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

Matthieu Vernier, Laura Monceaux, Béatrice Daille

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.

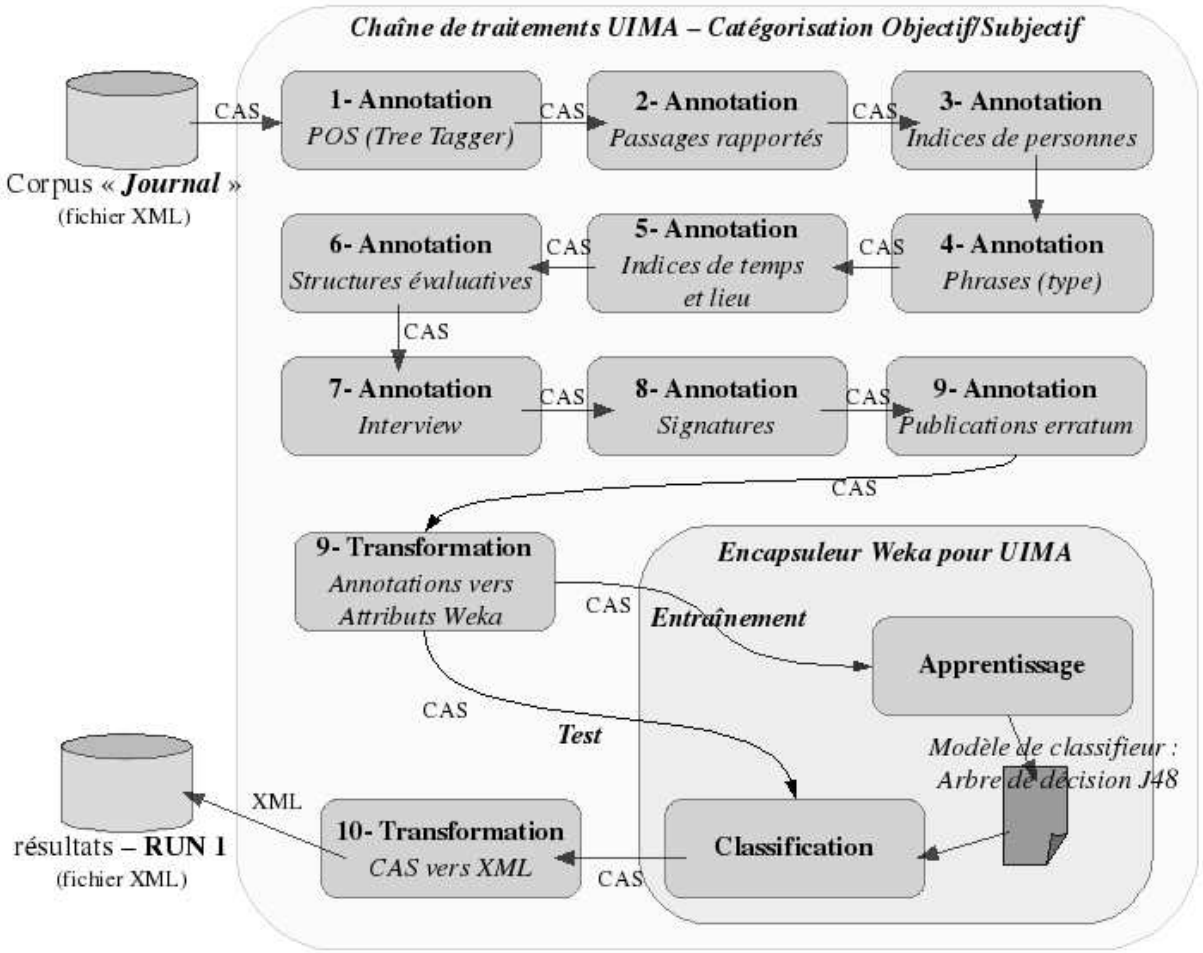

FIG. 1 – Chaîne de traitements UIMA : Annotations d'indices textuels pour la classification supervisée de textes Objectif/Subjectf.

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.

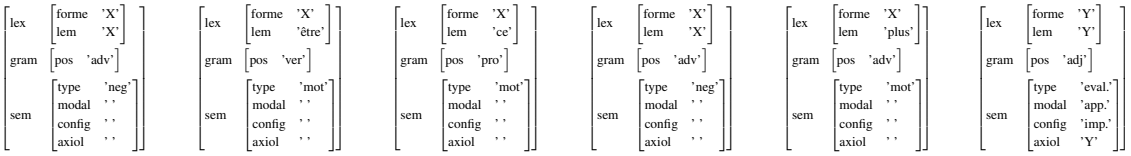

FIG. 2 – Structure évaluative apprise à partir de l'annotation de *n'est-ce pas plus original*.

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

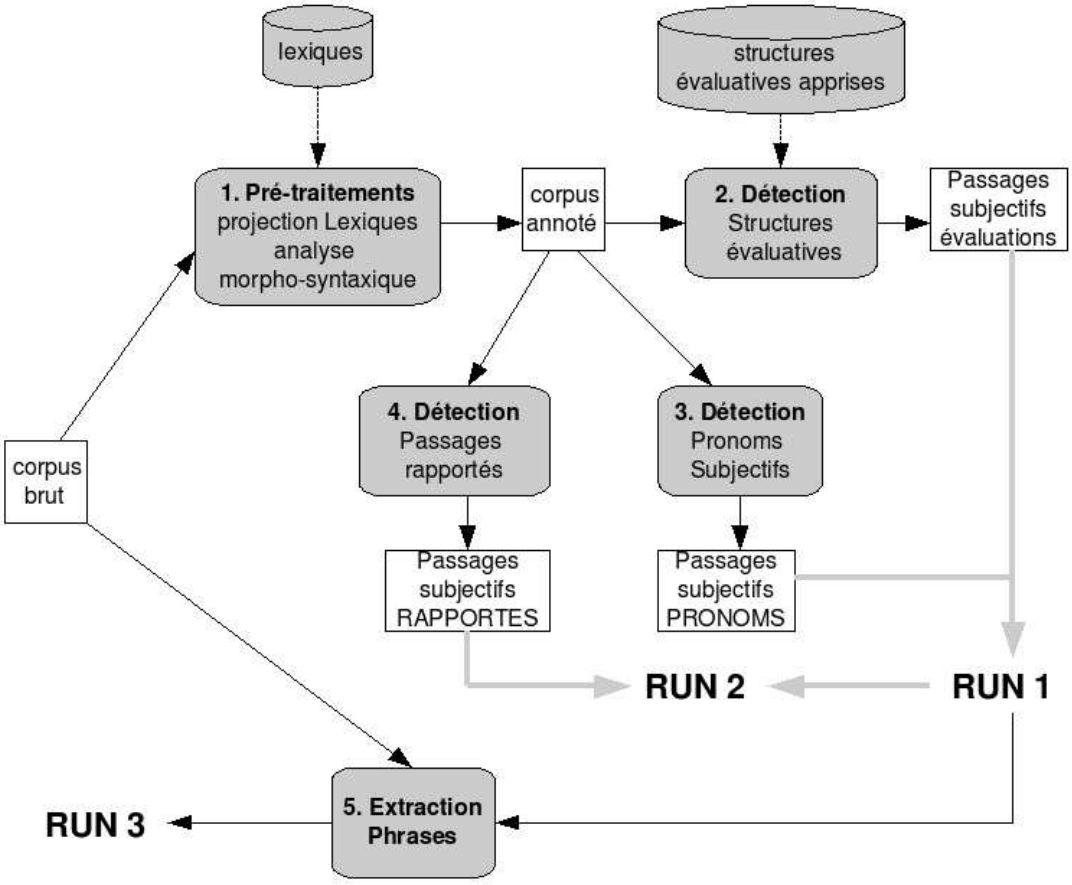

FIG. 3 – Chaîne de traitements pour obtenir les 3 fichiers de run pour la tâche 2

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*.