Identification automatique de marques d'opinion dans des textes

Aiala Rosá [1,2]

[1] Facultad de Ingeniería - Universidad de la República,
J. Herrera y Reissig 565, Montevideo, Uruguay
aialar@fing.edu.uy
[2] MoDyCo, UMR7114, CNRS - Université Paris X
200, avenue de la République, Nanterre, France

Résumé Nous présentons un modèle conceptuel pour la représentation d'opinions, en analysant les éléments qui les composent et quelques propriétés. Ce modèle conceptuel est implémenté et nous en décrivons le jeu d'annotations. Le processus automatique d'annotation de textes en espagnol est effectué par application de règles contextuelles. Un premier sous-ensemble de règles a été écrit pour l'identification de quelques éléments du modèle. Nous analysons les premiers résultats de leur application.

Abstract We present a model for the representation of opinions, by analyzing the elements which compose them and some properties. The model has an operating counterpart, implemented in the form of a set of tags. For the automatic application of these tags on Spanish texts, we work on the writing of contextual rules. A primary subset of rules was written for the identification of some elements of the model. We analyze the first results of their application.

Mots-clés : Identification d'opinions, Fouille de textes, Traitement automatique du langage naturel.

Keywords: Opinions Identification, Text Mining, Natural Language Processing.

1 Introduction

Dans le cadre de la recherche d'information, l'identification d'opinions correspondantes à de différents énonciateurs présente des difficultés du point de vue du traitement automatique. Savoir à qui attribuer les expressions présentes dans un texte, dire si ces opinions sont favorables ou pas envers un certain sujet, cela contribue à l'interprétation du texte et à y trouver des informations. Ce type d'analyse intéresse aussi les systèmes de questions-réponses, par exemple quand il s'agit de répondre à des questions du type (*Qu'elle est l'opinion de la France sur la guerre d'Irak?).*

Le travail que nous présentons a comme objectif l'identification de segments de texte qui contiennent des opinions. Nous visons à repérer des marques formelles qui puissent être reconnues par un système automatique. Nous identifions les éléments qui expriment l'opinion et nous proposons un modèle pour les représenter formellement. Ce modèle conceptuel donne lieu à un modèle opératoire qui servira de base au système d'identification automatique des opinions et des éléments qui les composent. Il s'agit d'un système basé sur des règles symboliques, qui peut être intégré à un outil pour l'application en chaîne de plusieurs modules de traitement automatique de textes.

Ce travail fait partie, d'une part, d'un projet qui vise à la définition d'un modèle du discours, et d'autre part, d'un projet pour le développement d'outils de recherche d'information et de navigation textuelle. Les deux projets sont menés en collaboration entre le laboratoire Modyco (Modèles, Dynamiques et Corpus) à l'Université Paris X, (France), et l'équipe PLN (Traitement du Langage Naturel) à la Universidad de la República, (Uruguay). En ce qui concerne le modèle du discours, nous nous occupons de l'identification des participants du discours et de l'attribution de chaque expression au participant correspondant. Quant à la navigation textuelle, l'identification des opinions permet de définir un parcours sur un texte suivant les opinions d'un certain participant ou selon la polarité envers un sujet déterminé.

2 Travaux reliés

Nous présentons ici quelques travaux sur l'identification d'opinions dans des textes qui sont proches de notre travail. Dans (Wiebe et al., 2005) les auteurs proposent un schéma pour la représentation d'opinions et d'émotions (en anglais, *private states*). Le schéma inclut la source de l'opinion, le segment de texte qui contient l'expression linguistique de l'opinion et quelques propriétés : intensité et polarité. Le schéma est utilisé pour l'annotation manuelle d'un corpus de textes en anglais. En appliquant des techniques d'apprentissage automatique sur ce corpus, les auteurs ont travaillé sur la classification d'opinions et d'émotions selon leur intensité (Wilson et al., 2006) et sur des systèmes de réponse à des questions sur des opinions (Stoyanov et al., 2004).

D'autres travaux, comme (Ku et al., 2005) pour le chinois, qui aborde le problème de la classification de documents selon leur polarité envers un sujet et (Bethard et al., 2005) pour l'anglais, qui s'occupe des opinions exprimées sous forme de proposition introduite par un verbe d'opinion (*dire*, *opiner*), proposent aussi un schéma pour annoter un corpus sur lequel ils appliquent des techniques d'apprentissage automatique.

Notre première approche se fonde sur l'écriture de règles pour l'identification d'opinions dans des textes en espagnol. Nous envisageons une deuxième étape où nous appliquerons des techniques statistiques. Notre modèle pour représenter les opinions ressemble aux schémas proposés dans les travaux mentionnés ci-dessus. Mais à la différence de (Wiebe et al., 2005), qui cherchent à repérer toute expression subjective dans un texte, même celles exprimées par l'auteur du texte, nous cherchons à identifier des segments contenant d'opinions provenant d'autres sources, rapportées par l'auteur du texte, pas nécessairement à connotation subjective. Par rapport au travail de (Bethard et al., 2005), nous repérons les opinions propositionnelles mais aussi des opinions exprimées au moyen d'autres constructions linguistiques, comme des phrases nominales où le nom est dérivé d'un verbe d'opinion (*les opinions de ...*) et des expressions introduites par *selon*, *d'après*, *à l'avis de*, etc.

En ce qui concerne la définition du modèle pour représenter les opinions, nous avons étudié des travaux de sémantique lexicale à partir desquels nous avons analysé les classes de verbes et de noms qui s'emploient normalement dans les expressions d'opinions ainsi que leur structure argumentale. D'une part, FrameNet (Subirats-Rüggeberg, 2003) modèlise cette information, mais la version pour l'espagnol n'est pas encore complète. D'autre part, nous avons consulté le projet ADESSE (García-Miguel, 2005), conçu pour décrire les verbes de l'espagnol : il y a une classification sémantique des verbes et une description des schémas syntaxiques d'apparition de chaque classe de verbes. Même si ADESSE ne s'occupe pas des noms, nous considérons que les noms dérivés des verbes qui nous intéressent héritent d'eux les propriétés sémantiques et syntaxiques.

3 Modèle conceptuel

Nous concevons l'opinion comme un objet conformé par plusieurs éléments, dont le principal est le **prédicat** (verbal ou nominal). Les autres éléments représentent les arguments de la prédication : la **source** (personne, document, publication) à laquelle on peut attribuer l'opinion, le **contenu** de l'opinion et le **sujet** sur lequel porte l'opinion. Nous avons défini deux propriétés pour caractériser l'opinion: la *polarité* et l'*intensité*. La valeur de ces propriétés dépend des éléments qui composent l'opinion, nous avons défini, donc, ces mêmes propriétés pour chaque élément de l'opinion. À partir des valeurs de *polarité* et d'*intensité* des éléments, les valeurs finales de *polarité* et d'*intensité* sont calculées.

L'image ci-dessous montre le modèle établi pour représenter les opinions.

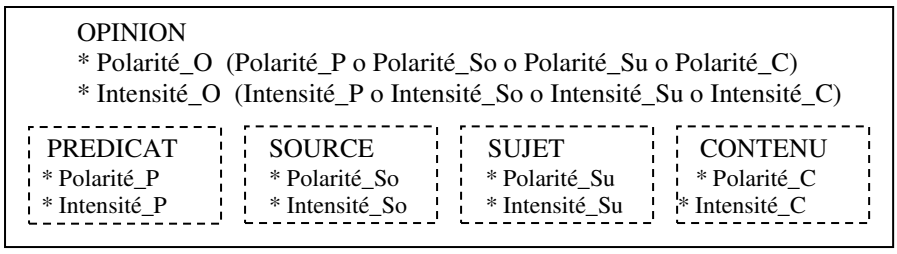

Image 1: Modèle conceptuel pour représenter une opinion

Le modèle proposé s'inspire principalement des exemples observés dans un corpus de textes de la presse uruguayenne[1]. Nous avons pris aussi quelques éléments du schéma proposé par (Wiebe et al., 2005) et nous avons consulté la classification, les schémas et la nomenclature du projet ADESSE. En particulier, nous avons travaillé avec la classe *cognición* (cognition), et ses sous-classes *conocimiento* (connaissance) et *creencia* (croyance), et avec la classe *comunicación* (communication), et ses sous-classes *valoración* (évaluation) et *mandato* (ordre).

La définition des deux propriétés, intensité et polarité, inspirée des travaux de (Mathieu, 2000), permet d'établir des relations entre les opinions d'un texte selon leur différent degré d'intensité (une opinion pourra être plus faible ou plus forte qu'une autre) et leur valeur de polarité (une opinion pourra être contraire ou pas à une autre).

Le tableau 1 illustre les éléments de l'image 1 en montrant quelques exemples extraits du corpus de travail. Dans les exemples, la **source** est marquée en gris, le **prédicat** est en caractères gras, le **sujet** est souligné et le **contenu** est en caractères italiques.

Seul le premier exemple contient tous les éléments qui intègrent l'opinion, pour le reste, le **sujet** et le **contenu** ne sont pas toujours présents. En espagnol, les verbes peuvent avoir un sujet syntaxique omis, comme dans *Declaró también que ...* (*(Il/elle) a déclaré aussi que...*). Dans ces cas il faut rechercher la **source** dans le contexte. Même si le verbe a un pronom personnel comme sujet syntaxique, *Él declaró también que ...*, il faut trouver l'antécédent du pronom pour obtenir la dénomination de la **source**.

Nous considérons tant des prédicats verbaux, exemples 1, 2, 3 et 5, comme des prédicats nominaux, exemple 4. Dans le cas des prédicats nominaux, il y a souvent un verbe support, comme *hacer* (*faire*) dans l'exemple 4, duquel la **source** est le sujet syntaxique. Dans d'autres cas, la **source** apparaît comme complément du nom, introduit par la préposition *de*, par exemple: *las opiniones de la Iglesia Católica ...* (*les opinions de l'Église Catholique ...*).

En ce qui concerne la polarité, l'opinion peut être *neutre*, *positive* ou *négative*, par rapport au sujet sur lequel on opine. Il faut regarder à l'intérieur de tous les éléments présents pour calculer la valeur de polarité de l'opinion toute entière. Le **prédicat** *reafirmaron* (*ont réaffirmé*) a une polarité positive et *rechazaron* (*ont refusé*) a une polarité négative, à partir de ces valeurs, on peut dire que l'opinion est positive pour le premier cas et négative pour le deuxième. Les **prédicats** *respondió* (*a répondu*), *opinión* (*opinion*) et *dijo* (*a dit*) ont une polarité neutre. Pourtant, pour l'opinion de l'exemple 4, le **sujet** introduit par *contraria a* (*contre*) a une polarité négative, l'opinion sera donc négative. D'autre part, pour l'exemple 5, à l'intérieur de l'élément **contenu** on trouve le mot *alentadoras* (*encourageantes*), qui donne au ***contenu*** et à l'opinion complète une polarité positive. On peut penser aussi à des éléments qui peuvent modifier la polarité de la **source**, par exemple, certains adjectifs comme *enthousiasmé* dans la phrase *Enthousiasmé par la proposition, le président a dit que ...*

[1] Le corpus est composé de 75 textes d'environ 500 mots.

	Texte originel	Traduction
Ex. 1	Consultado sobre la lentitud de los procesos judiciales uruguayos Carranza **respondió**: *"Hay una situación de un muy alto número de presos sin condena, hay que agilizar los procesos"*.	Consulté au sujet de la lenteur des processus judiciaires uruguayens Carranza **a répondu** : *"Il y a une situation d'un très grand nombre de prisonniers sans condamnation, il faut accélérer les processus"*.
Ex. 2	Las autoridades uruguayas **reafirmaron** hoy *su compromiso en el cuidado del agua y de su uso como un aporte a las futuras generaciones*.	Les autorités uruguayennes **ont réaffirmé** aujourd'hui *leur compromis dans le soin de l'eau et de son usage comme un apport aux générations futures*.
Ex. 3	Mientras que otros sectores como el Partido Comunista, Asamblea Uruguay, el Partido Socialista y Alianza Progresista **rechazaron** la propuesta.	Alors que d'autres secteurs comme le Parti Communiste, Assemblée Uruguay, le Parti Socialiste et l'Alliance Progressiste **ont refusé** la proposition.
Ex. 4	La Iglesia Católica también hizo pública **su opinión**, contraria a cualquier tipo de despenalización.	L'Église Catholique a aussi fait publique **son opinion**, elle est contre toute dépénalisation.
Ex. 5	El vicepresidente de Instituto Nacional de Carnes de Uruguay, Fernando Pérez Abella, **dijo** *que las conclusiones son muy alentadoras*.	Le vice-président de l'Institut National de Viandes de l'Uruguay, Fernando Pérez Abella, **a dit** *que les conclusions sont très encourageantes*.

Tableau 1: Exemples extraits du corpus

Pour l'intensité, nous avons défini trois degrés : faible, neutre et forte. L'intensité de l'opinion dépend de tous les éléments qui la composent. Les **prédicats** *opiner, dire, répondre* ont une intensité neutre, tandis que *réaffirmer* a une intensité forte. Nous n'avons pas trouvé dans le corpus des **prédicats** d'intensité faible. À l'intérieur de l'élément **contenu** il peut y avoir des éléments qui affectent l'intensité, dans l'exemple 5 on pourrait bien trouver *muy alentadoras* (*très encourageantes*). Dans ce cas, en plus d'une polarité positive, le **contenu,** et aussi l'**opinion**, auraient une intensité forte.

3 Modèle opératoire

Pour le traitement automatique nous avons défini, à partir du modèle conceptuel, un ensemble d'étiquettes pour d'indiquer la présence d'une opinion dans un texte sous la forme d'une annotation de type XML. Les éléments qui composent l'opinion sont

représentés comme des éléments de XML et les propriétés qui caractérisent les éléments sont représentées comme des attributs de XML.

Pour un texte donné, le but du traitement est d'y incorporer les étiquettes pour marquer les opinions et les éléments qui les composent, ainsi que d´établir les valeurs pour les attributs intensité et polarité.

5 Expérimentation

Lors d'une première étape, nous avons travaillé sur l'annotation manuelle d'un sous-ensemble de 7 textes, afin de valider le modèle défini. Sur le corpus entier, nous avons cherché des régularités qui ont permis d'écrire des règles pour l'identification des éléments et l'établissement des valeurs des attributs. Dans cette étape, nous nous sommes consacrés à la tâche de reconnaître le **prédicat** et la **source**.

Pour l'identification des autres éléments et l'établissement de la valeur de polarité et d'intensité il faudra écrire des nouvelles règles. L'incorporation de règles ne présente pas de problèmes, il suffit d'établir un ordre cohérent pour leur application.

Les règles sont écrites suivant le formalisme de règles contextuelles défini dans (Wonsever et Minel, 2001) augmenté grâce à l'incorporation d'un mécanisme qui permet d'imposer des conditions sur les éléments qui composent les règles. Voici la syntaxe des règles contextuelles :

étiquette# contexte gauche >> corps << contexte droit // conditions .

Cette règle établi que si un segment de texte qui s'unifie avec le corps est précédé par un segment de texte qui s'unifie avec le contexte gauche et est suivi d'un segment de texte qui s'unifie avec le contexte droit, alors ce premier segment sera étiqueté par l'étiquette spécifiée avant le signe #. S'il y a des conditions définies après le signe //, il faudra qu'elles soient satisfaites.

Pour écrire et tester les règles, on utilise la plate-forme informatique Lavinia (LAVINIA, 2007) développée par l'équipe PLN. Cet outil contient un module spécifique pour l'application de règles contextuelles qui permet de travailler d'une façon rapide et simple. Ce module nécessite l'application préalable d'un module pour l'étiquetage morpho-syntaxique qui utilise l'étiqueteur FreeLing (FREELING, 2007).

Pour le repérage des prédicats, il convient d'écrire une règle pour les verbes et une règle pour les noms. Chaque règle aura une condition pour exiger que le lemme appartienne à une liste de verbes et noms d'opinion. Ces deux règles ont les contextes gauche et droit vides. Il sera nécessaire d'enrichir ces règles pour résoudre des problèmes d'ambiguïté qui peuvent se poser pour certains prédicats. La désambiguïsation sera traitée par l'analyse des contextes.

Les règles, au nombre de 12, qui identifient la source s'appuient sur la présence d'un prédicat marqué par l'une des règles antérieures. Il faut tenir compte de plusieurs cas :

- Pour un prédicat verbal, la source est une phrase nominale qui apparaît avant le prédicat, *<Cores> dijo a Brecha ...* (*Cores a dit à Brecha ...*), ou bien après le prédicat, *...afirmó a Brecha <el arqueólogo Lenoel Cabrera> (a affirmé à Brecha l'archéologue Leonel Cabrera).*

- Pour un prédicat nominal, la source est une phrase prépositionnelle introduite par la préposition *de,* qui apparaît après le prédicat, *Una de las propuestas <de la OMPI>* (*L'une de propositions de l'OMPI ...*).

- Pour un prédicat verbal où le verbe est au participe passé (voix passive), la source est le complément d'agent, introduit par la préposition *par*, *... propuesto <por Seregni> (proposé par Seregni).*

Une autre règle a été écrite pour les opinions où il n'y a pas un verbe ou un nom comme prédicat, mais une expression comme *selon*, *pour* ou *à l'avis de*.

6 Premiers résultats et travails futurs

Nous avons appliqué les règles pour les prédicats et les sources sur un corpus d'environ 44.000 mots. Nous travaillons actuellement sur l'évaluation des résultats obtenus.

En ce qui concerne l'identification du prédicat, nous avons repéré des occurrences de verbes de notre liste dans des contextes où ils ne sont pas employés pour exprimer des opinions. C'est le cas, par exemple, de verbes qui se trouvent à l'intérieur de locutions comme *es decir* (*c'est à dire*) ou de verbes qui ont plusieurs significations possibles, comme dans l'exemple *La definición 41 dice que ...* (*La définition 41 dit que ...*). Il faudra écrire des règles pour résoudre ces ambiguïtés.

Nous n'avons pas encore travaillé sur l'identification de sources pour les verbes qui ont un sujet omis. Nous ne sommes pas arrivés non plus à identifier les sources des verbes en infinitif, comme dans l'exemple *Los científicos se niegan a especular* (*Les scientifiques ne veulent pas spéculer*) où le verbe d'opinion qu'on marque comme prédicat est *spéculer*.

Pour l'identification du contenu, puisque dans la plupart des opinions cet élément a une forme propositionnelle, on compte sur l'utilisation d'un outil pour la segmentation de phrases en propositions (Caviglia et al., 2006). Il s'agit d'un système basé sur des règles contextuelles qui sera bientôt intégré à la plateforme Lavinia.

En ce qui concerne les propriétés de l'opinion, il faudra travailler sur la classification des verbes selon leur valeur de polarité et leur degré d'intensité et sur le relèvement de mots, principalement des adverbes et des adjectifs, qui peuvent modifier les valeurs de ces propriétés pour chaque élément.

Les règles, écrites pour l'espagnol, pourraient être adaptées à des langues comme le français ou le portugais. Cette possible adaptation sera analysée vers la fin du travail.

Remerciements

Ce travail bénéficie du soutien de ECOS-Sud (U05H01), du PDT (Programa de Desarrollo Tecnológico) du Ministère de Culture de l'Uruguay (2006-2008) et d'une bourse de la Région Ile de France (2008).

Références

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The Berkeley FrameNet Project. *Proceedings of the COLING-ACL.*

BETHARD S., YU H., THORNTON A., HATZIVASSILOGLOU V. & JURAFSKY D. (2005). Automatic Extraction of Opinion Propositions and their Holders. *Computing Attitude and Affect in Text: Theory and Applications*. Editeurs: James G. Shanahan, Yan Qu, Janyce Wiebe, Springer.

CAVIGLIA S., COUTO J., ROSÁ A. & WONSEVER D. (2006). Un sistema para la segmentación en proposiciones de textos en español. *Letras de hoje* 144 (41). 89-101.

FREELING. (2007). http://garraf.epsevg.upc.es/freeling/.

GARCÍA-MIGUEL J. M., COSTAS L. & MARTÍNEZ S. (2005). Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. *Entre semántica léxica, teoría del léxico y sintaxis*. 373-384.

KU L.-W., WU T.-H., LEE L.-Y. & CHEN H.-H. (2005). Construction of an evaluation corpus for opinion extraction. *Proceedings of the 5th NTCIR Workshop Meeting*, 513-520.

LAVINIA.(2007). Accès en ligne. http://www.fing.edu.uy/inco/aplicaciones/Lavinia/.

LAVINIA. (2007). Documentation et fichiers pour l'installation. http://www.fing.edu.uy/inco/grupos/pln/recursos.html.

MATHIEU Y. Y. (2000). Les verbes de sentiment. De l'analyse linguistique au traitement automatique. Paris : CNRS Éditions.

STOYANOV V., CARDIE C., LITMAN D. & WIEBE J. (2004). Evaluating an Opinion Annotation Scheme Using a New Multi-Perspective Question and Answer Corpus. *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

SUBIRATS-RÜGGEBERG C. & PETRUCK M. (2003). Surprise: Spanish FrameNet!, *Proceedings of CIL 17*.

WIEBE J., WILSON T. & CARDIE C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39 (2-3), 165-210.

WILSON T., WIEBE J. & HWA R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence* 22 (2), 73-99.

WONSEVER D., MINEL J.-L. (2004). Contextual Rules for Text Analysis. Lecture Notes in Computer Science.