chocs et des idées chics. Il en est ainsi du « dialogue social », serpent de mer du débat public que les politiques défendent surtout quand ils n'ont pas à le pratiquer. Jacques Chirac en a parlé le 14 juillet. Dominique de Villepin redécouvre son existence après l'avoir superbement ignoré. L'UMP va achever, en septembre, des rencontres avec tous les partenaires sociaux, CGT comprise. Le PS recevra, à son Université d'été de La Rochelle, du 25 au 27 août, tous les syndicats et le numéro deux du Medef, Denis Gautier-Sauvagnac.

The Financial Times – objectif. THE Health and Safety Executive yesterday accused the European Community of imposing a mass of health and safety legislation on member states without adequate thought or consultation, Diane Summers writes. The executive is concerned that it is becoming a focus of complaints that UK businesses are increasingly subject to excessively bureaucratic regulation. It is anxious to remind the government that much of the regulation and perceived red tape originates from Brussels and not from it. Mr John Rimington, the executive's director-general, said the directives had been constructed in 'smoke-filled rooms in Brussels'. He blamed, in particular, the French for trying to get the directives to reflect their own domestic laws and said some of the provisions were 'incomprehensible'.

The Financial Times – subjectif. Denmark has a new government. The foreign minister has pledged himself to secure a Yes to Maastricht in the second Danish referendum on the subject, to be held probably 'before June'. All's right with the world. Some moaning Euro-minnies are still muttering about the dangerous precedent set by Denmark's special 'opt-outs', negotiated at last month's Edinburgh summit. Won't Conservative backbenchers try to obtain the same deal for Britain, as the price of ratification ? Won't candidates for EC membership, with three of whom formal negotiations are to start on Monday, demand that the same exemptions apply to them ? Isn't this the beginning of the a la carte union so dreaded by Mr Jacques Delors, president of the European Commission ?

Il Sole 24 Ore – objectif. Sarà la presidenza del Consiglio a dire l'ultima parola sulla riforma dei fondi pensione. Il ministro del Lavoro Nino Cristofori ha già annunciato per il mese di dicembre la presentazione del suo progetto di regolamentazione delle casse integrative aziendali, in attuazione della recente legge delega sul riordino della previdenza. Quello di Cristofori, però, sarà soltanto il materiale preparatorio per la stesura definitiva del decreto legislativo che darà il via libera alla riforma. A Palazzo Chigi si è costituito in questi giorni un gruppo di esperti con il compito di analizzare le proposte del Lavoro confrontandole con quelle provenienti da altri ministeri (Tesoro, Industria e Finanze), parti sociali e gruppi economici.

Il Sole 24 Ore – subjectif. L'economia va male e quindi i tassi di interesse si riducono. Questa apparentemente banale relazione è tornata a essere vera in questi giorni un pò in tutta Europa e quindi anche in Italia. Dopo i fuochi d'artificio di un mese fa, quando sembrava che le banche centrali di tanti Paesi europei fossero solo impegnate nella nobile gara a chi alzava di più i tassi di interesse, il buonsenso economico è tornato a prevalere.

4.2 Évaluation humaine de la tâche

Le test humain de la tâche 1 a été réalisé sur un ensemble de sept articles du *Monde*. Les résultats obtenus (voir tableau 2) se révélèrent assez élevés mais doivent sans doute être relativisés en tenant compte du nombre réduit de documents constituant notre corpus d'évaluation humaine.

Il faut cependant noter que le caractère objectif ou subjectif de chaque article a toujours été bien reconnu par la majorité des juges humains. Sur un vote majoritaire, le groupe des six juges humains auraient donc obtenu 100% de réussite.

Testeur	**1**	**2**	**3**	**4**	**5**	**6**
Rappel	0,71	0,83	0,67	0,88	1,00	0,88
Précision	0,71	0,90	0,83	0,88	1,00	0,88

TAB. 2 – Rappel et précision obtenus par les testeurs humains sur la tâche de qualification globale des articles.

4.3 Résultats

La première tâche proposée concernait donc la caractérisation globale d'articles de journaux parmi deux classes possibles : objectif ou subjectif. Cinq équipes se sont essayées à cette tâche, chacune sur le français, trois d'entre elles sur l'anglais, une seule pour l'italien.

Nous présentons dans le tableau 3 la F-mesure obtenue par chaque équipe, pour chacune des soumissions effectuées dans chacune des langues de cette tâche. Une F-mesure suivie d'une étoile renvoie à une soumission utilisant des indices de confiance (pondération des valeurs de résultat).

Langue	Équipe	F-mesures par soumission
Anglais	CHART	0,676 – 0,652*
Anglais	UCL	0,851
Anglais	UKP	0,822 – 0,769 – 0,814
Français	CHART	0,771 – 0,715*
Français	LINA	0,850
Français	UCL	0,925
Français	UdeM	0,757 – 0,778 – 0,781
Français	UKP	0,662 – 0,769
Italien	CHART	0,716 – 0,691*

TAB. 3 – F-mesures obtenues pour chaque soumission de chaque équipe dans chacune des langues sur la tâche 1. L'étoile indique qu'il s'agit d'une F-mesure pondérée.

Afin de faciliter la comparaison des résultats obtenus entre équipes, nous ne représentons dans le graphique suivant que la meilleure soumission obtenue par chaque équipe.

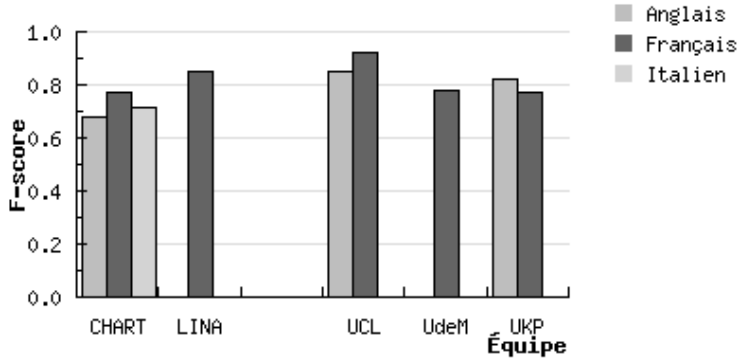

FIG. 1 – F-scores obtenus sur la meilleure soumission de chaque équipe sur la tâche 1.

Il apparaît que les meilleurs résultats ont été obtenus, pour les équipes francophones, sur le corpus en français tandis que l'équipe allemande (UKP), non francophone, a obtenu de meilleurs résultats sur l'anglais.

5 Tâche 2 : Détection des passages subjectifs d'un texte

5.1 Constitution des données de référence

5.1.1 Principe

Dans la mesure où nous ne disposions pas de données annotées en passages objectifs ou subjectifs, nous avons retenu le principe de constitution des données de référence par vote majoritaire : les données de référence d'un corpus donné sont constituées a posteriori, à partir d'un vote majoritaire entre les résultats des participants. Ce principe a déjà été expérimenté dans des campagnes d'évaluation d'analyseurs syntaxiques (Paroubek *et al.*, 2008).

Dans une tâche de classification, les données de référence sont donc constituées par les catégorisations sur lesquelles la majorité des participants à cette tâche sont tombés d'accord. Par exemple si la phrase « *L'affaire ne devrait pas améliorer les relations entre Séoul et Pyongyang.* » a été considérée comme un passage subjectif par la majorité des participants, alors cette phrase sera annotée comme subjective dans les données de référence. Et tous les mots de cette phrase, considérés seulement dans cette phrase évidemment, seront comptés comme mots subjectifs. Dans le cas contraire, cette phrase sera annotée comme objective.

Les données de référence sont donc, suivant ce principe, les données qui ont été classées de la même manière par les logiciels en compétition. Elles sont intéressantes car elles donnent un état des capacités des logiciels, mais nous sommes conscients qu'elles peuvent être contestées en tant que références. L'évaluation des logiciels par rapport à cette référence donne une bonne idée des écarts entre les logiciels, sans pour autant donner un classement fiable entre participants. Par ailleurs, la faible participation à cette tâche affaiblit également la valeur des données de référence.

5.1.2 Exemples

Nous avons expérimenté ce principe de constitution des données de référence lors de l'évaluation humaine de la tâche (voir la section suivante 5.2). L'exemple de texte qui suit (tiré d'un article du *Monde*) est extrait des données de référence constituées à partir des résultats de cette évaluation humaine sur le petit corpus de la tâche 1. Il est annoté en passages subjectifs et objectifs.

SUBJECTIF : Pourquoi ne pas fonder sans complexes une critique des feux d'artifice ? Trop vulgaires ? Puérils ? Naïfs ? Manque de dignité esthétique, en somme ? Allons, allons.
OBJECTIF : De Chantilly, en juin, à Saint-Sébastien (jusqu'au 17 août), en passant par Biarritz le 15 août, et toutes les nuits olympiques en Chine,
SUBJECTIF : la matière ne manque pas. Les trois vertus théologales de la critique moderne (affluence, âge, berlue) s'y hisseraient à leur zénith.
OBJECTIF : Exemple : plus personne ne se risque, en festival pyrotechnique, au bleu de Chine, aperçu pour la dernière fois en 1957.
SUBJECTIF : Chimiquement trop complexe et bien aléatoire. A Pékin, peut-être ?
OBJECTIF : Pendant les
SUBJECTIF : fameuses
OBJECTIF : fêtes de Pampelune (7-14 juillet), Mikel Pagola Erviti exerce la fonction de critique de feux d'artifice au Diario de Navarra. Quand une prestation déçoit Mikel Pagola Erviti - la Pirotecnia Vicente Caballer de Valence, cette année -, il égrène quatre motifs :
SUBJECTIF : composition peu discernable, manque de rythme, faiblesses des effets, banalité des couleurs.
OBJECTIF : Sur place, un jury
SUBJECTIF : pointilleux
OBJECTIF : s'aligne sur ces critères. Le lendemain à 9 heures, la chaîne Navarra 6 retransmet en boucle le festival pyrotechnique de la veille.
SUBJECTIF : Son infernal bruit de guerre s'y noie évidemment. Or, la vérité d'un feu, c'est son bruit.
OBJECTIF : A la fin des années 1950, la chaîne locale de la radio d'Etat diffusait en direct le feu d'artifice de Biarritz.
SUBJECTIF : Un feu d'artifice à la radio, formidable. Presque mieux qu'un concours de mime. A Pampelune, Mikel Pagola Erviti misait sur la Pirotecnia Zaragozana. Laquelle ressent la gloire qu'on lui ait confié l'ouverture et la clôture de l'Expo universelle.
OBJECTIF : Le gérant de la Zaragozana occupe la chaire de chimie à l'université de la ville.
SUBJECTIF : Ses cours sont très marrants. Mais le favori de Mikel Pagola Erviti, c'est " Gori ",
OBJECTIF : le fondateur de la Pirotecnia Gori à Valence. Gregorio Juan Moreno, dit " El Gori ", a effectué sa première et dernière présentation à Pampelune le 11 juillet.
SUBJECTIF : Le Gori ne fait rien comme les autres. Il se signale par un inimitable parfum des enchaînements, des durées et des surprises du temps. Le Gori est une légende. Il attendait de Pampelune la consécration pour se retirer dans le bouquet, tel un Cincinnatus de la belle bleue. Gare au Gori.

5.2 Évaluation humaine de la tâche

Pour la tâche 2, chaque testeur humain a encadré les passages objectifs et subjectifs de balises `<obj>...</obj>` et `<sub>...</sub>`, dans chaque article du corpus d'évaluation humaine de la tâche 1. Un passage est un

extrait de texte dont nous avons volontairement laissé indéfinies les limites. Celui-ci pouvait donc aller d'un mot à plusieurs phrases suivant l'estimation personnelle du juge humain. Parfois, c'est seulement un modifieur qui a été qualifié de subjectif, mais très souvent, c'est une proposition complète qui constitue un passage subjectif. Les annotations ont ensuite été alignées au niveau du mot. Pour chacun des mots du corpus, la valeur majoritairement attribuée par les testeurs à ce mot a constitué la référence. Les annotations des testeurs ont ensuite été évaluées sur la base de cette référence. Les résultats obtenus par les testeurs humains sur cette tâche se sont révélés très bons et encourageants pour l'organisation du défi.

Testeur	**1**	**2**	**3**	**4**	**5**	**6**
Rappel	0,81	0,92	0,82	0,90	0,89	0,78
Précision	0,77	0,81	0,73	0,79	0,78	0,70

TAB. 4 – Rappel et précision obtenus par les testeurs humains sur la tâche de détection des passages subsjectifs.

L'un des enseignements que nous pouvons tirer de cette expérience concerne le caractère personnel de ce qui constitue un élément subjectif par opposition à un élément objectif. Ainsi, le corpus que les six testeurs humains ont eu pour charge d'annoter comprenait 5036 mots. Sur ces 5036 mots, on observe une concordance stricte entre les six testeurs sur 2053 mots seulement, soit 41% du corpus. Parmi ces concordances, il existe 1447 concordances sur des mots objectifs et 606 concordances sur des mots subjectifs. Malgré ce désaccord apparent, les bons résultats des testeurs humains (voir tableau 4) montrent que chacun d'entre eux a finalement peu d'écart avec l'accord majoritaire.

Nous donnons ci-après le nombre de mots catégorisés « objectif » et « subjectif » par chacun des testeurs.

Testeur	**1**	**2**	**3**	**4**	**5**	**6**
Mots objectifs	3635	3050	3263	2936	2455	2186
Mots subjectifs	1401	1986	1773	2100	2581	2850

TAB. 5 – Nombre de mots catégorisés « objectif » et « subjectif » par chacun des testeurs humains.

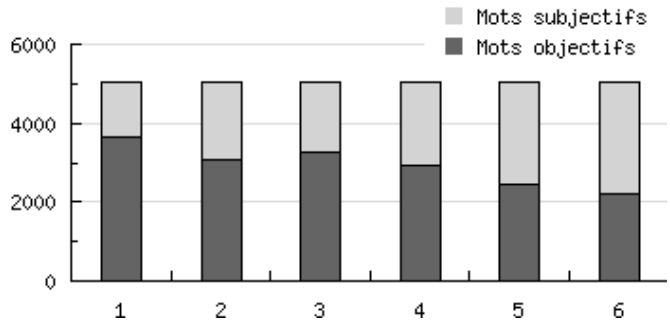

FIG. 2 – Variations personnelles des mots catégorisés « objectif » (blocs foncés) et « subjectif » (blocs clairs) pour chacun des testeurs humains (barres numérotées de 1 à 6), sur les 5036 mots du corpus testé.

Il apparaît ainsi que le testeur humain n° 1 considère le corpus comme étant majoritairement objectif (72% des mots du corpus sont catégorisés « objectifs ») alors qu'à l'opposé, le testeur humain n° 6 envisage le corpus sous un aspect nettement plus subjectif (57% des mots du corpus relèvent de passages catégorisés « subjectifs »).

Lien entre les tâches 1 et 2 Nous avons voulu évaluer également en quoi la qualification globale de l'article en subjectif/objectif d'une part, et la détection des passages subjectifs d'autre part, pouvaient se rejoindre. Nous avons là en effet deux points de vue sur la subjectivité d'un texte, l'un global et l'autre local. Pour chaque article, nous avons donc comparé la proportion de mots qualifiés de subjectifs en moyenne par l'ensemble des juges humains, à la référence de qualification globale de cet article en objectif ou subjectif. Le tableau 6 qui rassemble ces comparaisons montre qu'à l'exception de l'article 1656, et donc pour 6 articles sur les 7 du corpus donné aux juges humains, un article considéré comme objectif comporte une majorité de mots classés objectifs par les juges humains, et un article considéré comme subjectif comporte une majorité de mots classés subjectifs par les

juges humains. Cela montre une certaine cohérence, concernant la différenciation entre objectif et subjectif, entre l'impression globale sur un texte et la somme des impressions locales.

L'article 1656 en revanche constitue un contre-exemple. Cet article, qui est un éditorial à propos d'une biographie, a été jugé globalement de caractère subjectif par la majorité des juges humains (quatre juges sur les six). Malgré cela, ces mêmes juges ont trouvé localement peu de passages subjectifs dans cet article. Les liens entre le caractère subjectif global d'un texte et les passages subjectifs qu'il contient sont parfois complexes et mériteraient une analyse approfondie.

Article (id)	**4415**	**2628**	**1662**	**1935**	**1656**	**3998**	**4123**
Caractère de l'article	objectif	subjectif	objectif	objectif	subjectif	subjectif	objectif
Mots subjectifs	12%	67%	22%	33%	28%	71%	39%

TAB. 6 – Proportion de mots subjectifs dans chaque article, suivant les juges humains, comparé à sa qualification globale de référence en objectif/subjectif.

5.3 Résultats

Dans la présentation proposée sur le site Internet du défi, nous avions définie cette tâche de la manière suivante : « *Un texte peut être segmenté 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. Cette deuxième tâche consiste donc à repérer les passages subjectifs d'un texte, que ce texte soit globalement subjectif ou objectif. Un passage peut aller d'un mot (par exemple un modifieur) à plusieurs phrases* ». Un passage est donc un extrait de texte pouvant aller d'un mot à plusieurs phrases. Dans la section 5.1.2, nous avons présenté un exemple de la variabilité de la taille d'un passage, extrait du test humain de la tâche.

Les deux participants à cette tâche ont pris, concernant la taille d'un passage, des options fixes et radicalement opposées. L'un a systématiquement pris une phrase comme passage, et l'autre a pris comme passage ce que nous appellerions un *déclencheur* de subjectivité, c'est-à-dire en général un mot, modifieur ou pronom personnel par exemple.

Pour pouvoir extraire des données de référence de ces résultats, nous avons dû harmoniser leurs notions différentes d'un passage. Pour cela, nous avons systématiquement étendu les passages du deuxième participant à une portion de texte comprise entre deux ponctuations.

Cette expérience nous a montré que cette tâche était sans doute trop ambitieuse par rapport aux moyens que nous pouvions mettre en œuvre. Car si les juges humains ont intuitivement annoté des passages de tailles bien différentes, ils l'ont fait suivant des critères difficiles à expliciter. Et il est clair que les logiciels demandent, en revanche, des critères bien définis.

Résultats des participants. Deux équipes ont participé à cette tâche et ont soumis chacune trois fichiers de résultats. Chaque soumission comporte à la fois le corpus des débats parlementaires et le corpus des articles de journaux. Le tableau 7 rassemble les résultats de ces soumissions. Les données de référence pour chaque corpus résultant de l'accord majoritaire entre les six soumissions, elles constituent un sous-ensemble de toutes les soumissions. Plus une soumission est proche de ce sous-ensemble, plus sa F-mesure sera élevée. Les résultats pour chaque soumission marquent donc avant tout son écart par rapport à cet accord.

Corpus	**Équipe**	**F-mesures par soumission**	**Précisions**	**Rappels**
Journaux	LINA	0,670 – 0,623 – 0,863	0.928 – 0.623 – 0.808	0.524 – 0.623 – 0.926
Journaux	LIPN	0,777 – 0,714 – 0,775	0.701 – 0.929 – 0.699	0.871 – 0.579 – 0.869
Parlement	LINA	0,648 – 0,648 – 0,909	0.805 – 0.804 – 0.903	0.543 – 0.543 – 0.916
Parlement	LIPN	0,799 – 0,678 – 0,797	0.806 – 0.816 – 0.805	0.791 – 0.580 – 0.789

TAB. 7 – F-mesures obtenues pour chaque soumission de chaque équipe dans chacune des langues sur la tâche 2.

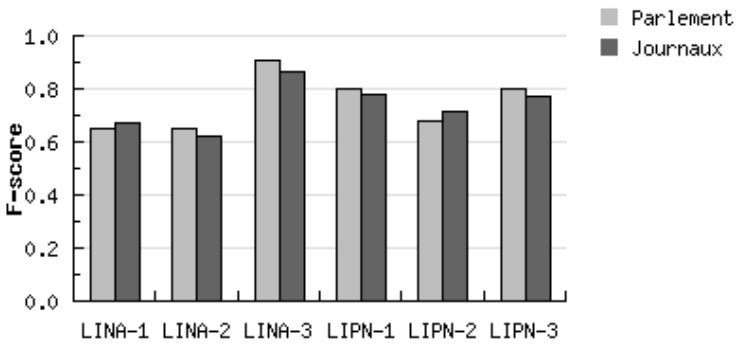

FIG. 3 – F-mesures obtenues sur les deux corpus de la tâche 2 pour chaque soumission de chaque équipe.

Accords et désaccords. La représentation graphique des résultats obtenus par les deux équipes permet d'apprécier les différences entre les différentes soumissions. Ces différences semblent légèrement plus accentuées sur le corpus du Parlement que sur le corpus des journaux. Quoi qu'il en soit, les données de référence constituées à partir des résultats marquent un état des lieux intéressant à analyser. En effet, une analyse de ces données permet de voir ce qui est actuellement repéré (avec d'ailleurs d'éventuelles erreurs), et ce qui a été omis par les logiciels.

Par exemple la phrase « *Madame la Présidente, il est indiqué à l'ordre du jour : vote de 12 heures à 13 heures, suite à 18 heures 30.* » a bien été considérée comme objective dans la majorité des soumissions, et la phrase « *Je pense que l'on devrait s'en tenir à l'ordre du jour.* » a bien été annotée comme subjective. En revanche la phrase « *Monsieur le Président, il est capital, selon moi, de disposer dorénavant d'une législation commune en matière de responsabilité environnementale.* » a été considérée comme objective dans la majorité des soumissions.

6 Tâche 3 : Détermination du parti politique auquel appartient l'orateur

6.1 Constitution des données de référence

6.1.1 Traitements des données

Ce corpus de débats a été constitué en récupérant, pour chaque séance parlementaire de la période de 1999 à 2004, l'ordre du jour, à partir duquel nous avons récupéré la retranscription des débats, dans les trois langues prévues pour le défi (pour rappel, français, anglais et italien).

Nous avons ensuite extrait de ces retranscriptions le nom de l'intervenant, son parti politique d'appartenance, la langue dans laquelle il s'est exprimé ainsi que son intervention. Le résultat de ces extractions a ensuite été aligné dans les trois langues de manière à disposer d'un corpus parallèle, chaque intervention dans une langue ayant sa traduction dans les deux autres langues, dans le même ordre des interventions pour chacune des trois langues.

Nous avons procédé à un nettoyage minimal de ce corpus aligné (suppression des fonctions de personnes, des codes de langues, et des caractères étranges) ainsi qu'à une anonymisation de base (par le remplacement des noms de groupes politiques[10] dans les textes par une balise <anonyme />).

Nous avons alors segmenté le corpus en deux parties, l'une réservée à la constitution du corpus d'apprentissage et comprenant 60% des interventions, l'autre dédiée à la réalisation du corpus de test, composée des 40% d'interventions restantes. Du fait de l'utilisation d'un corpus parallèle, nous avons la garantie que les interventions utilisées pour l'apprentissage dans une langue, ne pourront pas être présentes dans le corpus de test d'une autre langue.

Enfin, nous créons les corpus d'apprentissage et de test pour chaque langue, en mélangeant l'ordre d'apparition de chacune des interventions. Ainsi, les corpus d'apprentissage de chaque langue comprennent tous trois les mêmes

[10]Les expressions anonymisées sont, par exemple pour la français, les suivantes : ELDR, GUE/NGL, PPE-DE, PSE, Verts/ALE, Verts/Alliance libre européenne, Alliance libre européenne, gauche unie, gauche unitaire européenne, gauche verte nordique, parti des socialistes européens, démocrates-chrétiens, démocrates européens, parti populaire européen, chrétien-démocrate, gauche unitaire européenne, groupe libéral, sociaux-démocrates, gauche chrétienne, démocrate chrétien, démocrate européen, démocrates chrétiens, social-démocrate, national-démocrate-chrétien, social-démocrate.

interventions, mais présentées dans un ordre différent, fixé de manière aléatoire[11], et ce, afin d'éviter tout biais (tel que la possibilité de dupliquer les résultats pour une langue sur les deux autres langues). Il en est de même pour les trois versions du corpus de test.

Dans la perspective de la tâche 3 d'identification du parti politique de chaque intervenant, nous avons décidé de limiter le corpus aux cinq partis les plus représentés en termes de nombre d'interventions, soit : 3 346 interventions attribuées au groupe ELDR, 4 482 au GUE/NGL, 11 429 au PPE-DE, 9 066 au PSE et 3 961 aux Verts/ALE. Le parti politique d'appartenance de chaque parlementaire renseigné sur le site Internet du Parlement européen nous a permis de constituer les données de référence de la troisième tâche, en associant à chaque intervention le parti politique de son orateur.

6.1.2 Exemples

ELDR. Monsieur le Président, l'UE est le principal donateur des territoires palestiniens. Selon l'ONU, depuis les bouclages, plus d'un million de Palestiniens vivent en dessous du seuil de pauvreté, soit deux dollars par jour. La conséquence des bouclages sur le plan humain est encore bien pire. Comme le commissaire Patten l'a dit, les malades ne peuvent se rendre à l'hôpital européen de Gaza car même les ambulances ne franchissent pas les barrages. Il est donc essentiel que nous continuions d'apporter notre soutien à cette région. Je suis heureuse de constater que le Conseil et la Commission partagent ce point de vue.

GUE-NGL. Madame la Présidente, avec la proposition de modification de l'OCM des fruits et légumes, la Commission aggrave les problèmes de l'organisation des marchés actuelle, ainsi que les injustices de la PAC, et elle occasionne plus de difficultés aux producteurs de fruits et légumes. Les mesures visant à éliminer le prix minimum sont particulièrement graves, comme dans le cas de la tomate destinée à l'industrie ; de la réduction de la limite maximale de l'aide pour le volume des fonds opérationnels de 4,5 % à 3 % de la valeur de la production commercialisée de chaque organisation de producteurs ; de la réduction de 9,1 % du montant des aides à la première campagne après la réforme de l'OCM ; et de la réduction de la quantité susceptible d'indemnité communautaire de retrait pour les agrumes.

PPE-DE. Madame la Présidente, à mon grand regret j'ai dû voter contre le budget parce que je trouve absolument insuffisants, et même inexistants, tous les articles destinés à chercher à améliorer les conditions de vie des personnes âgées et des retraités. J'ai vu, en outre, que nombre de ces fonds sont destinés aux célèbres programmes d'action communautaire. Je crois que ces programmes n'exercent pas la fonction utile qu'ils devraient avoir dans l'utilisation des fonds communautaires. Je crois que l'Union européenne doit modifier complètement la façon dont elle dépense l'argent des quinze états membres de l'union.

PSE. Monsieur le Président, Mesdames et Messieurs, permettez-moi de limiter mon intervention au problème du VIH/sida. Le rapport sur l'état de la population mondiale, qui vient juste de paraître, contient des chiffres effrayants. 14 000 hommes, femmes et enfants meurent chaque jour, en moyenne, de cette maladie. Elle est devenue la première cause de mortalité en Afrique subsaharienne. Dans le monde, plus de 60 millions de personnes ont été contaminées par le virus du sida, environ 22 millions d'entre elles sont décédées. Sur les 40 millions de personnes contaminées à l'heure actuelle, 95 % vivent dans des pays en développement et presque trois quarts en Afrique. Des 580 000 enfants de moins de 15 ans morts du sida, 500 000 – près de 90 % – vivaient en Afrique. Je pourrais poursuivre indéfiniment cette énumération de statistiques édifiantes.

Verts/ALE. Monsieur le Président, hier, une fois de plus, onze personnes provenant de pays africains ont perdu la vie près des côtes espagnoles lorsque leur embarcation a fait naufrage. Ce drame, qui se produit très fréquemment, ne peut nous laisser indifférents : je crois que, malgré la difficulté, il faut rechercher une solution afin que ces gens ne soient pas obligés de tenter d'atteindre les terres européennes d'une manière aussi dramatique et tragique. Je crois que, en pareilles circonstances, il convient de le rappeler ici.

[11] À titre d'exemple, l'intervention figurant en première position dans le corpus d'apprentissage français se retrouve en position 8 399 dans le corpus d'apprentissage anglais et en position 16 074 dans le corpus d'apprentissage italien.

6.2 Évaluation humaine de la tâche

Le test a été réalisé sur le corpus des débats parlementaires européens et reposait sur l'identification de quatre partis politiques (contre cinq dans la version finale de cette tâche) : ELDR, PPE-DE, PSE et Verts/ALE. La répartition des interventions par parti dans ce petit corpus d'évaluation était la suivante : 12 ELDR, 20 PPE-DE, 25 PSE, 6 Verts/ALE.

Rappel/précision. Pour chaque testeur humain, nous avons calculé le rappel et la précision qu'il a obtenu pour chacun des partis à identifier et avons ensuite calculé un macro-rappel ainsi qu'une macro-précision. Le tableau 8 montre les valeurs de macro-rappel et macro-précision obtenues par chaque testeur humain.

Testeur	**1**	**2**	**3**	**4**	**5**	**6**
Rappel	0,41	0,35	0,39	0,23	0,47	0,35
Précision	0,42	0,34	0,37	0,27	0,42	0,34

TAB. 8 – Rappel et précision obtenus par les testeurs humains sur la tâche d'identification des partis politiques.

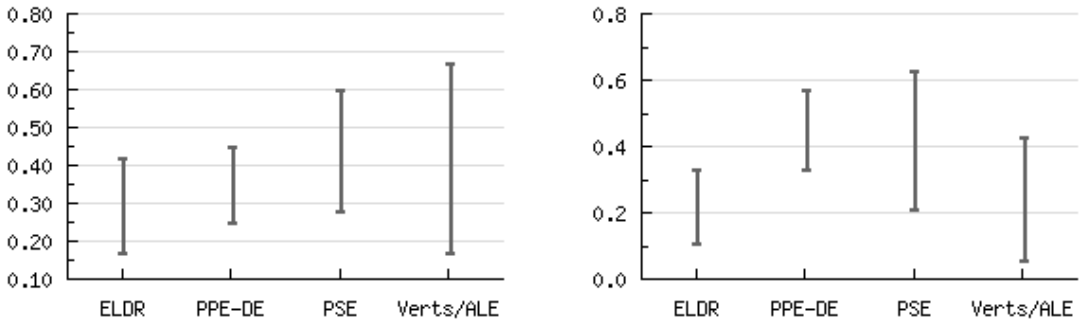

FIG. 4 – Valeurs minimales et maximales de rappel (graphique de gauche) et de précision (graphique de droite) obtenues par les testeurs humains pour chaque parti politique.

Ces résultats globaux masquent les différences qui peuvent exister entre les juges humains concernant la reconnaissance de chaque parti. Les graphiques de la figure 4 montrent ces différences en affichant pour chaque parti le minimum et le maximum des valeurs de rappel ou de précision obtenues par les juges humains. Les écarts entre les minima et les maxima montrent les désaccords entre juges, particulièrement accentués en ce qui concerne les valeurs de rappel pour le parti Verts/ALE. Il semble par ailleurs que les testeurs humains ont mieux réussi l'identification des partis PSE (précision maximale de 0,63) et PPE-DE (précision maximale de 0,57), deux partis correspondant au traditionnel clivage gauche/droite[12], que celle des autres partis. Les partis ELDR et Verts/ALE ont été moins bien identifiés, avec des valeurs minimum de précision tombant respectivement à 0,11 et 0,06). Par ailleurs, les juges humains semblent s'être mieux accordés sur l'identification du parti PPE-DE, l'écart entre minimum et maximum étant plus réduit que pour les autres partis.

Coefficient Kappa. Nous avons également confronté les résultats des différents testeurs au moyen du coefficient κ, défini par (Cohen, 1960) et repris par (Carletta, 1996). Ce coefficient permet de mesurer le taux d'accord entre deux juges. Sur cette tâche, le coefficient κ a varié entre -0,14 et 0,24, soit des accords qualifiés de très mauvais à médiocre, avec une majorité d'accords qualifiés de mauvais (pour des valeurs de κ comprises entre 0,02 et 0,17). Les accords entre juges sont donc mauvais sur cette tâche et leurs performances plutôt médiocres dans l'ensemble comme nous l'avons vu au paragraphe précédent (voir tableau 8).

Matrice de confusion. Enfin, à partir des matrices de confusion pour chaque juge humain, nous avons voulu retrouver quelles étaient les confusions entre partis les plus fréquentes. Pour cela, nous avons rassemblé dans le

[12]Le parti PSE – Parti Socialiste Européen étant à gauche tandis que le parti PPE-DE – Parti Populaire Européen (démocrates chrétiens)-Démocrates Européens est un parti de droite.

tableau 9 les répartitions, en pourcentage, de l'effectif de chaque parti parmi les autres partis, suivant les attributions effectuées par les juges humains. Ainsi, pour le parti ELDR (première ligne), les réponses effectivement attribuées à ce parti par les testeurs humains ont été au minimum de 17% et au plus de 42% de l'effectif réel des interventions de ELDR. Toujours concernant l'effectif réel des interventions du parti ELDR, les testeurs humains en ont attribué – à tort – entre 17% et 50%, suivant le juge, au parti PPE-DE, entre 0% et 33% au parti PSE et entre 0% et 42% au parti Verts/ALE. Les différences importantes qu'on observe entre minima et maxima d'attributions rendent évidents les désaccords entre juges. Tous ces éléments montrent la difficulté de cette tâche.

	Partis attribués par les juges humains aux interventions d'un même parti			
Parti	**ELDR**	**PPE-DE**	**PSE**	**Verts/ALE**
ELDR	17% < 42%	17% < 50%	0% < 33%	0% < 42%
PPE-DE	10% < 40%	25% < 45%	5% < 50%	5% < 30%
PSE	16% < 28%	8% < 32%	28% < 60%	4% < 28%
Verts/ALE	0% < 17%	0% < 5%	4% < 29%	17% < 67%

TAB. 9 – Valeurs minimales et maximales d'attribution de l'effectif de chaque parti, dans ce parti et les autres, résultant du test humain.

Une lecture globale de ce tableau permet néanmoins de mettre en évidence des « couples » de partis politiques pour lesquels certains juges ont rencontré des difficultés d'identification. Les erreurs d'identification qui ont été les plus importantes sont les suivantes : PPE-DE au lieu de ELDR (jusqu'à 50% d'interventions ELDR attribuées à PPE-DE), PSE au lieu de PPE-DE (jusqu'à 50% de mauvaises attributions), PPE-DE au lieu de PSE (jusqu'à 32% de mauvaises attributions), et enfin PSE au lieu des Verts/ALE (jusqu'à 29% de mauvaises attributions).

Si l'on établit une échelle des partis politiques, en classant les quatre partis testés de gauche à droite sur l'échiquier politique, nous obtenons la représentation suivante : Verts/ALE (gauche écologique) – PSE (gauche) – ELDR (centre-droit) – PPE-DE (droite). En se référant à cette échelle, il apparaît qu'une partie des erreurs concernent des partis proches sur cette échelle. Mais on observe également des confusions entre les deux « gros » partis que sont le PSE et le PPE-DE, plus dans le sens PPE-DE pris pour des PSE que l'inverse. Enfin, les confusions entre partis situés aux extrémités de notre échelle sont fortement asymétriques : les interventions des Verts/ALE ont rarement été prises pour des interventions des PPE-DE (maximum 5%), en revanche les interventions des PPE-DE ont plus souvent été prises pour des interventions des Verts/ALE (30% maximum).

6.3 Résultats

Cette tâche de détermination du parti politique auquel appartient l'orateur d'une intervention s'est révélée difficile, tant pour les testeurs humains que pour les participants au défi.

Trois équipes ont initialement fait part de leur intention de participer à cette tâche. Seule l'équipe de l'Université de Montréal (D. Forest et al.) a finalement soumis des fichiers de résultats, au nombre de trois. Cette participation repose uniquement sur le corpus en français. Le tableau 10 met en évidence les valeurs de rappel et précision obtenues par cette équipe pour chaque parti politique ainsi que la F-mesure de chacune des soumissions.

Soumission	**Type de valeurs**	**ELDR**	**GUE-NGL**	**PPE-DE**	**PSE**	**Verts/ALE**	**F-mesure**
Nombre de documents attendus		1338	1794	4571	3626	1585	
1	Rappel	0,189	0,393	0,437	0,360	0,233	0,320
	Précision	0,210	0,345	0,447	0,365	0,226	
2	Rappel	0,231	0,332	0,498	0,394	0,207	0,339
	Précision	0,236	0,422	0,452	0,370	0,252	
3	Rappel	0,202	0,376	0,462	0,383	0,243	0,334
	Précision	0,205	0,384	0,462	0,369	0,255	

TAB. 10 – Rappels et précisions obtenus par parti politique pour chacune des trois soumissions de la tâche 3 par l'équipe de Montréal.

Des partis mieux identifiés que d'autres. Nous pouvons constater que les valeurs de rappel et précision augmentent avec le nombre de documents attendus. Plus un parti aura eu de documents disponibles, plus l'apprentissage aura été efficace, et en conséquence meilleurs auront été les résultats du corpus de test. Cette tendance se vérifie sur les trois soumissions pour lesquelles les valeurs de rappel et précision les plus élevées se rapportent aux partis PPE-DE (4571 documents), PSE (3626 documents) et GUE-NGL (1794 documents).

Nous avions observé chez les évaluateurs humains de cette tâche des taux de rappel et précision plus élevés pour les deux principaux partis que sont le PPE-DE et le PSE que pour les autres partis (voir tableau 11).

Juges / Parti	1	2	3	4	5	6
ELDR	0,33/0,42	0,27/0,25	0,11/0,17	0,31/0,42	0,33/0,33	0,18/0,17
PPE-DE	0,57/0,40	0,37/0,35	0,33/0,25	0,56/0,45	0,39/0,45	0,40/0,30
PSE	0,63/0,48	0,42/0,44	0,57/0,32	0,53/0,36	0,41/0,28	0,50/0,60
Verts/ALE	0,14/0,33	0,43/0,50	0,06/0,17	0,29/0,67	0,22/0,33	0,29/0,33

TAB. 11 – Valeurs de rappel et de précision (rappel/précision) obtenues par les six évaluateurs humains pour chaque parti politique.

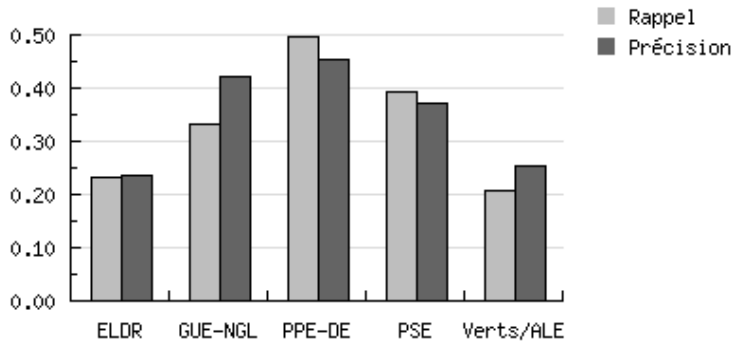

FIG. 5 – Valeurs de rappel et précision obtenues par parti sur la meilleure soumission (n° 2).

La figure 5 montre que les valeurs de rappel dépassent celles de la précision pour les deux principaux partis (PPE-DE et PSE), autrement dit ceux pour lesquels il y a eu plus de documents disponibles (tant dans le corpus d'apprentissage que dans celui de test), alors que c'est l'inverse pour les autres partis. Cela montre que l'attribution d'une intervention à ces partis a été préférentielle par rapport aux partis à plus faible effectif d'interventions.

7 Conclusion

L'édition 2009 du Défi Fouille de Textes a été pleine d'enseignements pour les organisateurs. L'évaluation humaine de la tâche a permis, comme d'habitude, de mieux cerner les problèmes, à défaut d'y trouver toujours des solutions.

À l'exception de la tâche 1 de classification de documents en *objectif* ou *subjectif*, les deux autres tâches se sont révélées difficiles et ont suscité peu de participations. Néanmoins, l'ensemble des données de référence sur les corpus de la tâche 2 de détection des passages subjectifs d'un texte nous semble devoir être intéressante pour la communauté concernée par ce type de problématique. Cet ensemble constitue une pré-annotation de corpus qui nous semble utile. À ce titre, les participants qui ont joué le jeu doivent en être remerciés.

Un autre aspect qui nous semble intéressant est le lien entre le nombre de passages subjectifs d'un texte et l'orientation subjective de ce texte. En effet, si l'abondance de passages clairement subjectifs suffit à donner l'impression que le texte est subjectif, cette abondance n'est cependant pas nécessaire d'après nos tests humains de la section 5.2.

Par ailleurs, les résultats médiocres de la tache 3 d'identification du parti politique d'un orateur montrent que, même dans un contexte parlementaire où les opinions sont supposées s'exprimer clairement, il n'est pas facile d'attribuer le bon parti à l'orateur. En soi, c'est un résultat intéressant.

8 Remerciements

Cet atelier bénéficie du soutien financier du projet CapDigital DoXa (traitement automatique des opinions et sentiments[13], convention DGE n° 08 2 93 0888). Nous exprimons notre gratitude envers le LIP6 (*Laboratoire d'Informatique de l'Université Paris 6*[14]) pour son soutien logistique.

Nous exprimons nos remerciements à la société ELDA (*Evaluations and Language resources Distribution Agency*[15]) pour son implication dans cette campagne d'évaluation au travers de la mise à disposition de ses corpus.

Nous remercions également les testeurs humains *(Arnaud, Béatrice, Cyril, Isabelle, Jean-Baptiste, Martine et Sarra)* qui ont bien voulu prendre un peu de leur temps pour tester les différentes tâches. Merci à Anne « la p'tite » pour la version italienne du site et à Jean-Baptiste pour la version anglaise.

Enfin, nous tenons à souligner combien nous avons apprécié la participation des différentes équipes de cette nouvelle édition, alors que cela représente une surcharge de travail conséquente et non financée dans le cadre de projets.

Références

Berthelin J.-B., Grouin C., Hurault-Plantet M. et Paroubek P. (2008). Human judgement as a parameter in evaluation campaigns. In *Coling 2008 : Proceedings of the workshop on Human Judgements in Computational Linguistics*, p. 17–23, Manchester, UK : Coling 2008 Organizing Committee.

Carletta J. (1996). Assessing agreement on classification tasks : the kappa statistics. *Computational Linguistics*, **2**(22), 249–254.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Nakache D. et Métais E. (2005). Evaluation : nouvelle approche avec juges. In *INFORSID*, p. 555–570, Grenoble.

Paroubek P., Robba I., Vilnat A. et Ayache C. (2008). EASY, Evaluation of Parsers of French : what are the Results ? In European Language Resources Association (ELRA), Ed., *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

[13] `http://www.projet-doxa.fr/`
[14] `http://www.lip6.fr`
[15] `http://www.elda.org/`

Session III - Catégorisation globale

Approche Multi-traces et catégorisation de textes avec Random Indexing

Yann Vigile Hoareau (1) & Adil El Ghali (2)

(1) CHArt – Université Paris 8
2, rue de la Liberté 93526 St Denis Cedex 02
vigilehoarau@gmail.com

(2) Edelweiss – INRIA
2004, route des Lucioles, 06902 Sophia Antipolis
adil.elghali@gmail.com

Résumé – Abstract

Nous présentons le travail réalisé dans le cadre de la participation à DEFT09 pour la tâche 1 en français, en anglais et en italien. L'approche envisagée consiste à appliquer les modèles utilisés en psychologie cognitive pour décrire la mémoire épisodique humaine à la catégorisation automatique de texte au moyen de *Word Vectors*. Nous détaillons un modèle de mémoire épisodique et l'utilisons pour fournir les hypothèses de travail concernant la réalisation des calculs de similarités impliqués dans la catégorisation de texte avec les *Word Vectors*. Le modèle de *Word Vector* utilisé n'est pas basé sur une approche statistique classique, mais relève des projections aléatoires. La chaine de traitements proposée est entièrement automatique.

The paper resumes the work realized in text-mining context DEFT09 for the task 1 in French, English and Italian. The approach consists in applying models used in cognitive psychology to describe human episodic memory on automatic text categorization using Word Vectors. A cogntive model of episodic memory will be detailled. This model provides working hypothesis used to drive the calculus of similarity involved in text categorization with Word Vectors. The Word Vector models used is not based on the classical statistic approach but on ramdon projection. The processing chain described is entirely automatic.

Mots-Clés – Keywords

Random Indexing ; Fouille de textes ; Approche cognitive ; Mémoire épisodique
Random Indexing; Text-mining; Cognitive approach; Episodic memory.

1 Introduction

Le modèle de l'Analyse de la Sémantique Latente (LSA) (Landauer & Dumais, 1997) est le plus connu de la famille de modèles à laquelle il appartient, la famille des *Word Vectors*. Les *Words vectors* ont pour caractéristique de représenter dans un espace vectoriel à grande dimension, la similarité entre mots ou concepts. Pour ce faire, ces modèles utilisent différentes méthodes qui permettent de compter les mots en prenant en compte l'environnement textuel dans lequel ils apparaissent pour construire une matrice qui renseigne sur la co-occurrence des mots dans des contextes donnés, typiquement des documents ou des paragraphes. Ces modèles reposent sur l'hypothèse distributionnelle selon laquelle des mots qui ont un sens similaire apparaissent dans des contextes similaires.

Les *Words Vectors* ont toujours montré une grande aptitude pour la catégorisation thématique de documents. Dans sa première version, LSA, appelé alors « *Latent Semantic Indexing* » (LSI) (Deerwester et *al*, 1990) était proposé spécifiquement pour l'indexation de documents textuels. La catégorisation thématique de textes au moyen des *Words Vectors* repose sur l'application directe de l'hypothèse distributionnelle : des mots qui apparaissent dans des *contextes similaires* partagent des *thématiques similaires*.

Dans le cas de la catégorisation de textes en fonction de l'opinion qu'ils expriment, l'application directe de l'hypothèse distributionnelle montre ses limites. Si l'on prend l'exemple du Défi de Fouille de Texte 2007 (DEFT'07), des articles exprimant une opinion sur des films, des jeux-vidéo ou encore des articles scientifiques devaient être catégorisés en fonction de l'avis qu'ils exprimaient : positif, neutre ou négatif. Dans une telle situation, l'application directe de l'hypothèse distributionnelle n'est pas possible, car les films qui ont reçu une bonne critique n'ont évidemment pas en commun une thématique singulière, particulière. La chose paraît évidente pour les articles scientifiques : il n'y a pas un domaine en particulier qui recevrait des avis positifs et un autre qui n'en recevrait que de négatifs. Autant l'hypothèse distributionnelle favorise les *Word Vectors* pour la catégorisation thématique, autant son application directe fait problème pour la catégorisation de jugement d'opinion.

L'algorithme que nous avons mis en œuvre a pour objectif de pallier les limites de l'application directe de l'hypothèse distributionnelle. Cet algorithme est dérivé de la recherche en psychologie cognitive sur la mémoire épisodique. Il est couplé à un modèle de *Word Vector* appelé *Random Indexing*. Dans la première partie, nous détaillons un modèle de mémoire épisodique comptant parmi les plus fameux de la littérature ; le modèle *MINERVA* 2 (Hintzman, 1986 ;1988). Puis nous exposons les éléments de *MINERVA 2* qui servent de base à l'algorithme de catégorisation de texte. Dans la deuxième partie, nous présentons l'algorithme implémenté, ainsi que le modèle *Random Indexing* (Kanerva et *al* 1998). Dans la troisième partie, nous présentons les résultats des exécutions de la tâche 1 pour les trois langues.

1.1 Un modèle de mémoire épisodique : MINERVA 2

L'intuition sous-jacente aux modèles Multi-Traces est qu'au cours de sa vie, un individu ne rencontre que des exemplaires des objets qui composent le monde et cela, à travers des épisodes discrets. C'est à partir de ces exemplaires isolés que les catégories et les concepts se

structurent. Ainsi, aucun individu ne rencontre jamais le concept de beauté, mais seulement des exemplaires successifs de choses considérées comme belles. Les modèles Multi-traces considèrent que chaque événement de la vie d'un système de mémoire est stocké en mémoire et que chaque événement est appréhendé en fonction de l'ensemble des expériences précédentes disponibles en mémoire.

Le modèle *MINERVA 2* (Hintzman, 1984 ;1986) compte parmi les plus fameux des modèles Multi-Traces. Selon ce modèle, chaque événement ou épisode de la vie du système de mémoire est représenté et stocké sous la forme d'un vecteur à D dimensions dont les valeurs peuvent être 0, (+1) ou (-1). Le modèle a fortement contribué à l'avancement de l'étude de l'effet de fréquence des épisodes sur la capacité mnésique dans des tâches de rappel libre, de rappel indicé, de reconnaissance d'items.

Pour rendre compte de l'activation de la mémoire par un épisode nouveau, appelé sonde, *MINERVA 2* met en œuvre un processus à deux étapes. Dans la première étape, un calcul de similarité est réalisé entre le vecteur-sonde et chaque vecteur-épisode stocké en mémoire (voir Eq 1).

$$S_i = \sum_{j=1}^{N} \frac{P_j T_{i,j}}{N_i}$$

Eq 1 Similirarité d'une trace i, où P_j est la valeur de la coordonnée j de la sonde, et $T_{i,j}$ la valeur de la coordonnée j dans la trace i

Les épisodes les plus similaires à la sonde seront affectés d'une valeur d'activation plus importante que les épisodes qui sont les moins similaires. Dans la deuxième étape, un calcul est réalisé à partir des traits de chaque épisode. Le calcul consiste à reconstituer un vecteur « écho » qui hérite des traits des épisodes ayant précédemment bénéficié des valeurs d'activation les plus élevées, y compris les traits qui n'existaient pas dans la sonde. L' « écho » dispose de deux composants. Le premier composant est l'intensité, appelée *I* (voir Eq 2).

$$\mathbf{I} = \sum_{i=1}^{M} A_i, \text{ où } \quad A_i = S_i^3$$

Eq 2 Intensité de l' « echo »

Le deuxième composant est le contenu. Il est obtenu par la sommation de toutes les traces de la mémoire pondérée par valeur d'activation (voir Eq 3). Le processus d'abstraction réalisé par « echo » est qualifié par Rousset (2000) de « re-création ».

$$C_j = \sum_{i=1}^{M} A_i T_{i,j}$$

Eq 3 Contenu de l' « echo »

1.2 Effet de Fréquence

L'étude de l'effet des différences fréquence des épisodes sur l'intensité de l'écho a montré que lorsqu'une sonde est similaire à de nombreux épisodes stockés en mémoire, alors l'intensité de l'écho est élevé et inversement. Lorsqu'il y a peu d'épisodes en mémoire qui sont similaires à la sonde, alors l'intensité de l'écho est faible. On peut considérer l'intensité de l'écho comme un indicateur de la familiarité de la sonde pour la mémoire. Pour cette raison, le calcul d'écho de *MINERVA 2* a constitué une heuristique pour la mise en place des algorithmes que nous décrivons par la suite.

En considérant, d'une part, la limite décrite plus avant de l'application de l'hypothèse distributionnelle pour la catégorisation de texte d'opinion et, d'autre part, le calcul d'écho de *MINERVA 2*, nous avons transféré une partie du cadre théorique de *MINERVA 2* pour raisonner sur la phase de catégorisation de texte en utilisant le paradigme de la mémoire épisodique. Le raisonnement est le suivant :

1. Il faut se représenter un texte que l'on souhaite catégoriser comme un vecteur-sonde dans le paradigme de la mémoire épisodique. Si cette sonde est d'une catégorie A et qu'elle est comparée à une mémoire épisodique qui regroupe toute les exemplaires de la catégorie A, alors d'après *MINERVA 2*, l'intensité de l' « écho » serait fort. De même, si on compare cette sonde de la catégorie A à une mémoire épisodique qui regroupe tous les exemplaires de la catégorie B, alors l'intensité de l' « écho » sera faible.

2. En poursuivant ce raisonnement, pour approcher le calcul du contenu du vecteur « écho », tous les épisodes (ie, les documents) appartenant à une même catégorie sont sommés au sein d'un même vecteur, appelé vecteur-cible. Pour approcher l'Intensité de l'écho, la sonde est comparée au vecteur « écho » au moyen du calcul du cosinus de l'angle formée par le vecteur-sonde et le vecteur-cible.

3. Le fait de regrouper les textes appartenant à la même catégorie permet d'accroître la saillance du contenu de l'écho de la mémoire ainsi constituée. Les mémoires épisodiques correspondant à chaque catégorie de textes sont homogènes. La catégorie attribuée à une sonde est celle qui correspond à la mémoire qui délivre l'écho le plus fort intense.

1.3 Effet de typicalité des épisodes

Les recherches sur l'activité de catégorisation ont mis en évidence que tous les exemplaires d'une catégorie ne sont pas équivalents et que certains sont plus typiques de la catégorie (Rosh & Mervis 1975 ; Cordier & Tijus, 2000). On définit la typicalité d'un exemplaire pour une catégorie par une forte similarité avec les autres exemplaires appartenant à la même catégorie et une faible similarité avec les exemplaires appartenant à d'autres catégories.

En intégrant le principe de typicalité à la construction des vecteurs-cibles précédemment décrits, nous pouvons considérer que pour une catégorie d'opinion donnée, il

existe plusieurs expressions possibles. Tel qu'il est décrit dans la section précédente, le mode de calcul des vecteurs-cibles les rend sensibles pour la détection des documents les plus typiques. Nous faisons cependant l'hypothèse qu'une catégorie d'opinion est diffuse et que la prise en compte des seuls exemplaires les plus typiques conduirait à rejeter un nombre important de documents qui ne présenteraient pas les critères de typicalité, mais qui appartiendraient effectivement à la catégorie[1].

Pour remédier au problème de rejet de la catégorie pour cause de défaut de typicalité, pour chaque catégorie, nous avons regroupé les exemplaires d'une catégorie en fonction de leur degré de typicalité : à partir des vecteurs-cibles correspondant à chaque catégorie, des sous-vecteurs-cibles homogènes ont été constitués. La méthode retenue pour regrouper les exemplaires les plus similaires a consisté à comparer la similarité de chaque vecteur-exemplaire qui compose un vecteur-cible au vecteur-cible. Les vecteurs-exemplaires sont par suite ordonnée en fonction de leur similarité avec le vecteur-cible. Une partition en *P* éléments de taille identique est réalisée pour réaliser *P* sous-vecteur-cibles homogènes.

1.4 Implémentation des vecteurs-cibles et *des* sous-*vecteurs*-cibles

Les modèles de *Words Vectors* constituent un cadre tout à fait adéquat pour l'implémentation des vecteurs-cibles et des sous-vecteurs-cibles. C'est à partir des représentations vectorielles des mots et des documents issus de la construction d'un espace sémantique que seront construits les vecteurs-cibles et les sous-vecteurs-cibles. Premièrement, nous décrivons le modèle de *Word Vector* auquel nous avons eu recourt. Il est s'agit de *Random Indexing*. Deuxièmement, nous présentons la méthode de construction des vecteurs-cibles et des sous-vecteurs-cibles à partir d'un espace sémantique. Troisièmement, nous présentons la méthode de catégorisation d'un vecteur-sonde à partir des sous-vecteurs-cibles de chacune des catégories.

1.4.1 Un modèle de Word Vector : Random Indexing

Si, l'hypothèse distributionnelle permet de proposer une piste pour résoudre la question de la similarité sémantique entre les mots, le nouveau problème qu'elle met à jour est celui de la circularité entre *mot* et *contexte*. Pour résoudre la question de la circularité, la plupart des modèles de *Word Vectors* font appel à des méthodes statistiques lourdes et complexes, comme c'est le cas pour la Décomposition en Valeurs Singulières avec LSA. Le modèle de *Word Vector* que nous avons utilisé dans le cadre de ce concours est appelé *Random Indexing (RI)* (Kanerva et *al* 1998). Il ne s'appuie pas sur les méthodes mathématiques habituelles de réduction (e.g. calcul de valeurs singulières dans LSA) , mais sur des méthodes de projection aléatoire.

Les *Word Vectors* ont en commun un certains nombre de principes que l'on peut résumer ainsi :

---

[1] Nous serions très satisfait de la capacité des vecteurs-cibles de détecter les exemplaires typiques si cette propriété ne risquait pas se révéler contre-productive par la conséquence du rejet d'un grand nombre de candidats à une catégorie donnée, sous le prétexte qu'ils n'auraient pas « le look de l'emploi ».

- Ils sont basés sur l'hypothèse distributionnelle

- Ils disposent d'une méthode de comptage des mots dans un contexte donné.

- Ils disposent d'une méthode d'abstraction de la signification des mots qui s'appuie sur la prise en compte des contextes dans lesquels ces derniers apparaissent.

- Ils utilisent une méthode de représentation vectorielle pour stocker, puis manipuler le signification des mots.

Le cas de RI est un peu particulier dans la famille des *Word Vectors*. En effet, dans les autres modèles, on peut dire que l'ordre dans lequel nous avons énoncé les principes qui régissent les modèles correspond par ailleurs aux différentes étape de construction des espaces sémantiques. Ce n'est pas du tout le cas pour RI. La méthode de construction d'un espace sémantique avec RI est la suivante :

- Créer une matrice A (*d* x *N*), contenant des vecteurs-Indexes, où *d* est le nombre de documents ou de contextes correspondant au corpus et *N,* le nombre de dimensions (*N* > 1000!) défini par l'expérimentateur. Les Vecteurs Index sont creux et aléatoirement généré. Il consiste en un petit nombre de (+1) et de (-1) et de centaines de 0.

- Créer une matrice B (*t* x *N*) contenant les vecteurs-Termes, où *t* est le nombre de termes différents dans le corpus. Pour commencer la compilation de l'espace, les valeurs des cellules doivent être initialisées à 0.

- Parcourir chaque document du corpus. À chaque fois qu'un terme *t* apparaît dans un document *d,* il faut *accumuler* le *vecteur-index* correspondant au document *d* au *vecteur-terme* correspondant au terme *t.*

À la fin du processus, les *vecteurs-termes* qui sont apparus dans des contextes (ou documents) similaires, auront accumulé des *vecteurs-index* similaires. Ainsi, RI ne fait appel à aucune méthode statistique telle la SVD ou l'analyse de régression (comme c'est le cas d'autres modèles de *Word Vector*) pour réaliser le processus d'abstraction de la signification des mots.

RI dispose par ailleurs d'une option d'apprentissage par cycle. Lorsque tous les documents qui composent le corpus ont été parcourus, la matrice B contient tous les vecteurs-termes finaux. Une matrice A' (*d'* x *N*), avec *d = d'* peut de nouveau être construite, non plus à partir de la génération aléatoire comme cela fut le cas initialement pour A, mais à partir des vecteurs-terms finaux de la matrice B. Le nombre de cycle d'apprentissage est un paramètre du modèle. Le processus d'apprentissage de RI est comparable à ceux décrits dans les approches connexionnistes pour les réseaux de neurones.

Le modèle a démontré des performances aussi convaincantes (Kanerva et *al*, 2000) si ce n'est plus convaincante (Karlgren and Sahlgren, 2001) que LSA pour le test de synonymie du TOEFL (Landaeur & Dumais, 1997).

1.4.2 Construction des vecteurs-cibles et des sous-vecteurs-cibles

Après avoir construit un espace sémantique avec *Random Indexing* à partir de l'ensemble des documents appartenant à toutes les catégories,

- Chaque mot du corpus est représenté par un vecteur à D-dimensions.
- Un vecteur-document est constitué de la sommation des vecteurs-mots dont il est composé.
- Un vecteur-cible est constitué pour chaque catégorie à partir de la sommation des vecteurs-documents qui composent la catégorie.
- Les *P* sous-vecteurs-cibles sont obtenus à partir de la partition de la liste ordonnée par ordre de similarité décroissante de chaque vecteur-documents qui compose une catégorie avec le vecteur-cible de la catégorie (voir la section1.3).

1.4.3 Catégorisation d'opinion avec des sous-vecteurs-cibles

Un fois les sous-vecteurs-cibles constitués pour chaque catégorie, il s'agit de proposer une méthodes pour attribuer une catégorie à un vecteur-sonde (un document que l'on doit catégoriser). Si l'on considère qu'il y a *p* sous-vecteur-cibles pour chacune de *C* catégorie. La méthode que nous avons retenue consiste :

1. à comparer la similarité entre le vecteur-sonde et les *p* sous-vecteurs-cibles de chaque catégorie, pour chacune des *C* catégories.
2. à attribuer à la sonde la catégorie dont les sous-vecteurs-cibles sont les plus similaires.

2 Déroulement du DEFT'09

2.1 Apprentissage

La première étape à consisté à compiler les espaces sémantiques pour les corpus en français, en anglais et en italien. La deuxième étape a été de construire les vecteurs-cibles, puis les sous-vecteurs-cibles.

Afin de pouvoir pré-tester et optimiser les différents paramètres intervenant à différentes étapes du processus de catégorisation, nous avons réservé 10% du corpus d'apprentissage pour les pré-tests.

Les paramètres sur lesquels a porté le travail d'optimisation sont le nombre du dimensions *d* et le nombre de cycles *N* pour *Random Indexing*, ainsi que le nombre de sous-vecteurs-cibles *P*. Les paramètres optimum d'après nos pré-tests sont *d*=5000, *N*= 50, *P*= 5. Ces paramètres ont été retenus pour la phase de test de la tâche 1 pour les trois langues.

Enfin, la chaîne complète de traitement entièrement automatique a été programmée en Java, la

librairie SemanticVectors[2] fournissant une implémentation efficace de RI en Java. Elle comprend la compilation de l'espace avec SemanticVectors, les constructions des P sous-vecteur-cibles pour les C catégories et l'attribution d'une catégorie à chaque document du corpus de test.

2.2 Tests et résultats.

Nous avons compilé l'espace sémantique de test à partir du corpus d'apprentissage et du corpus de test. Les sous-vecteurs-cibles ont été réalisés à partir des documents étiquetés issus du corpus d'apprentissage, mais recalculés à partir de la métrique propre à l'espace sémantique de test. Le résumé des exécutions pour la tâche 1 en français, anglais et italien figure dans le tableau I.

		Français		Anglais		Italien	
Nombre de documents	Appr.	25176		7866		1496	
	Test	16788		5245		999	
Taille (Ko) ~	Appr.	80000		25000		6000	
	Test	51000		16000		3500	
Nombre de dimensions		5120		4096		4096	
Nombre de Cycles		50		40		40	
Nombre de sous-vecteurs-cibles		5		5		5	
précision	Obj	0.740	0.941	0.720	0.515	0.710	0.828
	Subj		0.540		0.925		0.591
rappel	Obj	0.803	0.869	0.636	0.967	0.723	0.681
	Subj		0.738		0.306		0.765
F-mesure		0.771		0.676		0.716	

Figure 1 : Descriptions des corpora, des paramètres et des performances des exécutions de la tâche 1 en trois langues

[2] http://code.google.com/p/semanticvectors/

3 Discussion

La chaîne de traitement que nous avons proposée dans le cadre de DEFT'09 provient de l'utilisation et du transfert de la métaphore de la mémoire épisodique telle qu'elle est modélisée en psychologie cognitive. L'algorithme de catégorisation proposé se couple à un modèle vectoriel de représentation des connaissances. Comme nous l'avons énoncé plus haut, l'application de l'hypothèse distributionnelle classique pour un *Word Vector* dans une tâche de catégorisation de textes exprimant des opinions n'est pas envisageable. Les résultats obtenus avec l'algorithme de catégorisation démontrent cependant que les limites de l'hypothèse distributionnelle ont été dépassées. La capacité de dépasser les limites de l'hypothèse distributionnelle n'est pas à mettre sur le compte de quelconques paramétrages des différentes étapes de traitement : aucun pré-traitement n'a été réalisé sur le corpus.

La capacité pour un modèle de *Word Vector* de performer au-delà de la limite de l'hypothèse distributionnelle provient des traitements qui ont été réalisés lors de la construction des vecteurs-cibles et des sous-vecteurs-cibles. *Random Indexing* est ici utilisé comme un moyen particulièrement efficace de représenter les relations sémantiques entre les mots et les documents. Il n'est pas utilisé pour catégoriser les textes, car nous pensons qu'il n'a pas vocation à cela.

Inversement, l'algorithme que nous avons proposé n'a pas vocation à exprimer les relations sémantiques entre les mots, mais il a vocation à catégoriser des textes en se couplant à un modèle de représentation de la similarité entre les mots. Alors que *MINERVA 2* traite de l'information sub-symbolique, l'algorithme qui s'en inspire est capable de traiter de l'information symbolique d'un niveau d'abstraction très élevé, comme le montre sa capacité à reconnaître le caractère objectif ou subjectif d'un document.

Remerciements

Nous remercions tous les membres du projet Edelweiss et du laboratoire CHArt ainsi Dominique Widdows, le responsable du projet Semantic Vectors. Nous remercions particulièrement Charles Tijus, Denis Legros et Axel Gauvin pour leur soutien.

Références

Cordier F., Tijus C. (2001), Object properties: A typology. *Current Psychology of Cognition,* 20, 445-472

Hintzman, D. L. (1984), MINERVA 2: A simulation of human memory. *Behavior Research Methods, Instruments, & Computers*, 16, 96-101.

Hintzman, D. L. (1986), Schema abstraction in a multi-trace memory model. *Psychological Review,* 93, 411-428.

Karlgren J., Sahlgren M. (2001), From Words to Understanding, In Y. Uesaka, P. Kanerva, & H. Asoh (Eds.) *Foundations of Real-World Intelligence*, CSLI Publications, Stanford.

Landauer, T. K., Dumais, S. T. (1997), A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

Sahlgren M. (2006), The Word-Space Model: Using distributional analysis to represent syntagmatic and

paradigmatic relations between words in high-dimensional vector spaces. *Ph.D. dissertation*, Department of Linguistics, Stockholm University.

Sahlgren M., Cöster R. (2004), Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization, *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva.

Rosch E. H., Mervis C. B. (1975), Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology,* 7, 573-605.

Rousset, S. (2000), Les conceptions "système unique" de la mémoire: aspect théorique. *Revue de neuropsychologie*, 10(1), 30-56.

Un niveau de base pour la tâche 1 (corpus français et anglais) de DEFT'09

Yves Bestgen (1) et Guy Lories (2)

(1) PSOR/CECL – Université catholique de Louvain
Place du cardinal Mercier, 10 B-1348 Louvain-la-Neuve Belgique
yves.bestgen@psp.ucl.ac.be
(2) ECSA – Université catholique de Louvain
Place du cardinal Mercier, 10 B-1348 Louvain-la-Neuve Belgique
guy.lories@uclouvain.be

Résumé – Abstract

L'objectif de cette recherche était d'évaluer l'efficacité d'un classifieur simple de type SVM pour réaliser la première tâche de DEFT'09 sur les corpus français et anglais. Cette approche a été sélectionnée sur la base des informations disponibles à propos des principes qui ont gouverné l'affectation initiale des documents aux catégories objectives et subjectives et sur la base d'une analyse exploratoire des corpus. Une série de tentatives d'optimisation du classifieur n'ayant apporté que des gains négligeables, les performances obtenues peuvent être considérées comme des niveaux de base permettant de se faire une idée de la difficulté de la tâche.

The research goal was to assess the performance of a simple SVM classifier in task 1 of the DEFT09 competition on the french and english corpora. We were led to this approach by the information available regarding the initial categorization and by an exploratory analysis of the corpora. As various attempts to improve the classifier performance brought only negligible improvements the performance obtained can be considered as a base level assessment of task difficulty. A potentially original contribution stems from using discretized document length to bring about a slight performance improvement with the French corpus.

Mots-Clés – Keywords

Catégorisation de textes, Machines à support vectoriel, Discrétisation.
Text categorization, Support Vector Machines, Discretization.

1 Introduction

Le thème de DEFT'09, cinquième édition de la campagne d'évaluation en fouille de textes DEFT, est l'analyse multilingue d'opinion. Trois tâches étaient proposées dans trois langues : le français, l'anglais et l'italien. La première tâche se présentait comme un problème de détection du caractère objectif ou subjectif global d'un article de journal. La deuxième tâche visait à identifier les passages d'un texte qui sont subjectifs, par opposition à ceux qui sont objectifs. La troisième tâche exigeait l'identification du parti politique auquel appartient l'orateur d'une intervention dans le cadre de débats au parlement européen.

Nous avons choisi de nous concentrer sur la première tâche pour les corpus français et anglais. La sélection d'une procédure pour réaliser cette tâche dépend nécessairement de la manière dont la catégorisation de référence a été effectuée. Le tableau 1 reprend la totalité de l'information disponible à ce sujet. Une étape préliminaire d'analyse exploratoire des données était dès lors plus que

souhaitable. Les observations résultant de cette première étape sont présentées à la deuxième section de ce rapport. Sur la base de celles-ci, une technique classique, relativement opaque, mais bien connue pour son efficacité dans le cadre de la catégorisation supervisée de texte (Burges 1998 ; Joachims, 2002), a été choisie (section 3) et des essais ont été menés afin d'optimiser pour le corpus français une série de paramètres (section 4). La combinaison de paramètres sélectionnée est présentée à la section 5 et l'impact de chaque paramètre, considéré indépendamment, y est évalué sur le corpus de test. L'ensemble des résultats qui nous ont été transmis par les organisateurs est donné à la section 6. La conclusion synthétise ce que nous avons appris en participant à DEFT'09.

"*Détection du caractère objectif/subjectif global d'un texte :*

Le but de cette tâche est de pouvoir détecter si un texte est plutôt un texte d'opinion, subjectif, (comme une critique de film, ou un éditorial) ou plutôt un texte factuel, objectif, (comme une dépêche d'agence, ou des actualités). Nous nous en tenons ici strictement à l'explicite d'un texte, sans tenir compte de ce qui peut être sous-entendu, implicite."

"*Pour les tâches 1 et 3, les participants disposeront donc de références pouvant donner lieu à un apprentissage.*"

"*L'attribution des valeurs « objective » et « subjective » aux articles a été réalisée de manière différente selon les journaux* [suivent quelques exemples]"

Tableau 1 : Informations disponibles à propos de la catégorisation initiale, extraites du site web de DEFT'09.

2 Pré-traitement et analyse exploratoire

L'ensemble des traitements des textes a été mené en SAS (Statistical Analysis System, V6.12 sous MacOS 9 et V9.1 sous Windows 2000), section Base et Statistics, après prétraitement du corpus au moyen de TreeTagger (Schmidt, 1994). Il s'ensuit que, dans toutes les analyses rapportées ici, c'est la segmentation en termes (token) effectuée par Tree-Tagger qui a produit les descripteurs des documents. Il s'agit donc de mots ("a", "L'", "CNRS"), mais aussi de nombres ("2004", "3.5"), de symboles ("$", "&") et de signes de ponctuation (".", "?").

Les premières analyses ont porté sur la longueur des documents dans le corpus français. Il est immédiatement apparu que deux textes étaient anormalement courts (1 et 3 termes) et un autre anormalement long (plus de 90 000 termes, correspondant à une édition du journal). Il a été décidé de supprimer ces trois documents du corpus d'apprentissage, la catégorisation en objectif/subjectif d'au moins deux de ceux-ci étant sujette à caution.

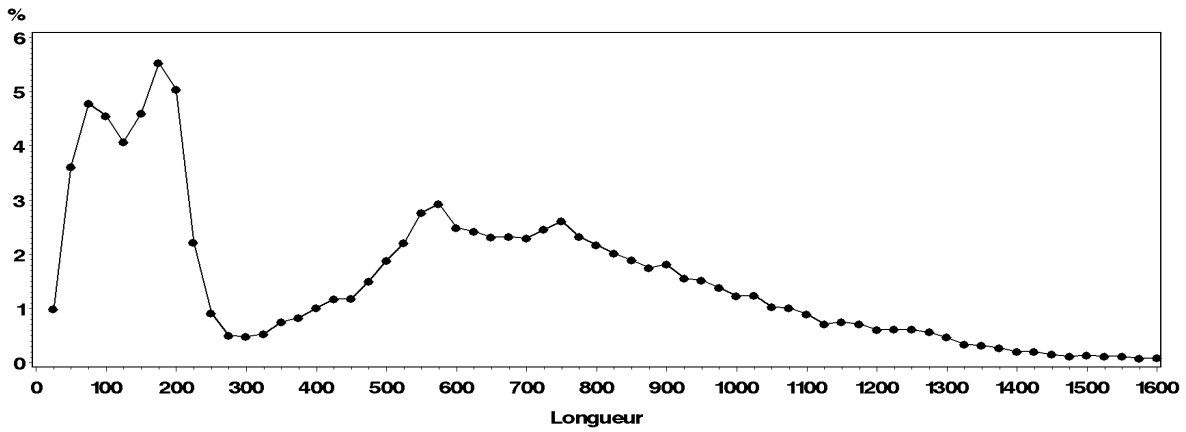

Figure 1 : Distribution des longueurs des documents pour le corpus français : pourcentage sur le nombre total de documents en ordonnée, longueur en nombre de termes en abscisse

La distribution[1] des longueurs des documents est donnée à la figure 1. Comme on peut le voir, cette distribution est nettement multimodale, ce qui n'est pas très étonnant puisque les longueurs des articles de journaux dépendent fréquemment de multiples contraintes tant éditoriales que matérielles. On note aussi une proportion relativement importante de textes courts (moins de 100 termes).

Une analyse détaillée des textes les plus courts montre que ceux-ci correspondent pour une part à des *erratas*, systématiquement considérés comme subjectifs et ce y compris lorsque l'erreur est factuelle et le document initial considéré comme objectif ainsi que l'atteste le tableau 1.

```
<doc id="1_fr:4528">
   <PROPRIETE valeur="SUBJECTIF" confiance="1" />
  <texte>
   <p>Dans l'article intitulé « Anschluss : l'Eglise d'Autriche fait son autocritique », le titre du livre d'Irene Harand publié en 1935 est Son combat, « réponse à Hitler », et non Mon combat, la réponse à Hitler, comme nous l'avons écrit par erreur.</p>
   <p>Par ailleurs, l'orthographe du cardinal Theodor Innitzer était erronée : il s'agit bien du cardinal Innitzer et non Innitzert.</p>
  </texte>

<doc id="1_fr:16181">
   <PROPRIETE valeur="OBJECTIF" confiance="1" />
  <texte>
   <p>Le palais de l'archevêché de Vienne accueille, samedi 12 mars - jour anniversaire de l'Anschluss, l'annexion par l'Allemagne nazie, en 1938 -, un événement que l'Eglise d'Autriche qualifie d' « historique ». Des religieux, artistes, écrivains et scientifiques doivent y lire à haute voix, pendant douze heures, le livre publié en 1935 par la militante catholique Irene Harand, l'une des rares à avoir dénoncé, à l'époque, l'antisémitisme et l'idéologie nationale-socialiste. Intitulé Mon combat, la réponse à Hitler, ce texte oublié vient d'être réimprimé avec un commentaire sans ambiguïté de l'archevêque de Vienne, le cardinal Christoph Schönborn : « Etre chrétien et antisémite sont deux positions inconciliables. »</p>
...
... Mais elle est aussi une critique indirecte des autorités ecclésiastiques sous le nazisme, coupables, au mieux, d'aveuglement, tel le cardinal Theodor Innitzert, au pire, de complicité avec le régime.
...
```

Tableau 1 : Document du type "Errata" catégorisé comme subjectif et article original catégorisé comme objectif

On notera aussi que des erreurs commises par d'autres entités que le journal lui-même sont considérées comme objectives (voir tableau 2).

```
<doc id="1_fr:4207">
   <PROPRIETE valeur="OBJECTIF" confiance="1" />
  <texte>
   <p>Une demande visant à ce que leur pays soit retiré de la liste des membres de la coalition pour la guerre en Irak a été formulée par les autorités slovènes. Le premier ministre, Anton Rop, avait réclamé, la semaine dernière, des explications à Washington, où le département d'Etat avait finalement reconnu qu'il avait cité par erreur la Slovénie comme faisant partie de la liste actuelle des 49 Etats qui soutiennent l'intervention armée.</p>
  </texte>
```

Tableau 2 : Document "Objectif" rapportant une erreur non commise par *Le Monde*

[1] Les distributions de longueur présentées dans ce rapport ont été obtenues en agrégeant les valeurs brutes par tranche de 25 termes et en censurant la distribution vers 1600 termes. Au-delà, les fréquences deviennent très faibles.

On trouve aussi parmi les textes très courts des légendes d'illustrations (tableau 3).

```
<doc id="1_fr:15946">
  <PROPRIETE valeur="SUBJECTIF" confiance="1" />
 <texte>
  <p>Paru dans « El Imparcial »cartoons@courrierinternational.com</p>
 </texte>
```

Tableau 3 : Document "Subjectif" donnant la référence d'une illustration

L'analyse de la longueur des documents pour l'anglais ne met pas en évidence de documents anormalement longs ou courts. La distribution des longueurs, donnée à la figure 2, se distingue de celle obtenue pour le corpus français (figure 1) par la présence d'un seul pic manifeste aux alentours de 300.

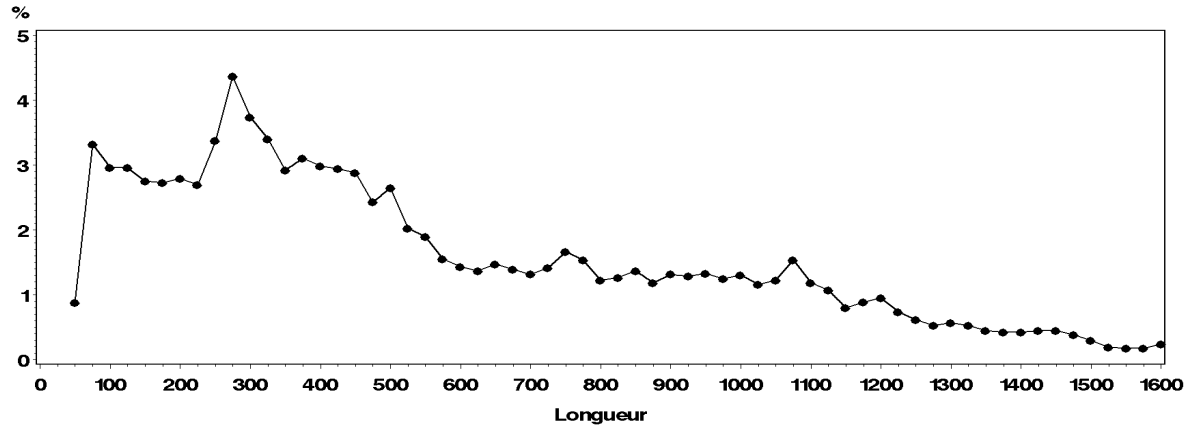

Figure 1 : Distribution des longueurs des documents pour le corpus anglais : pourcentage sur le nombre total de documents en ordonnée, longueur en nombre de termes en abscisse

Une analyse des textes courts du corpus anglais montre que ceux-ci correspondent en partie à des lettres de lecteurs qui sont une fois sur deux considérées comme objectives[2] (voir l'exemple donné dans le tableau 4). Cette affectation fréquente à la catégorie objective, qui ne se retrouve pas dans le corpus français, ne s'accorde pas aisément avec les conceptions courantes dans ce champ de recherche (Hurault-Plantet, 09.04.2009, dia n°9).

```
<doc id="1_en:832">
  <PROPRIETE valeur="OBJECTIF" confiance="1" />
 <texte>
  <p>Sir, I was astonished to read the comments by Professor Patrick Minford in Peter Marsh's article 'Salmon succeeds beer and sandwiches' (May 24). If an adviser to the Treasury can state, 'People from business are invariably a nuisance when it comes to talking about the economy', it explains much about the malign neglect and ignorance which our manufacturers have had to endure for many years.</p>
  <p>Does the professor believe that wealth is created from thin air, and where, if not from taxed wealth, does this bemused academic's salary come from?</p>
  <p>Campbell Dunford,</p>
  <p>chairman,</p>
 ...
```

Tableau 4 : Lettre d'un lecteur catégorisée comme objective

[2] Sur la base de l'analyse des 40 premiers documents du corpus d'apprentissage commençant par "<p>Sir, ".

Au vu de l'information disponible concernant les principes qui ont gouverné l'affectation initiale des documents aux deux catégories et des observations rapportées ci-dessus, il nous a semblé judicieux de nous fixer pour objectif d'évaluer le niveau de performance que peut atteindre une technique de catégorisation simple et relativement opaque (les machines à support vectoriel). L'intérêt de ce travail se limite donc à fournir un niveau de base auquel d'autres approches, plus informées, pourront être comparées.

3 Machines à support vectoriel

L'algorithme d'apprentissage utilisé est une machine à support vectoriel (SVM). De manière générale un tel algorithme apprend à classer un ensemble de vecteurs de R^n en deux catégories. Il s'applique cependant également à des vecteurs à composants binaires. Ici les vecteurs représentent un ensemble de propriétés des textes: longueur, présence ou fréquence éventuellement transformée de tel ou tel mot, appartenance à une catégorie déterminée par ailleurs etc... Il en existe une version linéaire et diverses versions « non linéaires » (à noyau).

L'algorithme *linéaire* catégorise les vecteurs en identifiant un hyper-plan qui sépare, si possible parfaitement, les exemples positifs des exemples négatifs. Il s'agit donc d'une technique qui peut évoquer les techniques linéaires traditionnelles comme l'analyse discriminante cependant l'hyper-plan est défini et calculé de manière particulière.

Dans le cas de *parfaite séparation*, l'hyper-plan de séparation passe entre les points positifs et négatifs les plus proches les uns des autres et il est choisi de sorte que la marge de séparation soit la plus grande possible. La figure 3 représente un tel cas en deux dimensions; la marge maximisée est la distance du plan aux points les plus proches, placés sur les droites en pointillé.

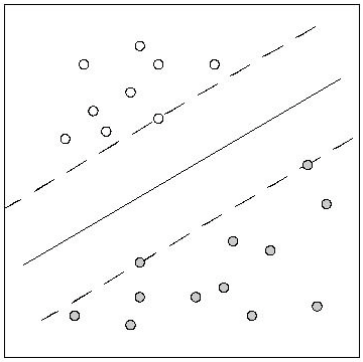

Figure 3 : Séparation parfaite en deux dimensions

Le plan ainsi défini peut être calculé à partir d'un sous-ensemble particulier de points appelé *ensemble de vecteurs de support* du plan. En réalité, le plan est défini par une somme pondérée de ces vecteurs de support (des coordonnées de ces exemples). On définit le plan en annulant l'équation 1 ou au moyen d'un seuil. Chaque vecteur x_i représente un point de support. Les y_i sont les valeurs à prédire et valent +1 ou -1. On voit que w est un vecteur de poids dont sont affectés au bout du compte les différents traits considérés; néanmoins, l'algorithme apprend en affectant chaque exemplaire de support (indicé par i) d'un poids a_i non nul.

$$\vec{w}\,\vec{x} = \sum_{i=1}^{n} a_i y_i (\vec{x}_i \vec{x})$$

Les poids sont choisis de manière à minimiser (à annuler en cas de parfaite séparation) le nombre d'erreurs de classification. La sortie de l'algorithme est constituée de ces coefficients pour chacun des vecteurs de support. Il permet de vérifier la qualité de l'apprentissage supervisé ainsi réalisé et de faire une prédiction pour un nouvel ensemble de points.

Le cas de *séparation imparfaite* peut évidemment, selon les domaines, apparaître plus ou moins fréquemment. Une solution plus flexible a donc été imaginée dans laquelle un point peut être éloigné arbitrairement du plan de séparation dans la direction qui aboutit à une classification correcte. La somme de ces éloignements est cependant soumise à optimisation. L'algorithme minimise alors une somme de deux termes l'un relatif à la marge et l'autre à la somme des ajustements consentis. Selon le poids C, plus ou moins grand, accordé à cette dernière, l'optimisation accorde donc une importance plus ou moins grande aux ajustements autorisés. Il s'en suit que pour des valeurs plus petites de C, le système produit plus d'erreurs de catégorisation durant l'apprentissage, c'est-à-dire s'ajuste moins aux particularités de l'échantillon et vice-versa. Un ajustement très précis aux particularités de l'échantillon d'apprentissage fait évidemment courir un plus grand risque de chute de performance lors de la généralisation à de nouvelles données. Le paramètre C est donc un élément important pour la généralisation ultérieure, tout gain de performance étant susceptible d'entraîner une perte de généralisation.

Un raffinement appelé *algorithme de transduction* peut être employé. Il consiste lors d'une épreuve de généralisation à utiliser l'information disponible dans les exemplaires du nouvel échantillon pour affiner l'apprentissage bien que les catégories de ces nouveaux exemplaires soient inconnues. Bien que leur classification ne soit pas disponible, il est clair que ces données fournissent une information sur l'occupation de l'espace. Cette information peut être utilisée pour éliminer des solutions qui ne s'adapteraient pas à cette occupation. Ceci allonge considérablement le temps de calcul, mais permet d'espérer un surcroît de précision.

Les versions *à noyau* de l'algorithme sont fondées exactement sur le même principe, mais ici une fonction, éventuellement non linéaire, (noyau) des produits vectoriels accumulés dans l'équation (1) est utilisée. Il peut être utile d'observer que dans l'équation 1 les poids sont déterminés sur la seule base des produits vectoriels entre exemplaires de support et exemplaires à classer. Une transformation de ces produits vectoriels suffit à projeter les exemplaires dans un espace différent. Ceci aboutit à identifier un hyperplan qui sépare non plus les exemplaires dans l'espace d'origine, mais, implicitement, leurs projections dans un espace différent. Il n'est pas considéré utile et il n'est donc pas courant d'utiliser ces techniques non-linéaires dans la classification de vecteurs représentant des textes. Nous nous sommes limités pour ce travail à la version linéaire.

4 Essais d'optimisation du classifieur pour le corpus français

L'ensemble des analyses rapportées ici a été réalisé au moyen du programme SVMLight V6.01 (Joachims, 2002). Afin de sélectionner les options de pré-traitement donnant lieu aux meilleures performances du classifieur, le corpus d'apprentissage français a été divisé en un corpus d'entraînement composé de 60% du corpus original et un corpus de test composé des 40% restants. Plusieurs répartitions aléatoires ont été effectuées. Les descripteurs initiaux correspondent à l'ensemble des termes (*token*) identifiés par TreeTagger dans les documents.

4.1 Paramètres classiques

Sur la base des procédures de pré-traitement courantes en traitement automatique du langage et de celles spécifiques au domaine de la catégorisation automatique, nous avons évalué les options suivantes :

Sélection et pré-traitement des descripteurs :

- Descripteurs : unigrammes, bigrammes et trigrammes.
- Seuil de fréquence minimale : 2, 3, 5, 10.
- Suppression de mots fonctionnels (articles, pronoms).
- Lemmatisation au moyen de Tree-Tagger.
- Sélection des descripteurs potentiellement les plus pertinents sur la base du test *t*, du test du *Chi-carré* et du test de *Wilcoxon-Mann-Whitney* (Paquot et Bestgen, 2009).

Attribution de valeurs numériques aux descripteurs :

- Pondération des fréquences des descripteurs dans les documents : fréquences brutes, binaire, log(freq+1) et LSA, la formule classique utilisée en analyse sémantique latente (Landauer et al., 1998 ; Piérard et Bestgen, 2006) qui combine des pondérations locale et globale au moyen de la formule suivante, dans laquelle f_{ij} fait référence à la fréquence brute du terme *j* dans le document *i* :

$$f_{ij}' = \frac{\log(f_{ij}+1)}{-\sum_j \frac{f_{ij}}{\sum_j f_{ij}} \log(\frac{f_{ij}}{\sum_j f_{ij}})}$$

- Normalisation des valeurs d'indice d'un document. Chaque valeur d'indice est divisée par la somme des valeurs pour ce document.

Paramètres de SVMLight

- Paramètre C.
- Algorithme de transduction.

On notera que certaines combinaisons de paramètres mènent à des difficultés, comme l'usage d'une pondération de type "fréquence" en l'absence de normalisation.

Étant donné le caractère partiel des analyses effectuées et le fait que les F-scores obtenus varient en fonction du rééchantillonnage effectué, nous ne rapportons ici que les informations générales qui en ont été extraites. Des données partielles, mais plus précises, basées sur le véritable corpus de test, sont données à la section suivante.

Ces analyses ont montré que l'utilisation d'un seuil de fréquence n'améliorait pas les analyses basées sur les unigrammes, mais bien celles sur les N-grammes, ce qui semble logique. Toutefois, ces analyses ont aussi montré que les bigrammes et les trigrammes pris indépendamment donnaient lieu à des performances moins bonnes que celles obtenues avec les unigrammes. Une combinaison des unigrammes, bigrammes et trigrammes n'est pas plus efficace que les unigrammes seuls.

La suppression des mots fonctionnels n'améliore pas les performances. La lemmatisation n'améliore pas non plus les résultats obtenus sur la base des unigrammes. Elle a un petit effet bénéfique pour les N-grammes, mais celui-ci est insuffisant pour égaler les performances des unigrammes non lemmatisés.

Les procédures de sélections des descripteurs testées n'ont pas permis d'améliorer les performances. Il faut cependant noter que nous n'avons probablement pas testé les indices les plus performants (Forman, 2003 ; mais voir (Gabrilovich, Markovitch, 2004 ; Joachims, 1998) pour une discussion de l'utilité de telles procédures).

L'analyse des différents schémas de pondération indique que la pondération dite « LSA » présentée ci-dessus est plus efficace que la pondération logarithmique et que l'absence de pondération. Ces essais ont aussi montré que la pondération LSA était plus efficace que TfIdf.

La normalisation des valeurs d'indice améliore nettement les performances, comme on pouvait s'y attendre en raison de la grande variabilité des longueurs des documents (Forman, 2003).

En ce qui concerne le paramètre C, nos essais indiquent qu'il a un impact important sur l'efficacité de l'apprentissage et qu'une valeur de 300 semble être optimale. L'apprentissage transductif donne lieu à une égalisation du nombre de documents mal classés dans les deux catégories, ce qui semble être légèrement bénéfique.

4.2 Prise en compte de la longueur des documents

Parmi les stratégies testées, la prise en compte de la longueur des documents mérite une attention toute particulière en raison de la nature même du corpus (articles de journaux) et de la tâche (catégorisation de référence effectuée, selon toute vraisemblance pour le corpus français, sur la base des rubriques dans lesquels les articles sont publiés).

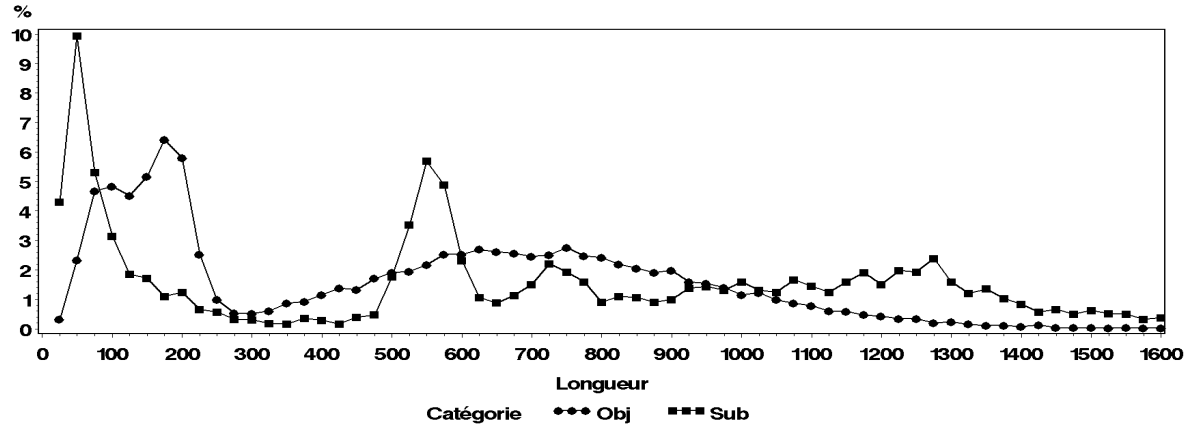

Figure 4 : Distribution des longueurs des documents pour le corpus français selon la catégorie : pourcentage sur le nombre de documents dans chaque catégorie en ordonnée, longueur en nombre de termes en abscisse

Les distributions des longueurs des documents selon la catégorie à laquelle ils ont été affectés sont très différentes comme l'indique la figure 4. Cette situation présente un défi intéressant pour l'emploi d'un algorithme comme SVM. En effet, celui-ci utilise une variable continue comme composant du vecteur représentant l'exemplaire et cette valeur entre dans les produits vectoriels de l'équation 1. Or, la comparaison des deux distributions montre que des plages disjointes de longueur correspondent à des taux de documents subjectifs très élevés. La longueur n'est donc pas liée de manière monotone à la probabilité que le document soit subjectif. Afin de permettre au classifieur de tirer idéalement parti de ces multiples zones, nous avons opté pour une discrétisation de la variable longueur (Liu, Setiono. 1997), représentée dès lors par un certain nombre de variables indicatrices. Discrétiser permet de transformer une variable "continue" en une série d'intervalles contigus qui sont employés pour recoder cette variable sous une forme discrète ou sous la forme d'une série de variables binaires. (Lustgarten et al., 2007) ont observé, dans un cadre d'analyse de données biomédicales, que la discrétisation ne permettait pas d'améliorer l'efficacité d'un classifieur SVM lorsqu'elle est appliquée à un très grand nombre de variables continues en raison du nombre de variables indicatrices ainsi créées. Pour réaliser celle-ci, plusieurs procédures sont envisageables comme :

- segmentation en intervalles de longueur constante, par exemple de 25 termes comme pour les graphiques,

- algorithme de type C4.5, tel qu'implémenté dans SAS Enterprise Mining,
- méthode de (Fayyad et Irani, 1993), basée sur une mesure d'entropie, telle qu'implémentée par exemple dans Weka.

On notera que la première approche est non supervisée alors que les deux autres le sont (Lustgarten et al, 2007). Nos analyses montrent que ces trois approches sont quasiment équivalentes les unes aux autres, les différences maximales observées lors des tests étant inférieures à 0.0025 point de F-score. Elles montrent aussi que la discrétisation de la longueur permet d'améliorer les performances du classifieur (voir la section suivante pour des détails). Il est cependant indispensable de noter que la longueur des documents est une variable très particulière puisqu'elle est liée de manière non monotone à la probabilité de classification subjective, ce qui d'ailleurs a mené à cette discrétisation. Une telle relation a, a priori, peu de chances de se produire avec beaucoup d'autres descripteurs. C'est ce que confirment une série d'analyses qui montrent que, parmi les autres descripteurs, seuls quelques signes de ponctuation montrent ce genre de profil et qu'ils apportent des gains négligeables.

La figure 5 montre qu'une discrétisation de la variable longueur semble nettement moins pertinente en anglais qu'en français (figure 4) pour distinguer les deux catégories. Force est donc de constater que l'intérêt potentiel de la procédure de discrétisation employée ici est limité à des problèmes de catégorisations très spécifiques.

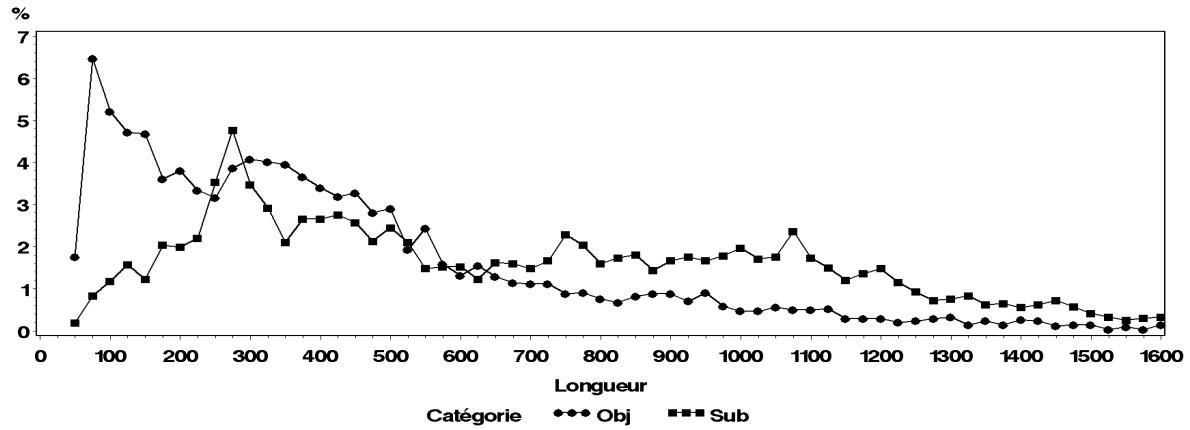

Figure 5 : Distribution des longueurs des documents pour le corpus anglais selon la catégorie : pourcentage sur le nombre de documents dans chaque catégorie en ordonnée, longueur en nombre de termes en abscisse

5 Analyse des effets spécifiques des paramètres

Les expérimentations brièvement rapportées ci-dessus ont conduit à sélectionner, comme optimaux pour le corpus français, les pré-traitements suivants : unigrammes, seuil de fréquence à 2, sans lemmatisation, pondération LSA, normalisation et longueur sous forme discrète. Cette solution, appliquée au corpus français et au corpus anglais de la tâche 1, a été soumise lors de la phase de test. Elle sert aussi de point de référence pour une analyse de modifications *indépendantes* de chacun de ces paramètres sur les performances du classifieur. Le paramètre C de SVMLight a été fixé à 300 et il a été décidé de ne pas utiliser l'algorithme de transduction qui accroît le temps calcul d'un facteur supérieur à 100.

La première ligne du tableau 5 donne les performances de la procédure que nous avons considérée comme optimale. Les lignes suivantes présentent les performances de procédures basées sur les mêmes paramètres que la procédure optimale à l'exception d'un seul. Celui-ci est le seul paramètre mentionné explicitement pour cette ligne. Une cellule vide, quelle que soit sa position dans le tableau, signale donc que la valeur considérée comme optimale de ce paramètre a été employée. Les lignes sont ordonnées en fonction du F-score pour le corpus français.

Lemme	Pondération	Normalisation	Seuil de fréquence	Longueur	F-score Fr	F-score En
Non	LSA	Oui	2	Discrète	0.9271	0.8513
			3		0.9269	0.8502
			5		0.9267	0.8488
			10		0.9260	0.8464
	Binaire				0.9254	0.8445
	Log				0.9209	0.8507
				Continue	0.9199	0.8491
				Aucune	0.9171	0.8485
Oui					0.9171	0.8456
	Fréquence				0.9131	0.8469
		Non			0.8969	0.8408

Tableau 5 : Résultats pour les corpus de test français et anglais en fonction des paramètres.

On note, en tout premier lieu, qu'aucun des paramètres évalués n'a un impact important sur les performances dans les deux langues. Pour l'anglais, la différence la plus importante est d'à peine 0.01 point de F-score. Pour le français, la normalisation semble nécessaire, les autres paramètres n'ayant qu'un impact inférieur à 0.015 point de F-score. C'est cette petitesse des écarts qui nous a conduits à présenter les F-score avec 4 décimales.

Dans ce paysage sans relief, on notera, néanmoins, que la prise en compte de la longueur sous une forme discrétisée apporte un gain de 0.007 point de F-score en français alors que le gain n'est que de 0.002 point de F-score en anglais. Même pour le français, le bénéfice est petit, mais il faut garder à l'esprit qu'il est obtenu par l'entremise de la discrétisation d'un seul descripteur alors que plus de 100 000 autres descripteurs étaient à la disposition du classifieur. Cette absence de relief confirme la possibilité d'utiliser les résultats comme une sorte de niveau de base.

La figure 6 permet de se faire une idée de la manière dont le classifieur a tiré profit de la discrétisation des longueurs des documents. L'ordonnée affichée à gauche sert de référence pour la courbe qui présente les rapports entre les fréquences des longueurs des documents pour les deux catégories obtenus au moyen de la formule suivante :

$$Rapport = 100 \times \frac{Freq_{Sub}}{Freq_{Sub} + Freq_{Obj}}$$

L'ordonnée à droite sert de référence pour la courbe qui présente les poids que le classifieur donne aux descripteurs de la longueur discrétisée, une valeur négative correspondant à un descripteur propre à la catégorie objective et une valeur positive à un descripteur typique de la catégorie subjective. Ces valeurs ont été obtenues au moyen de la procédure proposée par Joachims (http://www.cs.cornell.edu/People/tj/svm_light/svm_light_faq.htm). À titre de comparaison, le descripteur correspondant au mot *erreur* reçoit un poids de 25.87 qui peut être mis en relation avec la fréquence des erratas "subjectifs" dans la classe de longueur *25* ("*par erreur*", voir section 2). Le même mot, mais au pluriel, ne reçoit qu'un poids de 6.51. Comme les poids attribués à un descripteur dépendent des poids de l'ensemble des autres descripteurs, il est normal que la correspondance entre les deux courbes ne soit pas parfaite. Celle-ci est néanmoins suffisante pour montrer l'usage fait par le classifieur de l'information apportée par la discrétisation de la longueur des documents.

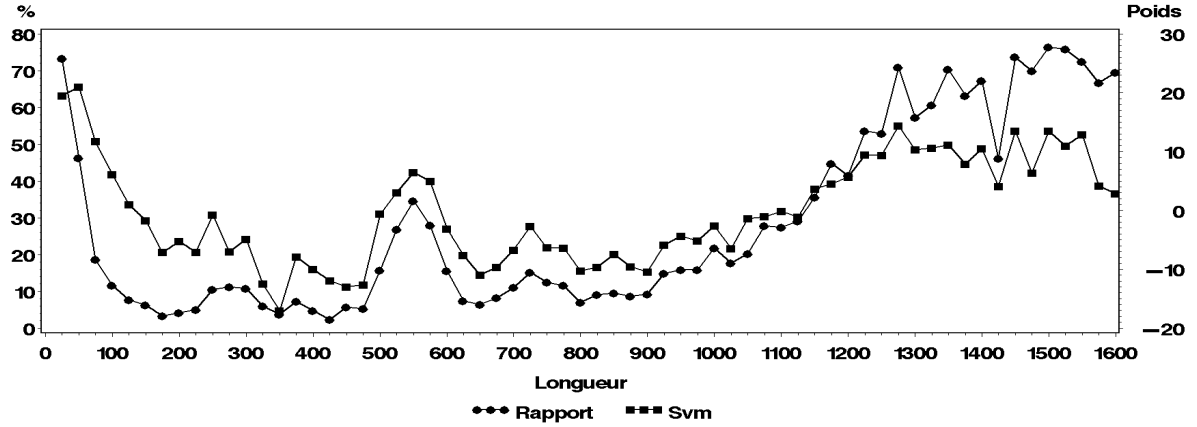

Figure 6 : Comparaison entre le rapport des fréquences des longueurs des documents pour les deux catégories et les poids des descripteurs correspondant à la discrétisation des longueurs (corpus français).

6 Résultats finaux

Les résultats finaux de cette campagne d'évaluation, tels qu'ils nous ont été transmis par les organisateurs, sont un F-score de 0.925 pour le corpus français et de 0.851 pour le corpus anglais. Étant donné que ces valeurs ont été obtenues au moyen d'un algorithme classique en catégorisation de textes et que la section 5 a montré que nos tentatives d'optimisation des paramètres n'avaient apporté que des gains négligeables, ces valeurs nous semblent pouvoir être considérées comme des niveaux de base permettant de se faire une idée de la difficulté de la tâche 1 de DEFT'09 pour les corpus français et anglais et pouvant servir, si nécessaire, de points de comparaison pour les autres participants. Le seul autre commentaire possible est que, contrairement à nos attentes, la procédure de transduction a donné lieu à une infime baisse de la performance pour le corpus français (0.002).

Remerciements

Yves Bestgen est chercheur qualifié du F.R.S-FNRS.

Références

Burges C.J.C. (1998), A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, Vol. 2, 1-47.

Fayyad U.M., Irani K.B. (1993), Multi-interval discretization of continuous-valued attributes for classification learning, Proceedings of the *13th International Joint Conference on Artificial Intelligence,* 1022-1027.

Forman G. (2003), An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research*, Vol 3, 1289-1305

Gabrilovich E., Markovitch S. (2004). Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5, Proceedings of the *21st International conference on machine learning*, 8 p.

Hurault-Plantet M. (2009), Détection des opinions dans les textes, *Séminaire INALCO*, 09/04/2009, (http://www.limsi.fr/Individu/mhp/seminaire/inalco-detection-opinion.pdf).

Joachims T. (1998). Text categorization with support vector machines: Learning with many relevant features, Proceedings of *ECML'98*, 137-142.

Joachims T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*, Dordrecht, Kluwer.

Landauer T.K., Foltz P.W., Laham D. (1998), An introduction to latent semantic analysis, *Discourse processes*, Vol. 25, 259-284.

Liu H, Setiono R. (1997). Feature selection via discretization. *Knowledge and data engineering*, Vol. 9, 642-645.

Lustgarten J.L., Gopalakrishnan V., Grover H., Visweswaran S., (2008), Improving classification performance with discretization on biomedical datasets. Proceedings of *AMIA 2008 Symposium*, 445-449.

Paquot M., Bestgen Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction, In Jucker A.H., Schreier D., Hundt M, (Eds), *Corpora: Pragmatics and discourse* (pp.243-265), Amsterdam: Rodopi.

Piérard S. Bestgen Y. (2006), Validation d'une méthodologie pour l'étude de deux types marqueurs de la segmentation dans un grand corpus de texte, *Traitement automatique des langues,* Vol. 47, 89-110.

Pomikalek J., Rehurek R. (2007), The Influence of preprocessing parameters on text categorization, *Proceedings of world academy of science, engineering and technology*, Vol. 21, 430-433.

Schmidt H. (1994). Probabilistic part-of-speech tagging using decision trees, Proceedings of the *International conference on new methods in language processing*, 9 p., (revised version).

Impacts de la variation du nombre de traits discriminants sur la catégorisation des documents

Dominic Forest avec la collaboration de Astrid van Hoeydonck,
Danny Létourneau et Martin Bélanger

Université de Montréal – École de bibliothéconomie et des sciences de l'information
C.P. 6128, Succursale Centre-Ville, Montréal, Québec, Canada, H3C 3J7
dominic.forest@umontreal.ca

Résumé – Abstract

Cet article fait état des résultats générés par deux démarches de fouille de textes. La première démarche a été réalisée afin d'assister l'identification du caractère objectif ou subjectif d'un corpus d'articles de journaux. La seconde démarche a été appliquée afin d'assister l'identification du parti politique d'un parlementaire. Dans les deux cas, les démarches ont été réalisées en utilisant l'algorithme des k plus proches voisins (*k-nearest neighbor*). Cet article porte sur la dimension informationnelle de la démarche de fouille de textes. Il présente l'impact de la variation du nombre de traits discriminants sur les résultats de la catégorisation automatique des documents.

This paper presents the results of two text mining processes. The first process aimed at assisting the identification of the objective or subjective characteristic of a corpus of newspaper articles. The second process aimed at assisting the identification of the political party of a corpus of political interventions. Both processes were accomplished using the k-nearest neighbor algorithm. This paper focuses on the informational dimensions of text mining processes. It reports on the impact of the variation of features on automatic text categorization results.

Mots-Clés – Keywords

Catégorisation automatique de documents, traits discriminants, variation, application.

Automatic text classification, feature selection, variation, application.

1 Introduction

Depuis une dizaine d'années, le domaine de la fouille de textes s'avère un territoire de recherche des plus actifs. Ainsi, de nombreuses initiatives de recherche ont tenté de développer des applications informatiques permettant d'extraire des patrons d'informations caractéristiques de documents, d'en identifier le contenu thématique ou même d'en extraire certaines informations précises. Les applications développées intègrent des concepts et des techniques provenant de plusieurs disciplines académiques bien établies. Ainsi, les récents développements dans le domaine de la fouille de textes reposent très souvent sur des concepts et des traitements éprouvés issus des domaines de l'intelligence

artificielle et de l'apprentissage machines, de la linguistique informatique, des sciences de l'information, etc.

Plusieurs facteurs ont motivé les recherches dans ce domaine. Parmi ces facteurs, on retrouve au premier plan le nombre croissant de documents disponibles en format numérique. Nous disposons désormais de très volumineux corpus de documents textuels en format numérique, dont l'exploitation ne peut être réalisée sans avoir recours à des techniques évoluées de traitement de l'information. Ainsi, les conséquences des initiatives de numérisation ont des répercussions directes sur le développement d'applications visant à assister la recherche, l'analyse, la structuration, la gestion et la diffusion de l'information présente dans les documents textuels.

Les développements technologiques dans le domaine de la fouille de textes ont pris la forme de logiciels propriétaires permettant de combiner différentes opérations de fouille afin d'effectuer des traitements plus ou moins complexes sur des corpus de documents textuels non structurés. Parmi les principaux logiciels de fouille de textes commerciaux disponibles, on retrouve entre autres *Text Analyst* (Megaputer), *Text Miner* (SAS), *PASW Modeler* (SPSS). Plus récemment, plusieurs efforts ont été consacrés au développement de plates-formes de fouille de données modulaires et flexibles offrant aux utilisateurs la possibilité de combiner différents modules afin de générer rapidement des chaînes de traitement plus adaptées à leurs besoins (*RapidMiner*, *Weka*, etc.). Ces plates-formes ont d'abord été conçues pour traiter des données structurées. Cependant, elles sont souvent accompagnées de modules supplémentaires (prétraitement, filtrage linguistique, conversion numérique, etc.) leur permettant de traiter efficacement des documents textuels non structurés.

Les développements technologiques dans le domaine de la fouille de textes ont été notables au cours des dernières années, tant en ce qui a trait aux algorithmes de traitement, qu'aux performances et aux interfaces utilisateurs. Cependant, malgré ces avancées technologiques, nous constatons que certaines difficultés demeurent en ce qui a trait à l'utilisation des techniques de fouille de textes. Au-delà des particularités et des difficultés propres à chaque corpus, deux difficultés font toujours obstacle à l'utilisation des outils de fouille de textes. La première de ces difficultés est d'ordre méthodologique. Dans de nombreux cas, les utilisateurs ont une idée relativement précise des informations qu'ils souhaitent extraire à partir des données textuelles dont ils disposent. Cependant, la démarche grâce à laquelle ils pourraient extraire optimalement les informations recherchées est trop souvent inconnue. Il existe certes un ensemble d'opérations potentiellement utiles afin d'atteindre l'objectif souhaité, mais la méthodologie à déployer pour atteindre l'objectif n'a pas été identifiée. Dans plusieurs projets de fouille de textes, les questions suivantes demeurent souvent sans réponse satisfaisante : quels sont les algorithmes à appliquer afin d'assister le plus efficacement la réalisation d'une tâche spécifique ? Par exemple, quelle serait la démarche méthodologique optimale à mettre en œuvre si l'on souhaite assister l'extraction d'opinions à partir de large corpus bilingues ? Est-il utile d'appliquer un algorithme de lemmatisation sur les données ? Si oui, est-il préférable de le faire après ou avant la suppression des mots fonctionnels ? À quelle étape doit être appliqué l'algorithme d'extraction des entités nommées ? Est-il nécessaire de procéder préalablement à une opération de marquage morphosyntaxique des données linguistiques initiales ? Voilà autant de questions d'ordre technique et méthodologique auxquelles il est actuellement très difficile de fournir des réponses éclairées, *a fortiori* lorsqu'elles sont posées dans des contextes applicatifs bien spécifiques. Ainsi, nous constatons que la dimension méthodologique de la fouille de textes n'a pas été une préoccupation importante dans ce domaine. Nous disposons de techniques de fouille efficaces, mais nous n'en connaissons malheureusement trop peu sur les modalités d'application des différents algorithmes afin d'assister efficacement les tâches auxquelles nous sommes confrontées.

La seconde difficulté est étroitement liée à la première. Elle concerne les paramètres à spécifier au niveau de chaque étape du processus général de fouille de textes. À l'instar de la dimension méthodologique de la fouille de texte, les paramètres à spécifier pour chaque algorithme sont inévitablement dépendants de l'objectif à atteindre, des particularités inhérentes aux documents à traiter et du contexte dans lequel le traitement est réalisé. Cependant, malgré la particularité de chaque

contexte applicatif, peu d'informations sont connues concernant les paramètres optimaux à spécifier pour chaque étape du processus de fouille. Par exemple, peu d'informations pratiques sont disponibles concernant les paramètres de prétraitement et de filtrage à mettre en œuvre dans les premières étapes du processus. Quels sont les paramètres qu'il serait souhaitable de mettre en œuvre pour effectuer un filtrage statistique des données linguistiques extraites du corpus que l'on souhaite analyser ? En dessous de quel seuil de fréquence les mots retenus comme traits discriminants doivent-ils être supprimés afin d'accroitre les performances d'un algorithme de classification automatique ? Voilà autant de questions auxquelles il est actuellement difficile de répondre. Dans le domaine de la fouille de données (que les données soient de nature textuelle ou non), la majorité des opérations que l'on exécute impliquent que l'on spécifie un ou plusieurs paramètres. Il en va ainsi tant au début du processus (au moment du filtrage des données) qu'à l'étape finale de validation. Souvent, trop peu d'informations rigoureuses sont connues concernant les paramètres optimaux à mettre en œuvre afin de réaliser des tâches en contexte spécifique.

2 Objectifs

Dans cet article, nous présentons la démarche que nous avons employée, ainsi que les résultats que nous avons obtenus lors de notre participation à l'édition 2009 du DEfi Fouille de Textes (DEFT'09). En outre, nous présentons nos travaux sous l'angle du second problème que nous avons identifié concernant l'utilisation des processus de fouille de textes en contexte applicatif réel. Plus spécifiquement, l'objectif de notre démarche dans ce projet consiste à identifier, dans des contextes expérimentaux précis, l'impact de la fluctuation du nombre de traits discriminants retenus pour représenter des documents qui sont soumis à une opération de catégorisation automatique.

Dans le cadre de DEFT'09, nous avons exécuté deux tâches qui nous ont permis d'explorer l'impact de la variation du nombre de traits discriminants sur les résultats de la catégorisation automatique. La première de ces tâches consiste à prédire le caractère objectif ou subjectif d'un corpus d'articles de journaux. L'identification du caractère objectif ou subjectif d'un article de journal est une opération importante dans le domaine de la fouille de textes. En effet, elle est à la base de plusieurs applications complexes de traitement de l'information, parmi lesquelles figure au premier plan l'identification automatique d'opinions (*sentiment analysis*) (Lui, 2007).

La seconde tâche consiste à identifier le parti politique d'un parlementaire. Cette tâche est comparable à celle qui consiste à prédire l'auteur d'un document (*authorship attribution*). Plutôt que d'identifier l'auteur d'un document, l'objectif de cette seconde tâche consiste à identifier automatiquement le parti politique auquel se rattache d'auteur d'une intervention politique. L'identification d'auteur a fait l'objet de nombreux travaux dans le domaine de la fouille de textes. Comme l'a clairement souligné (Juola, 2008), l'identification d'auteur est une tâche complexe qui fait intervenir plusieurs dimensions dans la description des données textuelles. Malgré les récents développements dans ce domaine, cette tâche pose encore de nombreux problèmes, tant théoriques que pratiques.

3 Méthodologie

La démarche méthodologique que nous avons employée pour accomplir les deux tâches est inspirée de celle que l'on retrouve au cœur de nombreux projets de fouille de données. Cette démarche est principalement de nature numérique. Pour des raisons théoriques et pratiques, nous avons volontairement réduit au minimum le nombre d'opérations faisant intervenir des dimensions linguistiques. La démarche repose sur le modèle vectoriel pour le traitement des documents (Salton, 1988 ; Memmi, 2000). Elle est composée de cinq principales étapes.

La première étape de tout processus de fouille de données réside dans le développement ou la constitution d'un corpus de documents. Il est essentiel que le corpus de documents soit constitué en tenant compte des objectifs à atteindre par le processus de fouille. À l'étape de constitution du corpus, quatre grandes familles de caractéristiques doivent être évaluées et prises en considération. La constitution d'un corpus à des fins de fouille de textes implique certains choix en ce qui concerne les caractéristiques 1) générales (provenance, taille, date de création, etc.), 2) technologiques (support, format, etc.), 3) informationnelles (thématiques et sujets abordés) et 4) linguistiques (langue, genre, registres, etc.) des documents. Dans le cadre de DEFT'09, la constitution du corpus a été entièrement prise en charge par le comité organisateur, avec le concours de l'Agence pour l'Évaluation et la Distribution des Ressources Linguistique (ELDA). Nous n'avons apporté aucune modification aux documents qui nous ont été fournis par le comité organisateur.

La seconde étape de la démarche a consisté à extraire, à filtrer et à normaliser le lexique du corpus. L'opération de filtrage du lexique est composée traditionnellement de plusieurs sous-opérations. La première d'entre elles consiste à supprimer certains mots non pertinents pour l'analyse. Le filtrage du lexique peut être effectué à l'aide de plusieurs techniques, certaines étant de nature linguistique, d'autres de nature statistique. Une première opération a pour but de supprimer l'ensemble des mots fonctionnels présents dans le texte. Ce processus est réalisé en retirant les termes figurant dans une liste prédéfinie de mots fonctionnels. Il est aussi souhaitable d'appliquer certains filtres statistiques au lexique du corpus afin d'en éliminer les unités qui, tout en ne figurant pas dans la liste des mots fonctionnels, ne sont pas pertinentes pour l'analyse. La pertinence des termes est très étroitement associée à leur potentiel discriminant. Ainsi, il importe de supprimer les mots dont la fréquence est supérieure ou inférieure à certains seuils (souvent déterminés empiriquement). Dans un dernier temps, il est d'usage d'appliquer un processus de généralisation sensible aux variantes sémantiques et syntaxiques présentes dans le corpus. Il importe alors d'appliquer au lexique du corpus une opération de lemmatisation. L'opération de lemmatisation est réalisée généralement d'abord en effectuant un marquage morphosyntaxique des différents lexèmes à analyser, ensuite en comparant ceux-ci à un dictionnaire. Ce processus permet de dégager une liste de lemmes propres à une langue donnée. S'il est impossible de lemmatiser les données à traiter (par manque de ressources linguistiques, par exemple), il est souhaitable de recourir à un processus d'amputation des terminaisons (*stemming*), lequel génère une liste de *stems* (racines).

La troisième étape de la démarche consiste à convertir le corpus initial dans un format pouvant être traité par les algorithmes de fouille. Cette opération est réalisée en structurant les documents du corpus en une matrice de vecteurs dans laquelle chaque document (ou segment de document) est représenté par l'absence ou la présence, binaire ou pondérée, de chaque unité lexicale retenue à l'étape précédente.

C'est à la quatrième étape de la démarche que sont réalisées les opérations permettant plus spécifiquement d'extraire et de structurer les informations présentes dans le corpus. Dans une perspective de fouille de textes, la majorité des opérations d'extraction et de structuration des informations sont réalisées en utilisant des algorithmes développés dans les domaines de l'intelligence artificielle et de l'apprentissage machine.

Les tâches de catégorisation automatique – qui consistent à attribuer une ou plusieurs catégories à chaque document d'un corpus – sont traditionnellement accomplies en utilisant des algorithmes d'apprentissage supervisés. La principale particularité des techniques supervisées réside dans leur capacité à projeter certaines caractéristiques des documents préalablement connues et apprises par le système sur un ensemble de documents pour lesquels les mêmes caractéristiques ne sont pas encore connues. En vertu de cette particularité, les techniques supervisées impliquent donc d'abord une phase d'apprentissage (réalisée sur un corpus d'apprentissage) et, ensuite, une phase de test (ou d'application) lors que laquelle l'apprentissage effectué par le système est projeté sur de nouveaux documents (en contexte de test ou d'application concrète). Dans le cadre de DEFT'09, nous avons d'abord exploré deux algorithmes de classification : l'algorithme des k plus proches voisins (Manning

et Schütze, 1999) et un classifieur bayésien naïf (Manning et Schütze, 1999). Lors de la phase de test, seul l'algorithme des k plus proches voisins a été retenu.

La cinquième étape de la démarche réside dans l'interprétation, l'évaluation et l'intégration des résultats générés par les algorithmes de fouille de textes. Les opérations d'interprétation et d'évaluation sont des plus complexes, car elles sont dépendantes de plusieurs facteurs extrinsèques au processus de traitement des documents textuels. Les algorithmes supervisés peuvent être évalués selon les mesures classiques de rappel et de précision.

Par ailleurs, l'interprétation des résultats des algorithmes de fouille ne peut être dictée par aucun cadre théorique qui ferait abstraction du contexte dans lequel l'opération de fouille est réalisée. Finalement, les résultats doivent normalement faire l'objet d'un processus d'intégration à l'intérieur d'une application finale plus complexe dans laquelle le processus de fouille ne constitue d'une étape bien précise. Les applications finales intégrant des processus de fouille de textes sont de plus en plus nombreuses et variées. Parmi celles-ci, on trouve entre autres les applications de veille scientifique, de gestion électronique des documents et de recherche d'informations. La figure 1, inspirée de Fayyad *et al.* (1996), présente les principales étapes de la méthodologie générique de fouille de textes.

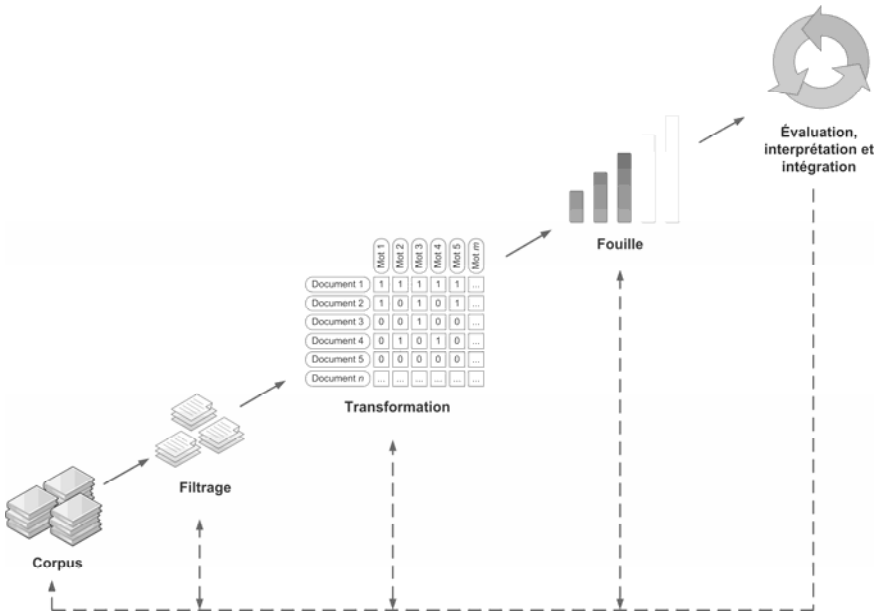

Figure 1. La démarche méthodologique (Forest, 2009).

4 Corpus

4.1 Corpus 1

Le corpus associé à la tâche de prédiction du caractère objectif ou subjectif a été divisé en respectant un ratio 2/3 – 1/3 classique dans les domaines de la recherche d'informations et du traitement automatique des langues (TAL). Les deux tiers du corpus ont été exploités à des fins d'apprentissage, alors que le tiers restant a été utilisé à des fins de test.

Le sous-corpus d'apprentissage est composé de 25 176 documents français. Il est composé de 12 511 590 mots (occurrences) et de 75 302 formes. Le lexique de ce sous-corpus a été lemmatisé. Nous en avons aussi supprimé les mots fonctionnels. Finalement, les hapax, ainsi que les formes présentes dans

plus de 25% des documents ont été supprimés. À la suite de ces prétraitements, le lexique restant était composé de 51 864 mots. Par la suite, nous avons utilisé une mesure de Chi2 pour identifier des sous-ensembles de mots fortement discriminants qui ont servi à décrire numériquement les documents.

Le sous-corpus de test est composé de 16 788 documents français. Il est composé de 8 499 931 mots (occurrences) et de 110 297 formes. Afin de catégoriser automatiquement les documents du corpus de test, les documents ont préalablement été convertis numériquement en utilisant les mots retenus à l'étape d'apprentissage.

4.2 Corpus 2

Le corpus associé à la tâche de prédiction du parti politique a été divisé en respectant aussi un ratio 2/3 – 1/3. Les deux tiers du corpus ont été exploités à des fins d'apprentissage, alors que le tiers restant a été utilisé à des fins de test.

Le sous-corpus d'apprentissage est composé de 19 370 documents français. Il est composé de 7 208 721 mots (occurrences) et de 53 531 formes. Le lexique de ce sous-corpus a été lemmatisé. Nous en avons aussi supprimé les mots fonctionnels. Finalement, les hapax, ainsi que les formes présentes dans plus de 50% des documents ont été supprimés. À la suite de ces prétraitements, le lexique restant était composé de 21 698 mots. Par la suite, nous avons encore une fois utilisé une mesure de Chi2 pour identifier des sous-ensembles de mots fortement discriminants qui ont servi à décrire numériquement les documents.

Le sous-corpus de test est composé de 12 914 documents français. Il est composé de 4 799 665 mots (occurrences) et de 46 242 formes. Afin de catégoriser automatiquement les documents du corpus de test, les documents ont préalablement été convertis numériquement en utilisant les mots retenus à l'étape d'apprentissage.

5 Résultats

Cette section présente les résultats que nous avons obtenus pour les deux tâches que nous avons réalisées. Nous présentons d'abord les résultats obtenus lors de la phase d'apprentissage, puis lors de la phase de test. Tel que nous l'avons mentionné précédemment, nous avons fait varier plusieurs paramètres lors de la phase d'apprentissage. Les résultats que nous présentons font état de ces variations. Les résultats obtenus lors de la phase d'apprentissage sont quantifiés en utilisant les mesures de rappel et de précision. Afin d'évaluer les performances du processus de catégorisation au moment de la phase d'apprentissage, une méthode d'échantillonnage croisée a été appliquée (*10-fold cross-validation*). L'ensemble des expérimentations menées dans ce projet a été réalisé avec le logiciel *WordStat* (Provalis Research).

En ce qui concerne les résultats de la phase de test, nous présentons les résultats que nous avons obtenus selon trois configurations expérimentales. Ces résultats ont été calculés par le comité organisateur de DEFT'09.

5.1 Tâche 1. Prédiction du caractère objectif ou subjectif

5.1.1 Phase d'apprentissage

Expérimentations d'apprentissage utilisant les paramètres suivants (figures 2 et 3) :

- Nombre de traits discriminants : entre 10 000 et 50 000 mots discriminants (avec un incrément de 10 000)
- Méthodes de représentation des traits : fréquence et fréquence pondérée par IDF
- Algorithmes employés : k plus proches voisins (k=1 et k=5) et classifieur bayésien naïf

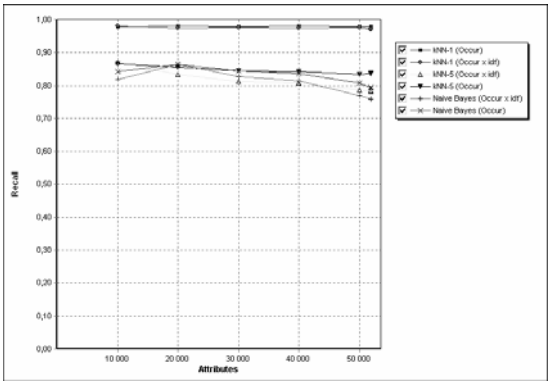

Figure 2. Tâche 1 – performances (rappel) du premier ensemble d'expérimentations d'apprentissage.

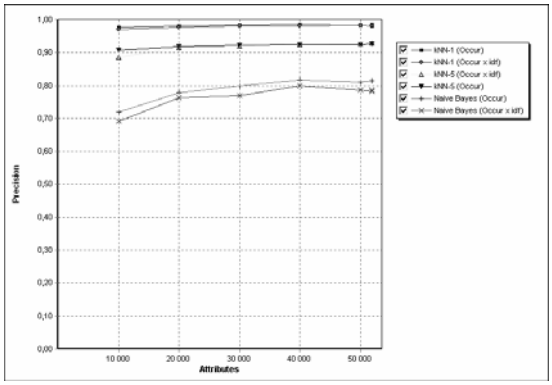

Figure 3. Tâche 1 – performances (précision) du premier ensemble d'expérimentations d'apprentissage.

Expérimentations d'apprentissage utilisant les paramètres suivants (figures 4 et 5) :

- Nombre de traits discriminants : entre 1 000 et 10 000 mots discriminants (avec un incrément de 1 000)
- Méthodes de représentation des traits : fréquence et fréquence pondérée par IDF
- Algorithmes employés : k plus proches voisins (k=1 et k=5) et classifieur bayésien naïf

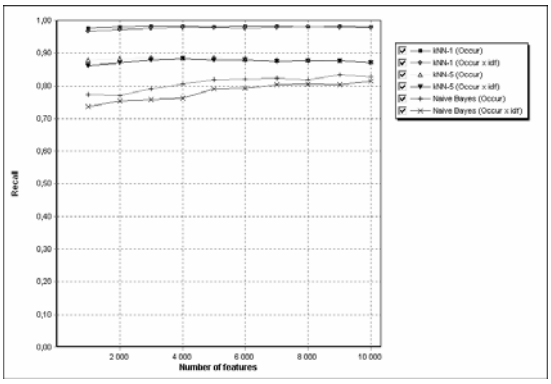

Figure 4. Tâche 1 – performances (rappel) du deuxième ensemble d'expérimentations d'apprentissage.

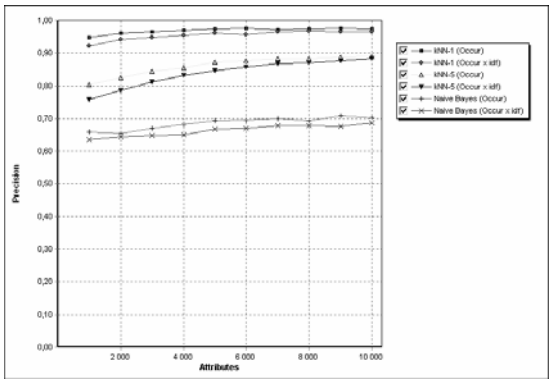

Figure 5. Tâche 1 – performances (précision) du deuxième ensemble d'expérimentations d'apprentissage.

Expérimentations d'apprentissage utilisant les paramètres suivants (figures 6 et 7) :

- Nombre de traits discriminants : entre 100 et 3 000 mots discriminants (avec un incrément de 100)
- Méthodes de représentation des traits : fréquence et fréquence pondérée par IDF
- Algorithmes employés : k plus proches voisins (k=1 et k=5) et classifieur bayésien naïf

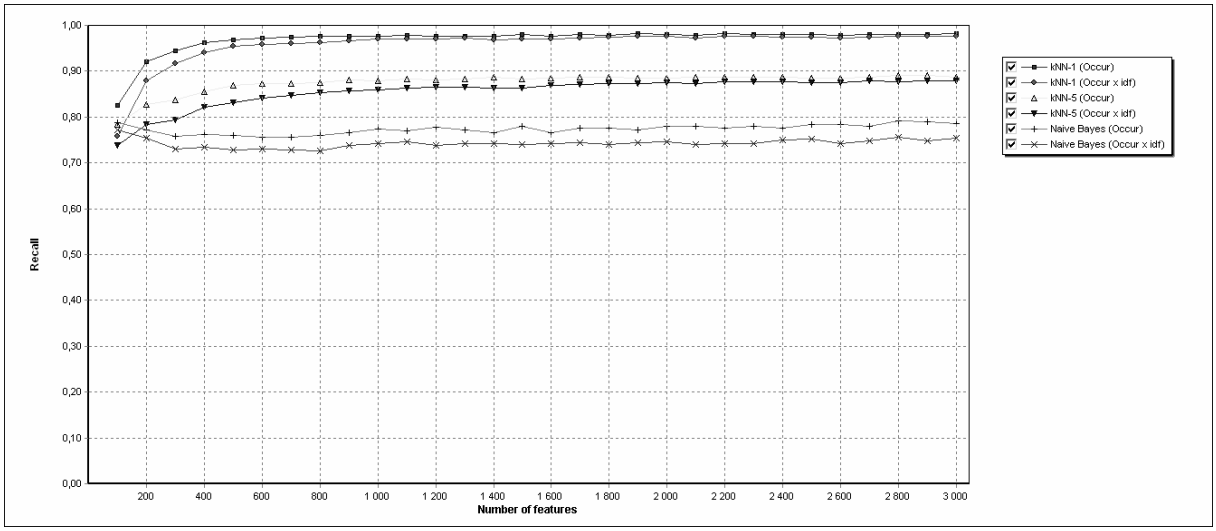

Figure 6. Tâche 1 – performances (rappel) du troisième ensemble d'expérimentations d'apprentissage.

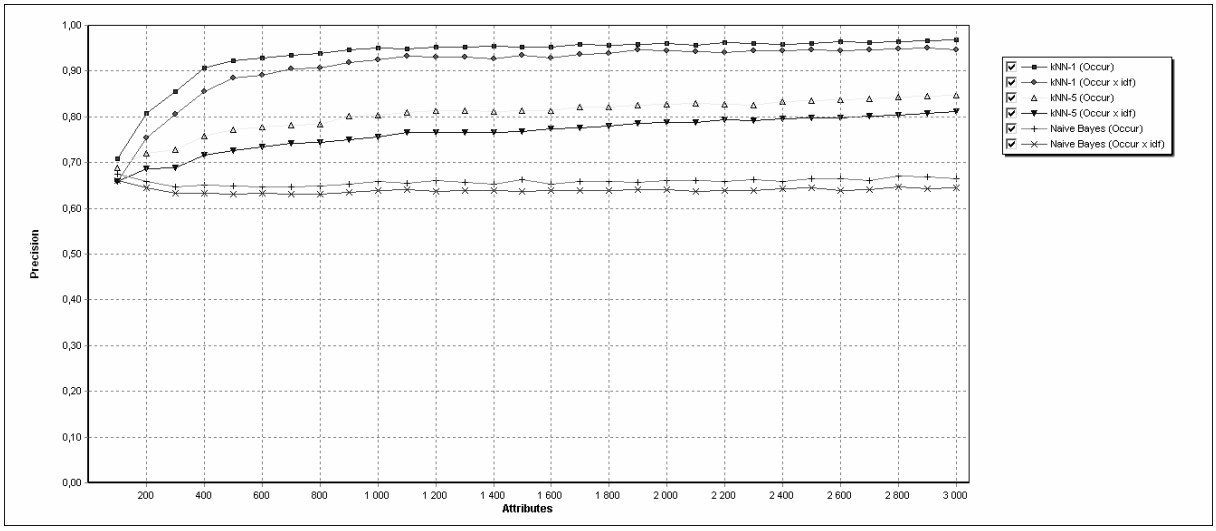

Figure 7. Tâche 1 – performances (précision) du troisième ensemble d'expérimentations d'apprentissage.

Durant la phase d'apprentissage, les meilleures performances ont été obtenues en utilisant 40 000 mots discriminants représentés uniquement par leur fréquence. Ces mots ont été employés pour décrire les documents qui ont été soumis à l'algorithme des k plus proches voisins avec le paramètre k=2. Ainsi, selon en utilisant cet algorithme avec ce paramètre, nous avons obtenu un taux de rappel de 98.09% et un taux de précision de 98.46%. En contrepartie, les pires performances ont été obtenues en utilisant 800 mots discriminants représentés par leur fréquence pondérée qui ont été soumis au classifieur bayésien naïf. Ainsi, selon en utilisant cet algorithme avec ce paramètre, nous avons obtenu un taux de rappel de 63% et un taux de précision de 72.59%.

5.1.2 Phase de test

Lors de la phase de test, nous avons mené trois exécutions en ne faisant varier que le nombre de traits discriminants. En effet, nous avons constaté lors de la phase d'apprentissage que les meilleurs résultats ont toujours été obtenus sans pondérer la fréquence des mots et en utilisant l'algorithme des k plus proches voisin avec le paramètre k=1. Les résultats que nous avons obtenus lors de ces exécutions sont les suivants :

Exécution de test 1 utilisant les paramètres suivants :

- Nombre de traits discriminants : 6 000
- Méthodes de représentation des traits : fréquence
- Algorithmes employés : k plus proches voisins (k=1)

Performances globales : rappel = 77.80%, précision = 73.80%, f-mesure = 75.70%

Performances spécifiques (catégorie *Objectif*) : rappel = 88.40%, précision = 92.80%

Performances spécifiques (catégorie *Subjectif*) : rappel = 67.30%, précision = 54.70%

Exécution de test 2 utilisant les paramètres suivants :

- Nombre de traits discriminants : 20 000
- Méthodes de représentation des traits : fréquence
- Algorithmes employés : k plus proches voisins (k=1)

Performances globales : rappel = 77.90%, précision = 77.60%, f-mesure = 77.80%

Performances spécifiques (catégorie *Objectif*) : rappel = 92.20%, précisions = 92.40%

Performances spécifiques (catégorie *Subjectif*) : rappel = 63.60%, précision = 62.90%

Exécution de test 3 utilisant les paramètres suivants :

- Nombre de traits discriminants : 40 000
- Méthodes de représentation des traits : fréquence
- Algorithmes employés : k plus proches voisins (k=1)

Performances globales : rappel = 77.30%, précision = 79.00%, f-mesure = 78.10%

Performances spécifiques (catégorie *Objectif*) : rappel = 93.40%, précision = 92.00%

Performances spécifiques (catégorie *Subjectif*) : rappel = 61.20%, précision = 65.90%

5.2 Tâche 2. Prédiction du parti politique

5.2.1 Phase d'apprentissage

Expérimentations d'apprentissage utilisant les paramètres suivants (figures 8 et 9) :

- Nombre de traits discriminants : entre 5 000 et 20 000 mots discriminants (avec un incrément de 5 000)
- Méthodes de représentation des traits : fréquence et fréquence pondérée par IDF
- Algorithmes employés : k plus proches voisins (k=1 et k=5) et classifieur bayésien naïf

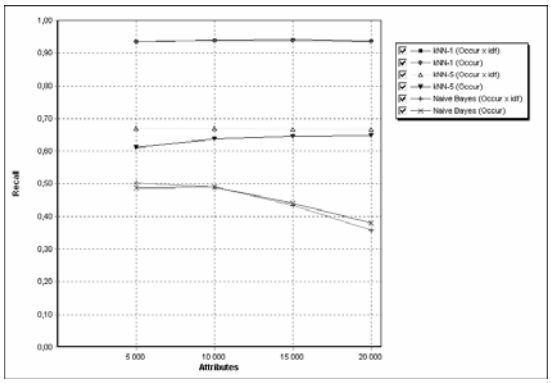

Figure 8. Tâche 2 – performances (rappel) du premier ensemble d'expérimentations d'apprentissage.

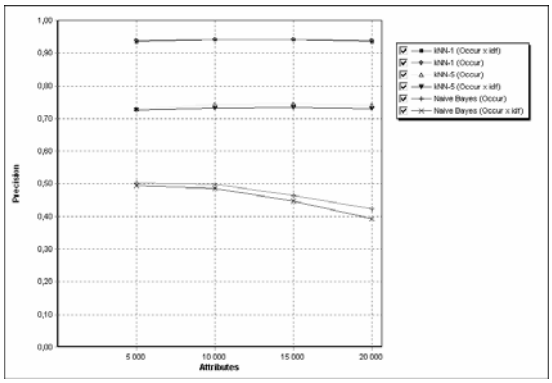

Figure 9. Tâche 2 – performances (précision) du premier ensemble d'expérimentations d'apprentissage.

Expérimentations d'apprentissage utilisant les paramètres suivants (figures 10 et 11) :

- Nombre de traits discriminants : entre 1 000 et 10 000 mots discriminants (avec un incrément de 1 000)
- Méthodes de représentation des traits : fréquence et fréquence pondérée par IDF
- Algorithmes employés : k plus proches voisins (k=1, k=2, k=3, k=4 et k=5)

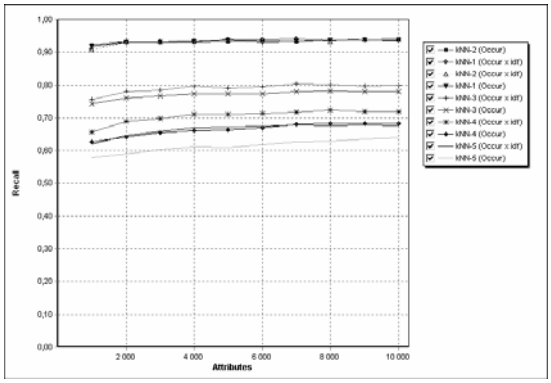

Figure 10. Tâche 2 – performances (rappel) du deuxième ensemble d'expérimentations d'apprentissage.

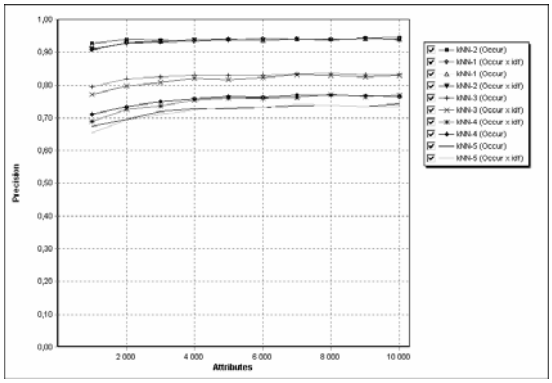

Figure 11. Tâche 2 – performances (précision) du deuxième ensemble d'expérimentations d'apprentissage.

Expérimentations d'apprentissage utilisant les paramètres suivants (figures 12 et 13) :

- Nombre de traits discriminants : entre 100 et 2 000 mots discriminants (avec un incrément de 1 00)
- Méthodes de représentation des traits : fréquence et fréquence pondérée par IDF
- Algorithmes employés : k plus proches voisins (k=1, k=2, k=3, k=4 et k=5)

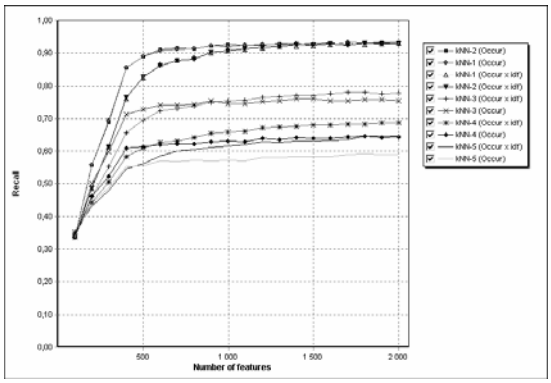

Figure 12. Tâche 2 – performances (rappel) du troisième ensemble d'expérimentations d'apprentissage.

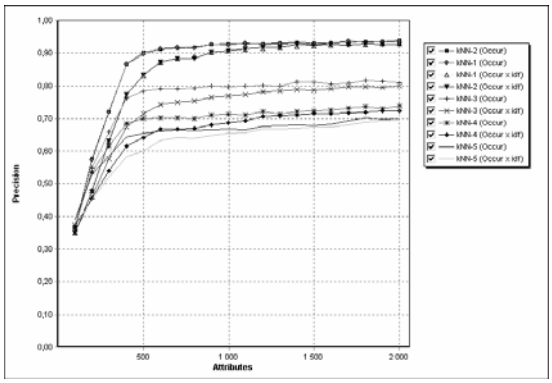

Figure 13. Tâche 2 – performances (précision) du troisième ensemble d'expérimentations d'apprentissage.

Durant la phase d'apprentissage de la tâche 3, les meilleures performances ont été obtenues en utilisant 10 000 mots discriminants représentés uniquement par leur fréquence. Ces mots ont été employés pour décrire les documents qui ont été soumis à l'algorithme des k plus proches voisins avec le paramètre k=2. Ainsi, selon en utilisant cet algorithme avec ce paramètre, nous avons obtenu un taux de rappel de 94.12% et un taux de précision de 94.53%. En contrepartie, les pires performances ont été obtenues en utilisant 100 mots discriminants représentés par leur fréquence. Ainsi, selon en utilisant cet algorithme avec ce paramètre, nous avons obtenu un taux de rappel moyen de 33.57% et un taux de précision moyen de 36.34%.

5.2.2 Phase de test

Lors de la phase de test, nous avons mené trois exécutions en faisant varier le nombre et la pondération des traits discriminants, ainsi que la valeur du paramètre *k*. Les résultats que nous avons obtenus lors de ces exécutions sont les suivants :

Exécution de test 1 utilisant les paramètres suivants :

- Nombre de traits discriminants : 5 000
- Méthodes de représentation des traits : fréquence pondérée par IDF
- Algorithmes employés : k plus proches voisins (k=1)

Performances globales : rappel = 32.20%, précision = 31.90%, f-mesure = 32.00%

Performances spécifiques (catégorie *ELDR*) : rappel = 18.90%, précision = 21.00%

Performances spécifiques (catégorie *GUE-NGL*) : rappel = 39.30%, précision = 34.50%

Performances spécifiques (catégorie *PPE-DE*) : rappel = 43.70%, précision = 44.70%

Performances spécifiques (catégorie *PSE*) : rappel = 36.00%, précision = 36.50%

Performances spécifiques (catégorie *VERT-ALE*) : rappel = 23.30%, précision = 22.60%

Exécution de test 2 utilisant les paramètres suivants :

- Nombre de traits discriminants : 10 000
- Méthodes de représentation des traits : fréquence
- Algorithmes employés : k plus proches voisins (k=2)

Performances globales : rappel = 33.20%, précision = 34.60%, f-mesure = 33.90%

Performances spécifiques (catégorie *ELDR*) : rappel = 23.10%, précision = 23.60%

Performances spécifiques (catégorie *GUE-NGL*) : rappel = 33.20%, précision = 42.20%

Performances spécifiques (catégorie *PPE-DE*) : rappel = 49.80%, précision = 45.20%

Performances spécifiques (catégorie *PSE*) : rappel = 39.40%, précision = 37.00%

Performances spécifiques (catégorie *VERT-ALE*) : rappel = 20.70%, précision = 25.20%

Exécution de test 3 utilisant les paramètres suivants :

- Nombre de traits discriminants : 15 000
- Méthodes de représentation des traits : fréquence pondérée par IDF

- Algorithmes employés : k plus proches voisins (k=1)

Performances globales : rappel = 33.30%, précision = 33.50%, f-mesure = 33.40%

Performances spécifiques (catégorie *ELDR*) : rappel = 20.20%, précision = 20.50%

Performances spécifiques (catégorie *GUE-NGL*) : rappel = 37.60%, précision = 38.40%

Performances spécifiques (catégorie *PPE-DE*) : rappel = 46.20%, précision = 46.20%

Performances spécifiques (catégorie *PSE*) : rappel = 38.30%, précision = 36.90%

Performances spécifiques (catégorie *VERT-ALE*) : rappel = 24.30%, précision = 25.50%

6 Discussions

Les résultats que nous avons obtenus lors de la phase d'apprentissage étaient des plus prometteurs (plus de 98% de rappel et de précision pour la tâche 1 et plus de 94% de rappel et de précision pour la tâche 3). Lors de la phase de test, les meilleurs résultats obtenus pour la tâche 1 sont de 77.30% au niveau du rappel et de 79.00% au niveau de la précision. Dans le cadre de la tâche 3, les meilleurs résultats obtenus sont bien en deçà des performances que nous étions en mesure d'obtenir lors de la phase d'apprentissage. Ainsi, pour la tâche 3, les meilleurs résultats obtenus sont de 33,30% au niveau du rappel et de 33,50% au niveau de la précision.

Les performances finales sont plutôt décevantes, surtout lorsque nous les comparons aux performances élevées obtenues lors de la phase d'apprentissage. Il est difficile d'identifier les causes exactes de ces performances finales. Il est cependant raisonnable de proposer que les modèles de catégorisation mis à l'épreuve lors de la phase de test sont caractérisés par un surapprentissage (*overfitting*). Ainsi, les modèles employés (surtout lors de la phase de test de la tâche 3) se sont avérés trop étroitement liés aux données initiales, peu généralisables et difficilement applicables aux données de test. Nous sommes d'avis que les performances observées découlent en partie des choix effectués lors du filtrage du lexique du corpus. Ainsi, il est probable que certains des mots qui ont été retenus pour décrire les documents en raison de leur présence élevée dans une des catégories (au moment de l'apprentissage) se sont avérés être beaucoup moins discriminants dans les données de test.

Une comparaison plus approfondie de la distribution des mots retenus dans les modèles de catégorisation à l'intérieur des données d'apprentissage et des données de test pourrait nous permettre de mieux comprendre pourquoi les performances initiales n'ont pu être reproduites sur les données de test. Nous sommes d'avis que cette piste d'explication pourrait être plus riche qu'une explication qui reposerait principalement sur l'algorithme de catégorisation employé.

En ce qui concerne les paramètres à spécifier lors de tâches de catégorisation automatique de documents textuels, les expérimentations que nous avons menées nous indiquent, dans un premier temps, que les performances des algorithmes sont en partie liées aux nombres de traits discriminants retenus pour décrire les documents à traiter. Ainsi, on constate une corrélation les performances et le nombre de traits discriminants employés. Plus le nombre de traits discriminants employés est élevé, meilleurs sont les performances du système. Au départ, cette amélioration est très prononcée. Elle devient plus subtile à partir de quelques milliers de traits discriminants (mais elle est néanmoins toujours présente). Ce phénomène a d'abord été observé sur les données d'apprentissage, puis, dans une moindre ampleur, sur les données de test.

Au niveau des algorithmes de catégorisation, l'algorithme des k plus proches voisins a toujours généré de meilleurs résultats que le classifieur bayésien naïf. Nos expériences nous portent donc à croire que

le classifieur bayésien naïf est peu efficace pour des tâches de catégorisation des documents textuels, lesquelles font inévitablement intervenir un nombre élevé de traits discriminants. En ce qui concerne l'algorithme des k plus proches voisins, nous constatons qu'il est optimal lorsque la valeur du paramètre *k* (paramètre spécifiant le nombre de « voisins » dans l'espace vectoriel auquel les éléments à catégoriser sont comparés) est très faible (idéalement 1 ou 2).

Références

Fayyad, U., G. Piatetsky-Shapiro et P. Smyth (1996), From data mining to knowledge discovery in databases, *AI Magazine*, vol. 1, pp. 37-54.

Forest, D. (2009, sous presse), Vers une nouvelle génération d'outils d'analyse et de recherche d'informations, *Documentation et bibliothèque*.

Juola, P. (2008), *Authorship attribution*. Now Publishers Inc.

Liu, B. (2007), *Web data mining: exploring hyperlinks, contents, and usage data*, London, Springer.

Manning, C. D. et H. Schütze (1999), *Foundations of statistical natural language processing*. Cambridge (Mass.): MIT Press.

Memmi, D. (2000), *Le modèle vectoriel pour le traitement de documents*. Grenoble, Cahiers Leibniz, no 2000-14.

Salton, G. (1989), *Automatic Text Processing*. Reading (Mass.), Addison-Wesley.

Document Level Subjectivity Classification Experiments in DEFT'09 Challenge

Cigdem Toprak and Iryna Gurevych

Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt
Hochschulstr. 10, 64289, Darmstadt, Germany
www.ukp.tu-darmstadt.de

Résumé – Abstract

Cet article présente nos expériences de classification supervisée pour la subjectivité au niveau des documents, pour l'anglais et pour le français, au cours du Défi DEFT'09 de fouille de textes. Nous avons testé des traits portant sur les *mots*, les *parties du discours* et sur des *vocabulaires spécialisés* pour faire fonctionner un classificateur SVM. Nos expériences sur les traits des *mots* examinent d'une part l'utilité de l'information contextuelle, et procèdent d'autre part à une comparaison, sur cette tâche, entre les représentations *binaires* et *tf*idf*. Nous montrons que des distributions différentes pour les classes privilégient des représentations différentes pour les traits. Puis, sur l'anglais, nous comparons trois vocabulaires spécialisés dans l'expression des opinions, deux d'entre eux étant bien connus. Ce sont les indices de subjectivité de (Wiebe et Riloff, 2005; Wilson et al., 2005), *SentiWordNet* (Esuli et Sebastiani, 2006), et une liste de verbes compilée à partir de (Santini, 2007 ; Biber et al., 1999). Malgré sa faible couverture, ce lexique de 156 verbes donne d'assez bons résultats pour l'anglais.

In this paper, we present our supervised document level subjectivity classification experiments for English and French at the DEFT'09 Text Mining Challenge. We experiment with the *word*, *POS*, and *lexicon-based* features using an SVM classifier. Our *word* feature experiments *(i)* investigate the utility of the context information, and *(ii)* compare the *binary* and *tf*idf* feature representations at this task. We show that different class distributions favor different feature representations. Furthermore, on the English collection, we compare three, two of which well-known, opinon lexicons at this task: the subjectivity clues from (Wiebe and Riloff, 2005; Wilson et al., 2005), *SentiWordNet* (Esuli and Sebastiani, 2006), and a list of verbs compiled from (Santini, 2007; Biber et al., 1999). We show that, despite its limited coverage, the verb lexicon, consisting of 156 verbs, establishes relatively good results in English.

Mots-clefs – Keywords

classification de textes, analyse supervisée de la subjectivité
text classification, supervised subjectivity analysis

1 Introduction

Distinguishing factual information from opinions plays a crucial role for many natural language processing applications in deciding which information to extract or retrieve or how to organize and present different types of information. For instance, an information retrieval system can aim at retrieving articles containing opinions in favor of a particular policy or decision, an information extraction system may need to extract only factual information, and a review aggregation system may require aggregating positive or negative opinions about a topic.

Subjectivity and sentiment analysis, a.k.a. opinion mining, are recent research directions focusing on the computational treatment of subjectivity, sentiments and opinions in text. Subjectivity analysis aims at classifying the content as objective vs. subjective. Sentiment analysis, on the other hand, involves several additional sub-tasks, such as: *(i)* determining the emotional orientation (polarity) of the subjective content, *(ii)* determining the strength of the polarity, *(iii)* determining the targets of the opinions in text, and *(iv)* determining the holders of the opinions in text.

Two of the DEFT'09 Text Mining Challenge tasks this year have focused on subjectivity analysis:

- **Task-1:** is a document level subjectivity classification task which required binary classification of the newspaper articles as *subjective* or *objective*.
- **Task-2:** is detecting the subjective parts of each individual document.

Our team participated in the first task in English and French. We used Support Vector Machine (SVM) classifiers (Joachims, 1998; Forman, 2003) as SVMs are shown to be among the top performer classifiers for high dimensional feature spaces as in the case of document level text classification. We utilized *word*, *part-of-speech (POS)*, and *lexicon-based* features in different configurations. *Lexicon-based* features were created using SentiWordNet (Esuli and Sebastiani, 2006), a list of subjectivity clues from previous works (Wiebe and Riloff, 2005; Wilson et al., 2005), and a list of verbs from (Santini, 2007; Biber et al., 1999). For our official submissions, we adopted a "kitchen sink" approach combining a variety of features. In this paper, besides our preliminary implementation employed for submissions, we report on additional experiments on the training and test corpora that investigate the contribution of various feature classes separately.

This paper is organized as follows: Section 2 introduces the related work in document level subjectivity classification. Section 3 explains our features. We discuss our experimental results in Section 4. Finally, we draw some conclusions in Section 5.

2 Related Work

Diverse features and classification algorithms have been investigated in document level subjectivity and sentiment classification tasks in previous works. Highest performance in document level subjectivity classification task for newspaper articles (F-measure 0.97) was established by (Yu and Hatzivassiloglou, 2003) using a Naive Bayes classifier with unigrams as features without stemming and stopword removal. (Wiebe et al., 2004) presents a detailed study for identifying *potential subjective elements*, i.e., subjective words and phrases, by clustering words according to their distributional similarity. They report accuracies up to 0.94 for document level subjectivity classification on a similar newspaper collection using the k-nearest neighbor algorithm based on the normalized counts of the *potential subjective elements* in each document. Similarly, we utilize *lexicon-based* features representing normalized counts of lexicon instances in a document.

(Pang et al., 2002) compared three classification algorithms, Naive Bayes, maximum entropy and SVM, with different feature configurations at a document level sentiment classification task for movie reviews. They show that using words as binary features performs better than using word frequencies as features. For French, we receive a better recall at the cost of a lower precision with the *binary* representation, however, for English frequency-based features (*tf*idf*) yield a better result than the *binary* features. Furthermore, they report that unigrams outperform bigrams in the same task. We confirm this finding for the English collection, however for the French collection, context information increases precision without damaging the F-measure. They also show that SVM is the best performer although not by a significant margin.

Subjectivity classification has its roots in genre classification. Similar to genre, subjectivity of documents can be regarded as orthogonal to the topic, i.e., an objective or a subjective document may have the same topic. (Finn and Kushmerick, 2003) view document level subjectivity classification as a genre classification task and aim at building domain independent subjectivity classifiers. They investigate the utility of three different types of features (bag-of-words, POS statistics and text statistics) across three domains for subjectivity classification. They show that bag-of-words performs best in single topic domains and worst in the cross domain experiments indicating that there are keywords conveying subjectivity within each topic domain. POS statistics yields the best results in cross domain experiments as it allows a better abstraction over a topic dependent model. We explore *POS* features in isolation and in combination with domain independent *lexicon-based* features. As our bag-of-word approaches, i.e., *word* features, outperform our domain independent lexicon or POS combinations, we also confirm that keywords play a crucial role at this subjectivity classification task.

In a document level sentiment classification task, (Généreux and Santini, 2007) explore the effect of different feature weighting schemes and the utility of macro-features called *linguistic facets*, which were shown to be effective in the Web genre classification by (Santini, 2007). *Linguistic facets* include features which can be functionally interpreted, e.g., high frequency of the first person pronouns indicate a argumentative style. We use some *linguistic*

facet features introduced in (Généreux and Santini, 2007) like the communication and mental verbs from (Biber et al., 1999) among our *lexicon-based* features.

3 Approach

We used an SVM^{perf} classifier[1][2] with a linear kernel. SVMs are *large margin* classifiers which aim to find a hyperplane (for two class problems) for separating the document vectors in one class from those in the other while keeping the separation, i.e., the *margin*, as large as possible. Classifying new documents is done by determining which side of the hyperplane they fall into.

Typically, in text classification documents are represented as vectors of feature counts. A feature can be as simple as the occurrence of a certain word or represent complex phenomena which can be observed in the document. For instance, a feature may represent the co-occurrence of a modal verb and a first person pronoun in the same sentence. There are different ways to represent feature counts. One way is to use a binary representation which indicates the presence (1) or absence (0) of the feature in the document. Another common approach is to represent each feature with a function of its frequency in the document. We explored both binary and frequency based representations in our experiments. For the frequency based representation, we used *tf*idf* (term frequency multiplied by inverse document frequency) as shown in the formula:

$$tf * idf = (1 + log(tf_{i,j}))log\frac{N}{df_i} \quad (1)$$

where $tf_{i,j}$ is the number of occurrences of $word_i$ in $document_j$, N is the total number of documents, df_i is the number of documents which $word_i$ occurs in.

We performed lemmatization, but applied no stop word removal. The documents are preprocessed with the TreeTagger[3] POS tagger (Schmid, 1994) and the Standford Named Entity Recognizer[4] (Finkel et al., 2005).

All of the features we used in our experiments can be grouped under three major classes as: *word features*, *POS features*, and *lexicon-based features*. Table1 illustrates all features used in our experiments. Next we explain each feature class in detail.

3.1 Word features

This feature class includes features representing each word as a feature. We investigated the contribution of context information as well as the effect of unigrams and bigrams in our different experiments. The context information is represented with the *word_window* feature, which encodes the previous and the next token of the current token. Feature *lemma_tfidf* represents the tf*idf values of lemmas as features. Similar to the *word_window* feature, *lemma_tfidf_window* represents the context of the lemma, but using tf*idf counts instead of the binary representation.

3.2 Lexicon-based features

Lexicon-based features are built based on three resources: the subjectivity clue lexicons from previous works (Wiebe and Riloff, 2005; Wilson et al., 2005), hereafter referred to as the *Wilson lexicon*, lexical semantic resource *SentiWordNet* created by (Esuli and Sebastiani, 2006) and a list of verbs taken from (Santini, 2007) which originates from (Biber et al., 1999), hereafter referred to as *C-M verb lexicon*.

Wilson lexicon consists of three lists of subjectivity clues: *(i)* the prior polarity lexicon, *(ii)* the intensifier lexicon, and *(iii)* the valence shifter lexicon. All three lexicons contain unigram as well as n-gram entries with *POS* and *stemming* attributes. The *POS* attribute indicates the POS of the subjectivity term. The *stemming* attribute indicates whether the look-up should be performed with lemmas or tokens. For instance, the look-up for the lexicon entry

[1] `http://svmlight.joachims.org/svm_perf.html`

[2] Classifier configuration: -c=1 -l=2 for English, -c=5 -l=2 for French. -c parameter represents the trade-off between training error and margin. -l parameter represents the loss function to use. We used the error rate, i.e., the percentage of errors in prediction vector as the loss function.

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

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

(word1=abuse pos1=verb stemmed1=y) should be performed with lemmas and match all the verb instances of the entry like "abused" (verb), "abusing" (verb), but not "abuse" (noun) or "abuses" (noun). Entries of the prior polarity lexicon also have the *prior polarity* and *reliability* attributes. *Prior polarity* represents the polarity of an entry out of the context with the possible values of *positive*, *negative*, *both* or *neutral*. The *reliability* attribute indicates whether the entry has a subjective usage most of the time *(strongsubj)*, or whether it has only certain subjective usages *(weaksubj)*. The intensifier lexicon contains a list of intensifier words such as "fierce, enormous, more, most". The valence shifter lexicon contains entries which shift the polarity of an existing opinion towards negative or positive including negation words.

In order to increase the coverage of the original lexicon described above, we looked up the verbs in the prior polarity lexicon in WordNet to check if they also existed as nouns. Eventually, we added 61 nouns with positive and 192 nouns with negative polarities to the original lexicon. Binary features generated from the Wilson lexicon contain the *reliability* and the *polarity* information. Real-valued features *Wilson_clue_count* and *Wilson_strongSubj_clue_ratio* represent the normalized count of lexicon instances in a document and the percentage of the *strongsubj* instances among all existing clues in a document respectively.

SentiWordNet (Esuli and Sebastiani, 2006) is a lexical resource which assigns a triplet of numerical scores for positivity *(PosScore)*, negativity *(NegScore)* and objectivity as *(1-(PosScore+NegScore))* to each synset in WordNet. Similar to the Wilson lexicon, *SentiWordNet* contains unigram as well as n-gram entries with the POS information besides the polarity scores. We used the *PosScore* and the *NegScore* of the first sense of the lexicon item as real-valued features. Similar to *Wilson_clue_count*, *SentiWN_count* represents the normalized count of the lexicon instances which have a non-zero *PosScore* or *NegScore* score for the first sense of the lexicon instance.

Communication and mental verbs (C-M verb lexicon): communication verbs include verbs like *say, claim, accuse etc.* which occur in reported speech and communication. Mental verbs include verbs conveying cognitive and emotional meaning such as *appreciate, love, judge etc.* C-M verbs have been taken from (Santini, 2007) who investigated them in the Web genre classification.

Wilson_1.person and *C-M_1.person* features assesses the co-occurrence of the lexicon instances and first person pronouns in the same sentence for the *Wilson* and the *C-M verb lexicons* respectively. Similarly, *Wilson_NE* and *C-M_NE* represents the co-occurrence of the lexicon instances and named entities in the same sentence.

Finally, the *C-M verb lexicon*[5] is manually translated to French, and it is the only lexicon used in the experiments for French.

Feature Class	Feature Name	Feature Type	Description
Word	word_lemma	binary	lemma of tokens
	word_window	binary	lemma of the previous and next 2 lemmas
	bigrams	binary	lemmas of the bigrams
	lemma_tfidf	real-valued	tf*idf value of the lemmas
	lemma_tfidf_window	real-valued	tf*idf of the previous and next lemma
POS	POS	binary	POS of the tokens
	POS_window	binary	POS of the previous and next token
	preceded_by	binary	whether token is preceded by an adj. or adv.
	POS_statistics	real-valued	number of pronouns, adj., adv. in each sent.
	modal_in_sentence	binary	existence of a modal verb in each sent.
Lexicon-based	Wilson	binary	existence of Wilson lexicon instances
	Wilson_NE	binary	Wilson word and named entity in the same sent.
	Wilson_1.person	binary	Wilson word and 1st person pr. in the same sent.
	Wilson_clue_count	real-valued	number of lexicon instances pro document
	Wilson_strongSubj_clue_ ratio	real-valued	ratio of strongSubj instances over all clues
	C-M	binary	existence of C-M lexicon instances
	C-M_NE	binary	C-M verb and named entity in the same sent.
	C-M_1.person	binary	C-M verb and 1st person pr. in the same sent.
	SentiWN_Scores	real-valued	positive and negative scores of the first sense
	SentiWN_count	real-valued	number of lexicon instances pro document

Table 1: An overview of features used in our experiments

[5]Both English and French versions of the C-M verbs can be found at: `http://www.ukp.tu-darmstadt.de/data/sentiment-analysis`

3.3 POS features

POS feature represents the POS of each token, and *POS_window* feature represents the POS of the previous and the next token as binary features. *Preceded_by* feature encodes whether a token is preceded by an adjective or an adverb. *Modal_in_sentence* feature looks for the existence of modal verbs in each sentence. Finally, *POS_statistics* feature assesses the number of pronouns, adjectives, and adverbs in each sentence.

4 Experiments

After the official submissions, we revised some parts of our system, performed additional experiments and evaluated them over the test collections. In this section, we present additional experiments as well as our submissions. Table 2 shows the number of documents in the training and test sets for both languages. The English collection had a more balanced class distribution (56% subjective vs. 44% objective) compared to the French collection (17% subjective vs. 83% objective). Documents were labeled as *subjective* or *objective* based on the sections they appeared in within the newspapers. Newspaper articles from the opinionated sections such as *letter from the editor*, *debates* and *analyses* were labeled as *subjective* and articles from the sections reporting facts such as *news in local and foreign politics and economy* were labeled as *objective*.

Language	Subjective	Objective	Total
English train	4426 (56%)	3440 (44%)	7866
English test	2977	2268	5245
French train	4338 (17%)	20838 (83%)	25176
French test	2894	13894	16788

Table 2: Document distribution for the training and test sets

Table 3 and Table 5 show our experimental results for the English and French collections respectively. Feature combinations used in our submissions for both languages are presented in Table 4. Precision for each class label is calculated as $P_i = \frac{correctly\ classified\ instances\ for\ class_i}{all\ instances\ classified\ as\ class_i}$ and recall for each class label is calculated as $R_i = \frac{correctly\ classified\ instances\ for\ class_i}{number\ of\ class_i\ instances\ in\ gold\ standards}$ where $class_i \in \{sub, obj\}$. Overall precision and recall is calculated as the arithmetic mean of the precisions and recalls for both class labels. *F-score* is calculated as $F = \frac{2XPrecisionXRecall}{Precision+Recall}$ using the overall precision and recall. Additionally, we report the accuracy for each experiment for enabling comparison to a random-choice baseline which would always assign the majority class to the instances.

We experimented with each feature class in isolation to understand their contribution for the specific classification task at hand. Next, we discuss our results for both collections.

4.1 Results

For both languages, we observe that *word* class features perform superior compared to the other classes. For the more or less balanced corpus of English, we observe that all groups, in isolation or combined, accuracy-wise outperform a random-choice baseline (56% accuracy). The French collection, on the other hand, has a high-skew class distribution in favor of the objective class, i.e., a random-choice baseline would already establish a high accuracy of 0.83. For the French collection, accuracy-wise only the *word* class features significantly exceed such a baseline.

Word class experiments: The experiments performed with *word* class features give us some insights about *(i)* how two different feature representations, i.e., *binary* vs. *tf*idf*, behave, and *(ii)* whether providing context information, i.e., using *bigrams* or *word_window* as opposed to using *unigrams* as features, would aid the document level subjectivity classification task.

*Binary vs. tf*idf:* For the English collection, we observe that the *tf*idf* representation (W4 in Table 3) outperforms the *binary* representation (W1 in Table 3) for all precision and recall values. On both collections, the *tf*idf* representation increases the subjective precision (Psubj), however, on the French collection at the cost of a dramatic decrease in the subjective recall (Rsubj) (F-W1 vs. F-W3 in Table 5). In other words, on the French collection which has a high-skew class distribution in favor of the objective class, the *binary* representation identifies more

instances of the minority class (high subjective recall), where the *tf*idf* representation is preciser. Frequency-based representations such as *tf*idf* are known to be effective for classical topic categorization tasks as they surface the content words, i.e., keywords. However, we speculate that for subjectivity classification on an uneven class distribution, binary representation may be surfacing the non-content words constituting evidence for the minority class, thus, increasing the recall for the minority class. However, this observation requires more investigation before drawing definite conclusions.

Context vs. no context: The binary features *word_window* and *bigrams* provide context information to the classifier. The *word_window* feature represents the previous and the next lemma as features. The *bigram* feature represents two consecutive lemmas as features. The experiments using *word_window* (W2 in Table 3) and bigrams (W3 in Table 3) are outperformed by the experiments using unigrams as features (W1). However, for the French collection, using *word_window* features (F-W2 in Table 5) increases the subjective precision and the objective recall compared to unigrams as features (F-W1). For the French collection, context information enables the classifier to make more precise decisions without damaging the recall too much. However, for the English collection, classifier does not benefit from the context information.

Lexicon-based class experiments: For the English collection, the experiments using the *lexicon-based* class features in isolation (L1-L7 in Table 3) aim at comparing three domain independent subjectivity lexicons. In other words, the lexicons contain no knowledge of the training or the test collections. The *Wilson lexicon* contains about 6850 unique entries from different POS classes, out of which 990 are multi-word expressions. The *C-M verb lexicon* has 156 verb entries. *SentiWordNet* assigns positivity and negativity scores to all synsets in WordNet Version 2.0, out of which around 9420 unigrams have non-zero subjectivity scores. As a result, *SentiWordNet* constitutes the largest resource, followed by the *Wilson lexicon*. Accuracy-wise, all lexicons perform well above the random-choice baseline proving their value for modelling subjectivity regardless of the domain. We see that the *Wilson lexicon* alone (L1) performs better than *SentiWordNet* alone (L5) and the *C-M verb lexicon* alone (L3). Adding complex features, which represent the co-occurrence of the named entities/first person pronouns and a lexicon item in the same sentence, improves the performance for the *C-M verb lexicon* (L3 vs. L4 in Table 3), but it does not contribute to the performance of the *Wilson lexicon* (L1 vs. L2 in Table 3). L3 and L4, the results obtained from the *C-M verb lexicon*, which is a limited lexicon restricted to verbs only, signal the potential of using verbs from certain semantic verb categories in subjectivity classification. From the results L1 (*Wilson lexicon*) and L5 (*SentiWordNet*), we observe that as the size of the lexicon increases, the noise introduced by the lexicon increases. This points out the importance of word-sense disambiguation in subjectivity classification. Finally, the best performance for the *lexicon-based* class experiments is obtained by combining all lexicons (L7 in Table 3).

For the French collection, we utilized the manual translations of the *C-M verb lexicon* (FL in Table 5). They proved to be insufficient for identifying subjective documents, establishing a low subjective recall.

POS class experiments: On both collections, the experiments with the *POS* class features (P1 in Table 3 and FP in Table 5) deliver similar results to the experiments using *C-M verb lexicon* (L4 in Table 3 and FL in Table 5). While for the English corpus, using *POS* class features alone establishes an accuracy significantly better than the random-choice baseline, for the unbalanced corpus of French, *POS* class behaves almost like a random-choice baseline.

Our submissions: We made three submissions for English and two submissions for French. The first submission in English (S1 for English in Table 4) combines one the best performing *word* class features, *lemma_tfidf*, with the *lexicon-based* class features. The second submission (S2 for English in Table 4) combines the *lexicon-based* class and the *POS* class features. The third submission is the best performing *word* class feature *lemma_tfidf_window*. For English, based on our revised system, S3 delivers the best results (SR_3 in Table 3). SR_1 in Table 3 (first submission) shows that the lexicons have almost no effect at all when combined with the *lemma_tfidf*. The results from the second submission (SR_2), which combines the *lexicon-based* class and the *POS* class, show that adding *POS* features damages the performance of the *lexicon-based* class. All lexicons contain POS information for their entries. Therefore, features created from lexicons already encode POS information rendering some POS features redundant which damages the performance of the classifier.

The first submission for French combines *lemma_tfidf* and the *lexicon-based* class features (S1 for French in Table 4). The *lexicon-based* features increases the subjective recall (F-SR_1 vs. F-W3 in Table 4). For the second submission, we combined the *lexicon-based* and *POS* classes, which performs better than each class in isolation.

Feature Class	Features	Psub	Rsub	Pobj	Robj	P	R	F	Acc
Baseline	assigning majority class	-	-	-	-	-	-	-	0.56
Word	W1: word_lemma	0.878	0.840	0.801	0.847	0.840	0.843	0.841	0.843
	W2: word_window	0.850	0.835	0.789	0.806	0.819	0.821	0.820	0.823
	W3: bigrams	0.876	0.831	0.792	0.845	0.834	0.838	0.835	0.837
	W4: lemma_tfidf	**0.896**	0.842	0.808	**0.872**	0.852	0.857	0.853	0.855
	W5: lemma_tfidf_window	0.893	**0.850**	**0.815**	0.866	**0.854**	**0.858**	**0.855**	0.857
Lexicon-based	L1: Wilson	0.848	0.801	0.757	**0.812**	0.802	0.806	0.803	0.806
	L2: L1, Wilson_NE, Wilson_1.person	0.835	0.814	0.764	0.790	0.800	0.802	0.801	0.804
	L3: C-M	0.738	0.722	0.645	0.663	0.691	0.693	0.692	0.697
	L4: L3, C-M_NE, C-M_1.person	0.748	0.737	0.661	0.675	0.705	0.706	0.705	0.710
	L5: SentiWN_scores	0.836	0.773	0.729	0.802	0.783	0.787	0.784	0.785
	L6: L2, L4	0.844	0.816	0.769	0.803	0.807	0.809	0.808	0.810
	L7: L5, L6	**0.849**	**0.829**	**0.783**	0.807	**0.816**	**0.818**	**0.817**	0.820
POS	P1: all POS group features	0.714	0.815	0.702	0.571	0.708	0.693	0.695	0.710
Submissions official	SO_1	0.836	0.863	0.812	0.778	0.824	0.821	0.822	-
	SO_2	0.791	0.819	0.751	0.716	0.771	0.767	0.769	-
	SO_3	0.783	0.931	0.880	0.662	0.832	0.796	0.814	-
Submissions revised	SR_1	0.876	0.841	0.801	0.843	0.838	0.842	0.840	0.842
	SR_2	0.839	0.824	0.775	0.792	0.807	0.808	0.807	0.810
	SR_3	0.893	0.850	0.815	0.866	0.854	0.858	0.855	0.857

Table 3: Experimental results for the English collection

Submission	English	French
S1	lemma_tfidf Wilson, Wilson_NE, Wilson_1.person C-M, C-M_NE, C-M_1.person SentiWN_Scores, SentiWN_count	lemma_tfidf C-M, C-M_NE, C-M_1.person
S2	Wilson_clue_count Wilson_strongSubj_clue_ ratio Wilson, Wilson_NE, Wilson_1.person C-M, C-M_NE, C-M_1.person SentiWN_Scores, SentiWN_count preceded_by, POS_statistics modal_in_sentence	C-M, C-M_NE, C-M_1.person preceded_by, POS_window modal_in_sentence
S3	lemma_tfidf_window	

Table 4: Features included in the submissions

5 Conclusions

In this paper, we presented our approach and experiments for the document level subjectivity classification task at the DEFT'09 Challenge which required the classification of the newspaper articles as *subjective* or *objective*. We experimented with an SVM^{perf} classifier using features from three different classes including *word*, *POS*, and *lexicon-based* features. We investigate how each feature class in isolation and combination with other classes performs at the subjectivity classification task on the two DEFT collections which have quite disparate class distributions.

Our experiments with the *word* class features reveal that different class distributions favor different feature representations at the document level subjectivity classification task. The English collection, which is almost balanced, benefits consistently from the *tf*idf* representation for all precision and recall values. For the unbalanced French corpus, the *binary* representation yields better subjective recall and the *tf*idf* representation yields better subjective precision. Additionally, with the *word* class experiments we assess the utility of the context information in the subjectivity classification task. We observe that for the English collection, context information does not contribute, wherein for the French collection, it increases the precision without damaging the F-measure.

The *lexicon-based* experiments investigate the utility of three domain-independent lexicons in the document level subjectivity classification task in English. The *Wilson lexicon* consists of the subjectivity clues from previous works (Wiebe and Riloff, 2005; Wilson et al., 2005). The lexical semantic resource *SentiWordNet* assigns a triplet of numerical scores for positivity *(PosScore)*, negativity *(NegScore)* and objectivity as *(1-(PosScore+NegScore))*

Feature Class	Features	Psub	Rsub	Pobj	Robj	P	R	F	Acc
Baseline	assigning majority class	-	-	-	-	-	-	-	0.83
Word	F-W1: word_lemma	0.787	**0.902**	**0.979**	0.949	0.883	**0.926**	**0.902**	0.941
	F-W2: word_window	0.861	0.816	0.962	0.972	0.911	0.894	**0.902**	0.945
	F-W3: lemma_tfidf	**0.920**	0.763	0.952	**0.986**	**0.936**	0.874	0.901	**0.947**
Lexicon	FL: C-M, C-M_NE, C-M_1st person	0.633	0.248	0.861	0.970	0.747	0.609	0.634	0.845
POS	FP: all POS group features	0.706	0.128	0.844	0.988	0.772	0.550	0.550	0.840
Submissions official	F-SO_1	0.783	0.122	0.844	0.993	0.814	0.557	0.662	-
	F-SO_2	0.527	0.752	0.943	0.860	0.735	0.806	0.769	-
Submissions revised	F-SR_1	0.886	0.789	0.957	0.978	0.921	0.884	0.901	0.946
	F-SR_2	0.676	0.359	0.878	0.964	0.777	0.661	0.694	0.859

Table 5: Experimental results for the French collection

to each synset in WordNet. The *C-M verb lexicon* constitutes a list of communication and mental verbs introduced by (Biber et al., 1999; Santini, 2007). For English, all three lexicons model document level subjectivity better than a random-choice baseline. Nevertheless, they lag behind the *word* class features due to: *(i)* the noise introduced by the lexicons, and *(ii)* the keywords which support the classification task and are incorporated by the *word* class features, but do not appear in the domain independent lexicons.

Acknowledgements

This work was supported by the German Research Foundation (DFG) as a part of the *Research Training Group on Feedback Based Quality Management in eLearning* under the grant 1223, and by the Volkswagen Foundation as a part of the Lichtenberg-Professorship Program under grant No. I/82806.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education Limited.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, pages 417–422, Genova, Italy.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, Ann Arbor, USA.

Finn, A. and Kushmerick, N. (2003). Learning to classify documents according to genre. In *Proceedings of the Workshop on Computational Approaches to Style Analysis and Synthesis at International Joint Conference on Artificial Intelligence*, Acapulco, Mexico.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Généreux, M. and Santini, M. (2007). Exploring the use of linguistic features in sentiment analysis. In *Proceedings of Corpus Linguistics Conference*, Birmingham, UK.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of European Conference on Machine Learning (ECML)*, pages 137–142, Chemnitz, Germany.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, Philadelphia, PA, USA.

Santini, M. (January 2007). *Automatic Identification of Genre in Web Pages*. PhD thesis, University of Brighton (UK).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing 2005: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 486–497, Mexico City, Mexico.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP'05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada.

Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 129–136, Sapporo, Japan.

Session IV - Subjectivité locale

DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique

Matthieu Vernier[(1)], Laura Monceaux[(1)] et Béatrice Daille[(1)]

[(1)]LINA - CNRS UMR 6241 – Université de Nantes
2, rue de la Houssinière BP 92208, 44322 NANTES CEDEX 03, France
Prenom.Nom@univ-nantes.fr

Résumé – Abstract

Nous présentons dans cet article le bilan de notre participation à la 5ème édition du *DÉfi Fouille de Textes* (DEFT'09). Nous participons à deux tâches parmi les trois tâches proposées dans le cadre de ce défi. La première consiste à catégoriser des textes journalistiques en deux classes : subjectif et objectif, et la seconde cherche à délimiter à un niveau de granularité le plus fin possible les passages subjectifs qui apparaissent dans des textes journalistiques et parlementaires. Pour réaliser ces tâches sur des textes en français, nous proposons deux méthodes basées sur la détection d'indices de différents niveaux linguistiques par une approche symbolique. Pour la tâche 1, nous utilisons ces indices comme attributs d'un texte dans une méthode d'apprentissage et de catégorisation automatique standard.
In this article, we present our contribution to the 5th *DÉfi Fouille de Textes* (DEFT'09). We take part in two tasks among the three tasks proposed in this challenge. The first task consist in a two classes text categorization : subjective and objective, and the second one try to achieve automatical annotations of subjective textual segments with a lower level of granularity. To realize these tasks on french texts, we propose two methods based on automatical annotations of linguistic clues with a symbolic approach, and on the use of these annotations as attributes in a standard classification algorithm.

Mots-clefs – Keywords

Subjectivité, fouille d'opinion, langage évaluatif, lexique, patron lexico-sémantique.
Subjectivity, opinion mining, appraisal language, lexical resource, semantic pattern.

1 Introduction

La cinquième édition de la campagne d'évaluation en fouille de textes DEFT porte principalement sur la fouille d'opinions en s'intéressant en particulier à la notion de subjectivité à travers deux tâches sur trois. L'opinion est un aspect fondamental dans notre société pour les personnes et les entreprises pour lesquelles l'avis du public est importante. Celles-ci ont besoin de se tenir au courant de l'évolution de leur image et des sujets qui intéressent la population pour s'adapter à leurs attentes et améliorer leur réactivité. Ces aspects impliquent particulièrement l'industrie des nouvelles technologies, la politique, la publicité, les médias ou la finance pour lesquels l'étude de l'opinion représente un enjeu et un pouvoir économique majeur. À l'heure du développement de la recherche d'informations, l'enjeu premier réside donc dans la création de programmes informatiques capables de détecter automatiquement les opinions ou évaluations émises à propos d'un sujet donné. Pour cela, avant même de déterminer automatiquement si une unité textuelle comporte une opinion, une première étape peut consister à observer si cette unité textuelle est exprimée de manière subjective (et donc naturellement propice aux opinions) ou objective.

Dans ce cadre applicatif, l'édition 2009 de DEFT propose trois tâches, réalisables dans trois langues (français, anglais, italien) :

– **Tâche 1** : la détection du caractère **objectif** ou **subjectif** de la *globalité* d'un texte. Cette tâche s'applique à des corpus d'articles de journaux français (*Le Monde*), anglais (*The Financial Times*) et italiens (*Il Sole 24 Ore*), Les articles sont extraits des rubriques : éditoriaux, débats, analyses, actualités en politique nationale/internationale

et économie. La référence est établie en suivant le type de rubrique ; la rubrique éditorial est par exemple considérée comme subjective car elle sert généralement à exprimer une opinion et à l'inverse, les actualités sont classées objectives car elles présentent des faits.
– **Tâche 2** : la détection des *passages* **subjectifs** d'un texte - que ce texte soit globalement objectif ou subjectif - s'applique aux mêmes corpus d'articles de journaux, et d'autre part à un ensemble de débats au parlement européen, en français, anglais et italien. La référence est établie par croisement entre les résultats des participants : les passages subjectifs sont les unités textuelles détectées comme telles par une majorité de participants. Le seuil de cette majorité est déterminé de manière empirique au vu des annotations produites par les outils des participants.
– **Tâche 3** : la détermination du parti politique auquel appartient l'orateur de chaque intervention dans le même ensemble de débats au parlement européen que précédemment. Le parti est à déterminer dans un ensemble fermé de partis européens.

Pour les linguistes et informaticiens-linguistes, un verrou scientifique majeur consiste à savoir comment modéliser la complexité du langage évaluatif et de l'expression de la subjectivité dans la langue, et plus complexe encore, comment en faire la détection et l'analyse automatique par des outils de traitements du langage. Dans le domaine du TAL, l'évolution des travaux en fouille d'opinions semble notamment guidée par une problématique : comment adapter des méthodes qui analysent un texte dans sa globalité vers des méthodes qui analysent séparement différents passages d'un texte avec un niveau de granularité plus précis ? En effet, les travaux de catégorisation de textes où il s'agît d'attribuer une étiquette Objectif/Subjectif ou Positif/Négatif/Neutre sont particulièrement classiques et s'adaptent bien à certains types de corpus monothématiques. Il peut ainsi s'agir de catégoriser des critiques de films, de livres, de produits technologiques (lecteurs MP3, ordinateurs portables, caméras, etc), de voitures, des album musicaux, des fiches de destinations de voyages touristiques selon la polarité positive, négative ou neutre de l'ensemble du document textuel. Ces textes, dont on sait à l'avance qu'ils vont être généralement subjectifs, évaluent un seul concept principal, cela a donc du sens de leur attribuer une étiquette dans leur globalité. En revanche, pour d'autres types de documents (des textes issus de blogs, de forums, d'émissions de télévisions, etc), il ne semble pas pertinent de chercher à les catégoriser dans leur globalité car leur contenu aborde différents sujets, alterne une énonciation subjective et objective et les opinions positives et négatives sont beaucoup plus facilement mélées. Quelques travaux un peu moins fréquents s'intéressent ainsi à catégoriser des unités phrastiques (Hu & Liu, 2004) ou intra-phrastiques (Whitelaw *et al.*, 2005) dans des problématiques de fouille d'opinions. Ce type de travaux, dans lequel nous nous positionnons, nécessitent de s'intéresser précisément à la nature des constituants du langage de l'évaluation et de la subjectivité pour pouvoir s'adapter à tout type de corpus.

Dans cet article, nous replaçons brièvement cette participation à DEFT dans le contexte de nos travaux actuels en fouille d'opinions en expliquant les motivations qui découlent naturellement pour ce défi. Nous rappelons également la définition théorique de la subjectivité dans la langue introduite par Benveniste (Benveniste, 1974). Cette définition a inspiré un courant de travaux francophones particulièrement riche (Charaudeau, 1992), (Galatanu, 2000), (Kerbrat-Orecchioni, 1997) en linguistique et nourrissent notre démarche pour accomplir du mieux possible la tâche 1 de catégorisation de textes Objectif/Subjectif et la tâche 2 de détection des passages subjectifs. Nous présentons et commentons les résultats obtenus par les deux méthodes que nous proposons sections 3 et 4.

2 Contexte motivant la participation à DEFT'09

2.1 Travaux reliés et tâches réalisées pour DEFT'09

La tâche 2, qui consiste à repérer les passages subjectifs d'un texte, suscite particulièrement notre intérêt. En effet, dans le cadre de travaux récents (Vernier *et al.*, 2009), nous cherchons à détecter des segments phrastiques ou intra-phrastiques exprimant une évaluation et à les catégoriser selon leur modalité (une opinion, un jugement, une appréciation, un accord, un désaccord), leur configuration d'énonciation (expression subjective explicite (prise en charge) ou expression subjective implicite (dissimulée)) et leur valeur axiologique (positive, négative ou ambigüe) tels que ces concepts sont définis dans les théories linguistiques de (Charaudeau, 1992) et (Galatanu, 2000). Un outil de détection et de catégorisation a ainsi été développé pour suivre l'évolution des passages évaluatifs exprimées dans les blogs francophones au fil des mois sur différents sujets et selon plusieurs problématiques :
– quels sont les sujets émergents de la blogosphère qui sont évalués positivement/négativement ?
– quel est précisément le vocabulaire évaluatif utilisé pour parler d'un sujet donné ?
– quels sont les sujets sur lesquels les internautes prennent en charge leur subjectivité ou au contraire cherche à la dissimuler ?

Dans ce cadre, la tâche 1 qui consiste à décider si un texte est globalement subjectif ou objectif nous intéresse également bien qu'étant un peu plus éloignée de nos problématiques actuelles. Elle nous semble néanmoins comporter quelques biais de part la nature du corpus considéré et le choix de la catégorie de référence : par exemple *s'agît t-il finallement de reconnaître automatiquement qu'un texte est subjectif ou bien de reconnaître qu'il s'agît d'un éditorial ?* Toutefois, la volonté d'adapter notre outil existant pour une tâche de catégorisation de textes et la curiosité d'observer l'utilité de la prise en compte de modèles théoriques sur la subjectivité nous amènent à proposer une première approche pour cette tâche.

Les notions d'évaluation et de subjectivité sont linguistiquement liées et il nous semble donc intéressant de réinvestir l'outil d'analyse des blogs dans le contexte proposé par DEFT'09 avec un minimum d'adaptations. L'objectif est ainsi de mesurer sa portabilité dans un tout autre genre de textes : les textes journalistiques et les débats parlementaires. Toutefois, les nuances entre évaluation et subjectivité imposent quelques adaptations en considérant et définissant précisément le concept de subjectivité.

2.2 Qu'est-ce que la subjectivité ?

La notion de subjectivité dans le langage a été découverte et introduite pour la première fois par Emile Benveniste (Benveniste, 1974). Pour Benveniste, la subjectivité dans le langage se définit comme « la capacité du locuteur à se poser comme sujet » dans son énoncé. La problématique de l'énonciation qu'il a développé, a rappelé la place de l'homme dans la langue : c'est dans et par la langue que l'homme se constitue comme *sujet* ; parce que le langage seule fonde le concept d'*ego*. Cette conception oriente l'auteur vers l'identification et l'analyse des marqueurs de subjectivité dans le discours. Les **déictiques**, indices de personnes, de temps et de lieu, retiennent alors son intérêt. Nous en détaillons une liste de marqueurs linguistiques dans la section 3. Toutefois, la langue offre de nombreuses autres possibilités, certes parfois moins explicites, pour mettre en scène le sujet dans sa relation à l'autre et au monde. Ces indices de construction identitaire et de prise en charge de l'énoncé appartiennent à la **modalité** et s'imposent à l'analyse comme traces de l'activité d'énonciation.

Dès 1932, le terme de modalité, initialement emprunté à la logique et récurrent dans la tradition grammaticale, a été introduit en linguistique. Les linguistiques soutiennent que l'énonciation d'un énoncé correspond à la communication d'une pensée distincte d'une pure et simple représentation. Le sujet pensant est indissociable de cette expression à laquelle il participe activement. Penser, « c'est donc juger qu'une chose est ou n'est pas, ou estimer qu'elle est désirable ou indésirable, ou enfin désirer qu'elle ne soit ou ne soit pas. On *croit* qu'il pleut ou on ne le *croit* pas, ou on en *doute*, on se *réjouit* qu'il pleuve ou on le *regrette*, on *souhaite* qu'il pleuve ou qu'il ne pleuve pas » (Bally, 1932). La modalité désigne donc l'attitude du locuteur dans l'activité d'énonciation.

Dans nos travaux en fouille d'opinion, nous nous sommes intéressés aux modalités du français plus récémment définies par (Charaudeau, 1992) et (Galatanu, 2000) mais qui suivent le courant initié par Benveniste sur la subjectivité et le langage évaluatif. Dans les exemples de modalités évaluatives du tableau 1, seul l'exemple 4 semble énoncé de manière objective. Bien que le verbe *mentir* soit un jugement axiologiquement négatif, le locuteur n'adopte pas d'attitude vis-à-vis de ce jugement et le présente de manière factuelle.

Exemple	Sur-modalité	Modalité
Je doute qu'il mente	Opinion faible explicite	Jugement implicite
Il est évident qu'il ment	Opinion forte implicite	Jugement implicite
Oui, c'est un menteur	Accord	Jugement implicite
Il ment		Jugement implicite
Je n'aime pas qu'il mente	Appréciation explicite	Jugement implicite

TAB. 1 – Exemple de discours évaluatif différent pour la même valeur axiologique *mentir*

A l'aide d'un lexique de 1115 termes axiologiques ou marqueurs de modalité et de 2830 patrons sémantiques, nous disposons d'un outil réalisant la détection et catégorisation de ces modalités. Nous revenons, au paragraphe 4.1, un peu plus précisément sur cet outil utilisé en particulier dans la méthode pour la tâche 2.

3 Tâche 1 : Catégorisation de textes Objectif/Subjectif

Afin de catégoriser automatiquement les textes du corpus « Journal » en deux classes (OBJECTIF et SUBJECTIF), l'approche que nous proposons se scindent principalement en deux axes :

– la représentation de chaque texte par un ensemble de descripteurs linguistiques,
– l'utilisation de ces descripteurs pour apprendre un modèle de classification.
Nous présentons les descripteurs considérés dans le paragraphe ci-dessous en décrivant leur pertinence par rapport au défi initial (reconnaître ce qui est subjectif de ce qui est objectif) et par rapport au biais induit par le corpus.

3.1 Choix des descripteurs

3.1.1 Descripteurs théoriques de la subjectivité

Notre point de départ consiste à suivre les théories linguistiques sur la subjectivité présentées dans la section 2 en considérant un certain nombre d'indices jouant un rôle dans l'expression de la subjectivité : les indices de personnes, les indices de temps et de lieu, les marqueurs de modalités, les valeurs axiologiques, les points d'exclamations et d'interrogations.

Les indices de personnes La construction des identités énonciatives dans le discours est le première indice de subjectivité selon Benveniste. Nous nous intéressons donc en premier lieu à la présence des pronoms et déterminants à la première personne dans le corpus :
– les pronoms personnels : *je, me, moi, nous* ;
– les pronoms possessifs : *le mien, la mienne, les miennes, le nôtre, la nôtre, les nôtres* ;
– les déterminants possessifs : *mon, ma, mes, notre, nos*
L'hypothèse consiste à considérer que ces marqueurs apparaissent plutôt dans des textes subjectifs. Les exemples suivants sont extraits du corpus :
– SUBJECTIF - *Ce constat n'est pas **le mien**, mais celui de Jean Hélène, que **j'**ai rencontré à Paris deux jours avant qu'il regagne la Côte d'Ivoire.*
– SUBJECTIF - *C'est **notre** proposition de sortie de crise.*
– SUBJECTIF - *il y a, à **mon** sens, le sentiment sous-jacent d'une menace apocalyptique*
La principale exception concerne le discours rapporté particulièrement présent dans les textes journalistiques. Des indices de personnes apparaissent également dans des textes OBJECTIF dans des passages entre guillemets.
– OBJECTIF - *« **Je** serai garant de l'intérêt européen et l'intérêt européen, c'est clairement un budget au-delà de 1% », a proclamé M. Barrot.*
– OBJECTIF - *« Ils ont détruit sa vie, et **la mienne** », a confié sa mère*
– OBJECTIF - *« Personne de **mon** village n'était entré au palais présidentiel »*
La nature du corpus fait qu'il existe d'autres exceptions que nous précisons dans le paragraphe 3.1.2.

Les indices de temps et de lieu relatifs Les indices de temps et de lieu relatifs, qualifiés d'ostension par Benveniste sont des unités linguistiques qui organisent les relations spatio-temporelles autour du JE, comme repère. On y trouve de nombreux termes ou unités comme : *ceci, ici* dont l'énonciation s'accompagne d'un geste de l'énonciateur, désignant l'objet dont il est question dans le discours produit par la subjectivité. Les unités linguistiques qui marquent le temps dans le discours (*maintenant, hier, l'an dernier*) n'ont d'existence que par rapport au présent d'énonciation et sont donc susceptibles de marquer la subjectivité.
– SUBJECTIF - *Grâce aux efforts qu'elle accomplira **d'ici là**, grâce aussi au soutien de ses amis européens, notamment de la France, elle sera au rendez-vous.*
– SUBJECTIF - *Il m'apparaît également indispensable que nous disions **dès maintenant** comment seraient utilisés les bénéfices éventuels qui résulteraient de l'organisation des Jeux de 2012.*
– SUBJECTIF - *Ceci expliquerait cela.*
Toutefois, le genre journalistique du corpus induit également la notion de regard du journaliste qui décrit les événements tels qu'il les voit. Ce regard est supposé objectif et l'usage d'indices de temps et de lieu relatifs est fréquent dans les articles classés OBJECTIF.
– OBJECTIF - *justifiant le statu quo de la BCE par les perspectives meilleures que prévu de croissance **cette année**.*
– OBJECTIF - *Le chef du NNP, Marthinus van Schalkwyk [...] devrait devenir membre officiel de l'ANC **d'ici quelques semaines**.*
Il est donc possible que ce type d'indice ne soit pas le plus discrimant pour effectuer la tâche 1 sur ce corpus.

Les modalités Nous avons présentés dans la section 2 en quoi certains verbes de modalité (*douter, penser, croire, reconnaître, être évident, etc*) jouent un rôle dans l'expression de la subjectivité dans la langue. Nous nous intéressons en particulier aux modalités d'**opinion**, d'**appréciation** et d'**accord-désaccord** décrites par Charaudeau. Dans le corpus « Journal » :
– SUBJECTIF - *Je le **regrette**.*
– SUBJECTIF - *on **doute** réellement de leur nécessité.*
– SUBJECTIF - *nous **croyons** que les prémisses d'un partenariat transatlantique fort consistent en une Europe stable.*
De la même façon, les passages rapportés entre guillemets contiennent également ce type d'indice.
– OBJECTIF - *« Je **doute** qu'il ait été convaincu par la seule force des arguments culturels »*

Les valeurs axiologiques L'axiologie recouvre la zone sémantique qui renvoie à l'idée de préférence et de rupture de l'indifférence. Elle est associée à une polarité positive/négative et comporte les évaluations référant aux champs d'expériences humaines : esthétique (beau/laid), pragmatique (utile/inutile, important/dérisoire, efficace/inefficace), cognitif ou intellectuel (intéressant/inintéressant), éthique ou morale (bien/mal, bon/mauvais), hédonique-affectif (agréable/désagréable, plaisir/souffrance).

Les termes axiologiques, qu'il s'agisse de noms (*richesse, élégance, luxe, éclat, mérite*) ou de verbes (*séduire, plaire, mentir*) servent donc à fournir un jugement de valeur. L'énonciateur se place dans un discours appréciatif. Cependant, un discours évaluatif appréciatif n'implique pas obligatoirement un discours explicitement subjectif. Les exemples du corpus en témoignent :
– OBJECTIF - *ce qui pourrait donner lieu à quelques **intéressants** apartés.*
– OBJECTIF - *Personnalité **séduisante**, sa proximité intellectuelle avec Jean Paul II [...] frappe tous les observateurs.*
Le journaliste présente les valeurs axiologiques *intéressant* et *séduisant* sans s'inclure dans l'énoncé, voire en prenant la précaution du conditionnel. Cependant la fréquence des termes axiologiques dans un même texte peut tout de même être un indice supplémentaire pour discriminer un texte subjectif, nous considérons donc ces indices comme des descripteurs potentiellement discriminants.

Les points d'exclamation et d'interrogation D'un point de vue discursif, l'interrogation et l'exclamation sont des marques de la présence du locuteur lorsqu'elles n'apparaissent pas dans des passages rapportés.
– SUBJECTIF - *Arrêtons, c'en est trop et gardons notre monopole !*
– SUBJECTIF - *Pourquoi faire croire que l'on fait oeuvre d'ouverture ou de compréhension en accueillant au sein de l'Eglise des intégristes patentés et qui le resteront ?*
Dans le corpus « Journal », la principale réserve que l'on peut émettre sur ce type d'indice de subjectivité concerne les articles de type *interview* pour lesquels un bon nombre de phrases interrogatives sont présentes sans pour autant qu'elles impliquent une subjectivité globale. Il s'agît là d'une particularité propre au corpus observé parmi plusieurs autres particularités que nous détaillons dans le paragraphe ci-dessous.

3.1.2 Descripteurs empiriques

Afin d'améliorer la catégorisation automatique de façon pragmatique, nous considérons également quelques caractéristiques supplémentaires pour décrire un texte. Ces caractéristiques s'éloignent quelque peu des définitions théoriques sur la subjectivité pour se rapprocher, de façon ad-hoc, des contraintes liées au corpus du Monde.

Les passages rapportés Comme nous l'avons observé précédemment, un locuteur utilise les citations lorsqu'il ne veut pas adopter d'attitude vis à vis d'un énoncé qui pourrait être axiologiquement positif ou négatif.
– OBJECTIF - *« Je voudrais que l'on comprenne bien que je n'ai aucun intérêt personnel. [...]»*
– OBJECTIF - *« Mugabe, assassin ! »*
Les unités textuelles issues de passages rapportés ou de citations ne doivent donc pas permettre de dire qu'un texte est globalement subjectif.

Les interviews Les interviews sont un type de texte du corpus qui perturbe grandement l'apprentissage. Ils sont en effet constitués d'un grand nombre d'indices subjectifs (phrases interrogatives, indices de personnes, modalités,

etc) mais sont pourtant classés comme étant OBJECTIF. De plus, les indices subjectifs n'apparaissent pas dans des passages rapportés entre guillemets dans les interviews.
– OBJECTIF - ***J'ai été particulièrement frappé*** *par un aspect du traité qui concerne les droits des salariés.*
– OBJECTIF - *Aucun homme au monde ne mérite ça* ***!***
Afin d'améliorer le modèle d'apprentissage, nous introduisons pour chaque texte un descripteur booleen indiquant s'il s'agit d'une interview ou non. Nous présentons dans le paragraphe 3.2.1 le module permettant de décider si un texte est une interview ou non.

Les courriers/éditoriaux signés Une partie des textes subjectifs correspondent à des courriers des lecteurs du Monde ou à des courriers de personnalités publiés en tant qu'éditoriaux ou articles longs. Ces textes sont en général signés par leur auteur qui exprime ainsi explicitement leur prise d'attitude par rapport à l'énoncé. Toutefois, il s'agît là d'indices de subjectivité valables sur ce corpus uniquement.
– SUBJECTIF - *SLAVOJ ZIZEK est philosophe, scénariste et psychanalyste slovène.*
– SUBJECTIF - *Pierre-Yves Gautier est professeur de droit civil à l'université Paris-II-Panthéon-Assas.*
– SUBJECTIF - *Fabio F.*

Les publications d'erratum du *Monde* Enfin, les textes où Le Monde prend l'attitude de reconnaître une erreur dans un article précédent sont très fréquents dans le corpus et sont classés subjectifs. Ces textes sont très courts (1 ou 2 phrases) et peuvent ne pas contenir beaucoup d'indices théoriquement subjectifs. Pourtant certains marqueurs sont assez efficaces pour repérer ce genre d'articles (3.2.1).
– SUBJECTIF - ***Contrairement à*** *ce que nous avons écrit dans Le Monde du 31 août*
– SUBJECTIF - *Silvio Berlusconi n'a pas promis d'abolir la taxe d'habitation, comme nous l'avons indiqué* ***par erreur***
– SUBJECTIF - ***Dans la légende*** *qui accompagnait l'article intitulé « Au Brésil, Trama ouvre de nouvelles pistes au disque » [...]*

3.2 Mise en oeuvre informatique

Pour la mise en oeuvre informatique des traitements sur le corpus « journal », nous utilisons la plateforme UIMA (Unstructured Information Management Architecture) avec laquelle nous avons précédemment développé l'outil d'annotation automatique des passages évaluatifs dans les blogs. Dans le paysage des solutions logicielles existantes qui offrent des moyens d'intégration, de développement et de déploiement, le « framework » Apache UIMA constitue l'une des solutions les plus avancées et des plus prometteuses. Son objectif est de permettre l'utilisation et la construction d'applications distribuées visant l'analyse de contenus multimédias non structurés. Initié par IBM (Ferruci & Lally, 2004), l'implémentation d'UIMA est aujourd'hui un projet en incubation au sein de l'ASF (Apache Software Foundation). Les principes de gestion de l'information non structurée (recherche sémantique et analyse de contenu) font l'objet d'un effort de standardisation de la part d'un comité technique de l'OASIS (Organization for the Advancement of Structured Information Standards). Nous présentons brièvement deux éléments de base de UIMA pour faciliter la compréhension de notre chaîne de traitement (voir figure 1) :

– les **composants d'annotations** sont utilisés pour analyser des documents afin de détecter des attributs descriptifs sous forme de métadonnées. Un document dans UIMA est une unité de contenu qui peut contenir soit du texte, de l'audio ou de la vidéo. Les métadonnées peuvent concerner des énoncés décrivant des régions d'une façon plus granulaire que le document source. Un composant d'annotations peut réutiliser les annotations apportées par les composants précédents.
– le **CAS** (Common Analysis Structure) est la structure qui permet de représenter et partager les résultats d'analyse entre les composants, il s'agit d'une structure de données pour représenter le document, les annotations et leur structure de traits correspondantes. UIMA fournit des types d'annotation de base mais peuvent être étendus par les développeurs pour aboutir à un schéma plus riche de types, appelé Type System (TS). Un TS est spécifique à un domaine ou une application, et les types dans un TS peuvent être organisés dans une taxonomie. Dans notre étude, nous possédons notamment les types d'annotations suivants : *paragraphe, phrase, mot, passage rapporté, indice de personne, indice de temps et lieux, structure évaluative, interview, signatures et erratum*.

Nous décrivons ci-dessous quelques composants d'annotations développées pour le défi.

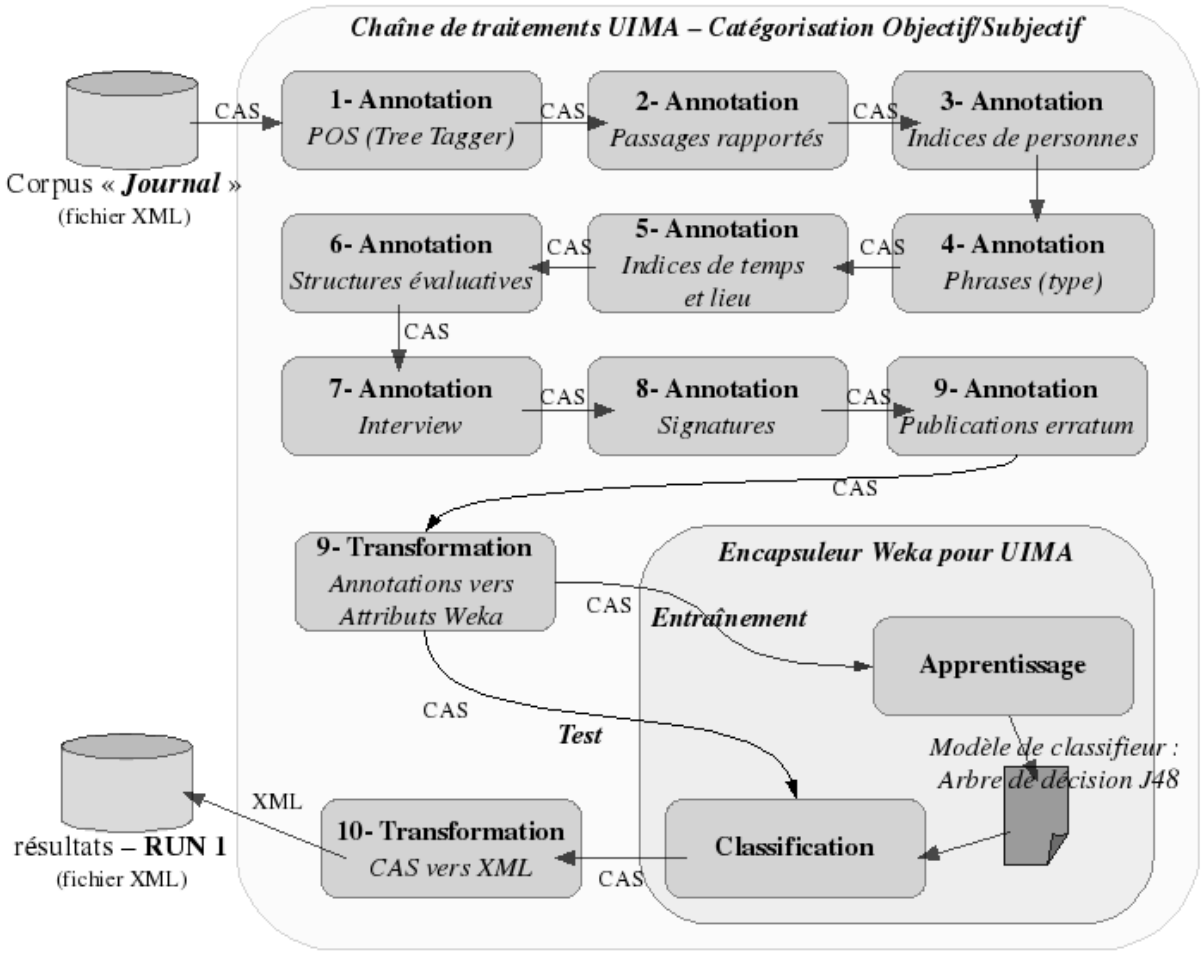

FIG. 1 – Chaîne de traitements UIMA : Annotations d'indices textuels pour la classification supervisée de textes Objectif/Subjectf.

3.2.1 Composants d'annotations UIMA

1- Etiquetteur grammatical Nous utilisons le TreeTagger de H. Schmid à travers un composant UIMA pour annoter les mots et leur associer un certain nombre de traits (catégorie grammaticale, lemme, temps, genre, etc). En sortie de ce composant, le CAS est donc constitué des annotations de type *mot* en plus du texte du corpus.

2- Passage rapporté Pour détecter les passages rapportés du corpus, le composant annote chaque passage contenu entre un guillemet ouvrant et un guillemet fermant. Le corpus du Monde est particulièrement peu bruité et en permet une détection efficace.

3-5 Indice de personne, de temps et de lieu Les indices de personne sont détectés en utilisant les annotations *mot* posées par le composant 1. Chaque mot est comparée à une liste de marqueurs de personne construite manuellement et comportant une dizaine d'entrée.
exemples : *je, nous, notre, le mien, etc*
De la même façon les indices de temps et de lieu sont comparés à la liste d'annotations de type mot ou à des suites de mots pour repérer les mots ou expressions qui appartiennent à une liste de marqueurs : exemples : *d'ici là, hier, maintenant, etc*

4- Phrases et paragraphe Les paragraphes sont annotées à partir des balises XML <p> du corpus original. Les phrases sont annotées à partir des signes de ponctuations et des annotations de type paragraphe (les sous-titres des articles du Monde ne contiennent pas de signe de ponctuation finale mais sont considérés comme des paragraphes).

6- Structure évaluative La détection des structures évaluatives est présentée pour la réalisation de la tâche 2 au paragraphe 4.1.

7- Interview Les interviews sont détectées grâce aux annotations de type *phrase* et *paragraphe* posées précédemment et quelques heuristiques correspondantes aux structures des interviews dans Le Monde. Nous considérons par exemple qu'une interview est composée :
– d'au moins 4 phrases interrogatives séparées par des phrases déclaratives,
– et que les phrases interrogatives doivent être réparties sur l'ensemble du texte et non dans un seul paragraphe.
À partir d'un échantillon de 200 textes contenant des phrases interrogatives, dont 100 interviews, nous avons évaluer la précision (0.96) et le rappel (0.76) de ce composant sur la tâche de catégorisation : interview/non interview. Les articles de type interview du Monde ont généralement une structure similaire, mais certaines interviews courtes (moins de 4 questions) ne sont pas détectées.

8- Signature Le composant de détection de signatures à la fin des textes s'appuie sur les annotations de type *phrase* et *paragraphe* et sur d'autres heuristiques. Il s'agît d'une signature si le dernier paragraphe d'un document contient moins de 2 phrases et comporte des marqueurs :
– noms avec majuscule,
– expressions (*est professeur, est philosophe, est ministre, etc*)
– adresse email (@, etc).

9- Erratum Selon le même principe, les textes publiés par le Monde signalant une erreur de publication sont des textes courts (un seul paragraphe) et doivent comporter des expressions spécifiques : *Contrairement à*, *dans la légende*, *par erreur* etc.

3.2.2 Encapsuleur Weka dans UIMA

Dans cette même chaîne de traitements, un dernier composant UIMA utilise l'API de weka[1] pour créer un modèle de catégorisation pendant la phase d'apprentissage et pour catégoriser les textes durant la phase de test. Pour créer ce modèle, nous transformons en attributs les annotations du corpus ajoutées au CAS. Chaque texte est ainsi représenté par un certain nombres d'attributs numériques normalisés par rapport au nombre de mots du texte :
– le nombre d'indices de personnes n'apparaissant pas dans un passage entre guillemets,
– le nombre de passages entre guillemets,
– le nombre de phrases interrogatives,
– le nombre de phrases exclamatives,
– le nombre de phrases interrogatives et exclamatives dans le dernier paragraphe,
– le nombre de modalités d'opinions,
– le nombre de modalités d'appréciations explicites,
– le nombre de modalités d'accord et de désaccord,
– le nombre de termes axiologiques,
– le nombre de termes axiologiques dans le dernier paragraphe,
et des attributs booléens :
– le texte est-il une interview ?
– le texte possède t-il une signature ?
– s'agît t-il d'une publication d'erreur du Monde ?

Résultats Afin d'évaluer l'efficacité de notre méthode durant la phase d'entraînement, nous utilisons la technique de validation croisée à 10 tours. Parmi les algorithmes de classification proposés par Weka (Witten & Frank, 2005), l'algorithme J48 est celui qui a obtenu les meilleurs résultats (voir tableau 2). J48 est une mise en oeuvre de l'algorithme d'arbre de décisions C4.5 de Quinlan (Quinlan, 1993). Finallement, les résultats obtenus sur le corpus test (voir tableau 3) restent assez stables bien que le rappel des textes subjectifs baisse. Cet algorithme permet également d'observer que les descripteurs les plus discriminants sont en premier lieu les interviews, les publications d'erreur du monde, les signatures, puis le nombre d'indices de personnes, le nombre de phrases interrogatives et d'indices de modalités d'opinion.

[1] http ://weka.sourceforge.net/doc/

Run	Précision	Rappel	FScore strict
1 - Corpus 1 (Journal)	90.8%	80.8%	85.5%
Objectif	91.7%	97.5%	-
Subjectif	89.8%	64.1%	-

TAB. 2 – Résultats obtenus par cross-validation pour la tâche 1 lors de l'entraînement

Run	Précision	Rappel	FScore strict
1 - Corpus 1 (Journal)	91.5%	79.3%	85.0%
Objectif	92.2%	98.7%	-
Subjectif	90.7%	60.0%	-

TAB. 3 – Résultats obtenus pour la tâche 1

4 Tâche 2 : Détection des passages subjectifs

4.1 Notre outil de détection des évaluations dans les blogs

Dans le cadre du projet ANR 2006 Blogoscopie, nous avons élaboré un outil de détection et de catégorisation de structures évaluatives dans un corpus de blogs multi-domaines ; tels que les opinions, les appréciations et les accord-désaccord, comme définit par (Charaudeau, 1992). Cet outil repose sur l'apprentissage de structures évaluatives à partir d'un corpus annoté de 200 billets issus de blogs multi-domaines où 4945 passages évaluatifs (chaînes symboliques) ont été annotés manuellement.

Apprentissage de structures évaluatives Afin de détecter le plus d'évaluations possibles, les chaînes symboliques issues de l'annotation manuelle ont été généralisées en suivant ces différentes étapes :

- Généralisation de la valeur des traits **axiol**, **forme** et de **lemme** (Y sur la fig. 2) pour tous les symboles de **type** *évaluation* et de **modalité** *appréciation*,
- Généralisation de la valeur du trait **lex** (X sur la fig. 2) pour certains symboles (adverbe, pronom ...) ainsi que le trait **lem** pour les symboles de type adverbe
- Ajout de l'opérateur standard ***** sur les symboles de **type** *intensité* et généralisation de la valeur des traits **forme** et de **lemme** de ces symboles,
- Ajout de l'opérateur standard **+** (une ou plusieurs fois) pour les symboles de **config** *explicite* et de **pos** *pronom* et généralisation de la valeur des traits **forme** et de **lemme** de ces symboles.

La figure 2 représente la structure évaluative généralisée apprise à partir de l'annotation *n'est-ce pas plus original* et qui permet par exemple lors de la détection d'annoter également les annotations suivantes : *n'est-ce pas plus banal*, *ne semble pas plus original*, etc. A l'issu de cette généralisation, nous disposons ainsi de 2830 structures évaluatives permettant de détecter les évaluations présentes dans un texte.

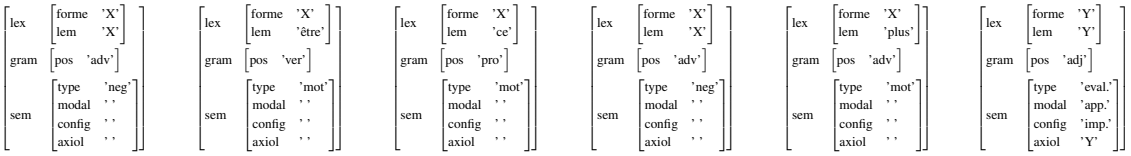

FIG. 2 – Structure évaluative apprise à partir de l'annotation de *n'est-ce pas plus original*.

Pour plus de détails sur la phase d'apprentissage, nous vous invitons à consulter (Vernier *et al.*, 2009).

Détection des évaluations Pour la détection des évaluations mais également pour l'apprentissage de structures évaluatives, trois ressources lexico-sémantiques ont été élaborées semi-manuellement à partir des annotations du corpus d'entraînement :

- un **lexique de l'évaluation** (1115 entrées), développé par Sinequa, contenant les termes évaluatifs, associées à leur catégorie grammaticale, leur modalité, leur énonciation et leur axiologie. ex : *machiste*, *chapeau bas*, *douter*,
- un **lexique de l'intensité** (21 entrées) ex : *particulièrement*, *très*,

– un **lexique de la négation** (15 entrées) ex : *pas*, *aucun*,

La détection des évaluations a pour objectif d'annoter les segments évaluatifs au niveau intra-phrastique. Avant de rechercher ces segments, le corpus est pré-traité par une projection des différents lexiques (évaluation / intensité / négation) et étiqueté morpho-syntaxiquement via le TreeTagger (composant 1 de la figure 3).

Pour chaque phrase du corpus à annoter, la stratégie du composant de détection (composant 2 de la figure 3) réside dans l'algorithme suivant :

Pour chaque phrase du corpus **Faire** :
– Transformation de la phrase en chaîne symbolique (n symboles dans la phrase),
– Recherche des chaînes évaluatives présentes dans la phrase :
  – i = 1 (i étant la position du symbole courant)
  – **Tant que** i <= n **Faire** :
    – Recherche d'unification d'une chaîne symbolique à partir du symbole courant (en position i) avec les structures évaluatives apprises (la plus longue possible)
    – **Si** unification possible entre i et j **Alors** annotation de la chaîne et i = j+1
    – **Sinon** i = i+1 (on regarde le symbole suivant) **FinSi**
  – **Fin Tant que**
FinPour

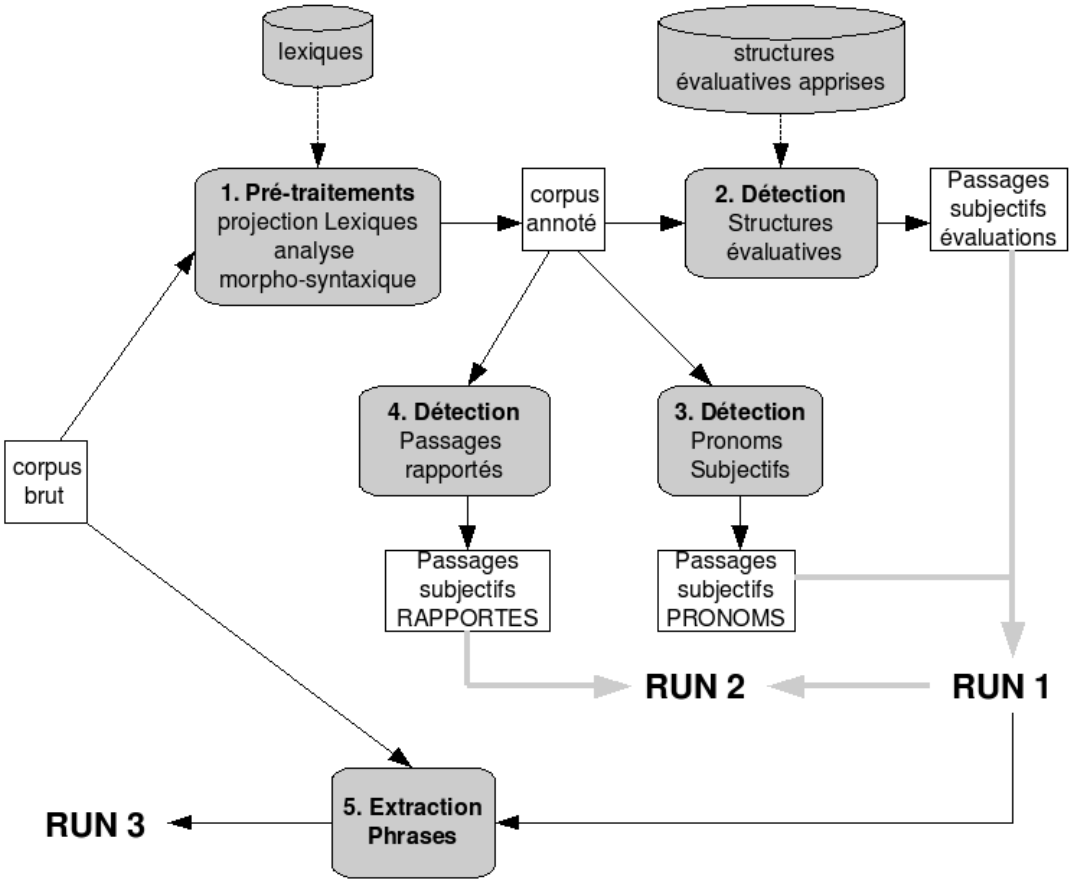

FIG. 3 – Chaîne de traitements pour obtenir les 3 fichiers de run pour la tâche 2

4.2 Adaptation de notre outil et résultats

Afin d'évaluer notre outil sur d'autres types de corpus que les blogs, nous avons décidé de participer à la tâche 2 de la campagne DEFT en adaptant notre outil (voir figure 3).

Adaptations Suite à l'étude du corpus d'apprentissage fourni par la campagne DEFT 2009, d'autres passages subjectifs que ceux définis dans notre outil de d'etection nous ont semblé pertinents à prendre en compte :
– tous les passages rapportés (passages entre guillemets) (composant 4 de la figure 3)
– tous les pronoms subjectifs (je,me,nous ...) (composant 3 de la figure 3)
Dans le cadre de la tâche 2 de la campagne DEFT, nous avons proposé trois runs. Le premier run que nous avons proposé correspond à l'union des passages évaluatifs détectés par notre outil de détection des évaluations et des pronoms subjectifs annotés par le composant 3 (RUN 1). L'outil de détection des passages rapportés n'étant à l'heure actuelle que les passages entre guillemets, nous l'avons inclus aux résultats du RUN 1 pour fournir un deuxième RUN (RUN 2).

Suite à la définition d'un passage subjectif dans la campagne DEFT : un mot ou une phrase ; nous avons proposé un dernier run (RUN 3) où chaque passage subjectif détecté dans le RUN 1 a été étendu à la phrase (composant 5 de la figure 3).

Résultats Le tableau 4 ci dessous représente les résultats obtenus de chaque RUN pour chaque corpus : le corpus journalistique (Corpus 1) et le corpus de débats parlementaires (Corpus 2).

Run	Précision	Rappel	FScore strict
1 - Corpus 1	92.8%	52.4%	67%
2 - Corpus 1	62.3%	62.3%	62.3%
3 - Corpus 1	80.8%	92.6%	86.3%
1 - Corpus 2	80.5%	54.3%	64.8%
2 - Corpus 2	80.4%	54.3%	64.8%
3 - Corpus 2	90.3%	91.6%	90.9%

TAB. 4 – Résultats obtenus pour la tâche 2

Même si nous ne connaissons pas les résultats des autres participants, les résultats obtenues par les 3 runs permettent plusieurs remarques :

– L'ajout des passages rapportés comme passages subjectifs ne semble pas efficace. Dans les débats parlementaires, cet ajout ne pose pas de problèmes car les passages rapportés y sont moins utilisés ; mais on constate pour les corpus journalistiques que la précision chute de manière importante par rapport à notre premier RUN (- 30 %), même si on trouve plus de passages subjectifs (+ 10 % en rappel). Toutefois, l'outil de détection des passages rapportés n'est pas optimisé à l'heure actuelle.
– L'extension des passages subjectifs du RUN 1 à la phrase (RUN 3) permet d'augmenter de manière impressionnante le rappel (+ 40 %) et cela quelque soit le corpus. Mais la précision fluctue selon le corpus de plus ou moins 10 %.

5 Conclusion

Dans ce défi nous nous sommes particulièrement intéressé aux constituants de la subjectivité dans le langage en prenant comme point de départ la théorie de Benveniste est les modalités définies par Charaudeau et Galatanu. La notion de subjectivité est encore particulièrement débattue y compris pour des analyses manuelles en linguistique, par conséquent l'analyse automatique de la subjectivité reste un défi important et inachevé. La question de l'évaluation des méthodes de détection et de catégorisation automatique se pose également : comment établir la catégorie de référence d'un texte si l'on souhaite s'intéresser à des corpus d'un autre domaine que journalistique ? La méthode d'évaluation de la tâche 2 proposée dans ce défi est certainement critiquable car les scores sont potentiellement influencés par le nombre de participants et le nombre de participants qui ont choisis des méthodes approchantes. Toutefois ce choix d'évaluation soulève un problème intéressant : est-il possible d'établir une référence fiable pour évaluer des méthodes d'annotations de passages subjectifs sans recourir à une phrase d'annotation manuelle coûteuse ? et dès lors, comment disposer de large corpus d'entraînements et de tests similaires à ceux disponibles pour la catégorisation de textes ?

En ce qui concerne les résultats de ce défi, sans connaître le nombre de participants et leurs résultats, il est difficile de se faire une idée précise de la difficulté de la tâche et de la réussite ou non de nos contributions. Néanmoins, nous observons que les résultats de tâche 2 de détection des passages subjectifs dans des articles du Monde et de débats parlementaires (F-Score strict entre 0.86 et 0.91) sont comparables aux résultats que nous obtenons en détectant les passages évaluatifs sur les blogs, voire meilleurs en terme de rappel. De ce point de vue, ces résultats

semblent intéressants pour notre problématique qui consiste à analyser les évaluations et la subjectivité dans des textes sans contrainte de domaine thématique.

Références

Bally C. (1932). *Linguistique générale et linguistique française*. Francke.

Benveniste E. (1974). *Problèmes de linguistique générale II*. Gallimard edition.

Charaudeau P. (1992). *Grammaire du sens et de l'expression*. Hachette Education, COMMUNICATION, PARA UNIVERSITAIRE.

Ferruci D. et Lally A. (2004). Uima : an architectural approach to unstructured information processing in the corporate research environment. In *Natural Language Engineering,10(3-4)*, p. 327–348.

Galatanu O. (2000). Signification, sens, formation. In *Education et Formation, Biennales de l'éducation, (sous la direction de Jean-Marie Barbier, d'Olga Galatanu)*, Paris : PUF.

Hu M. et Liu B. (2004). Mining and summarizing customer reviews. *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining (KDD)*, p. 168–177.

Kerbrat-Orecchioni C. (1997). *L'Énonciation, de la subjectivité dans le langage*. Colin (réédition 2002).

Quinlan R. (1993). C4.5 : Programs for machine learning. In *Morgan Kaufman Publishers*.

Vernier M., Monceaux L., Daille B. et Dubreil E. (2009). Catégorisation des évaluations dans un corpus de blogs multi-domaine. In *Numéro spécial de la revue RNTI (Revue des Nouvelles Technologies de l'Information) - fouille de données d'opinion, à paraître*.

Whitelaw C., Garg N. et Argamon S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*, p. 625–631 : ACM.

Witten I. H. et Frank E. (2005). Data mining : Practical machine learning tools and techniques. In *2nd Edition, Morgan Kaufmann*.

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.

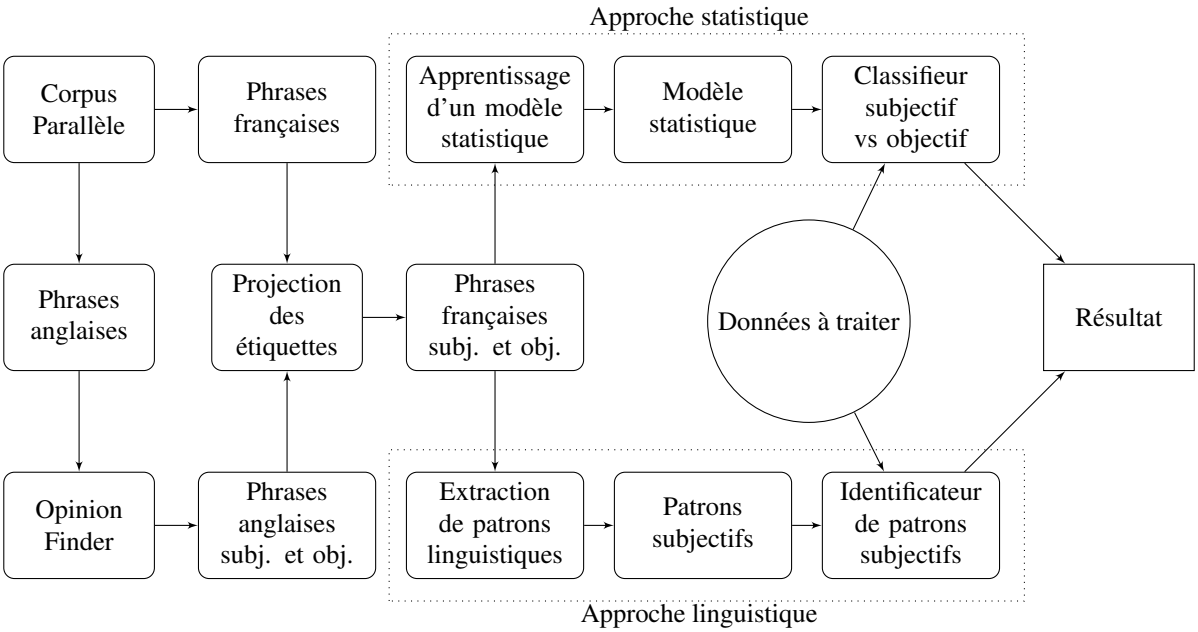

Figure 1: Une approche mixte pour la détection de subjectivité

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.

Index des auteurs

Index des mots-clés