A Hybrid Method for Sentiment Analysis

Sigrid Maurel, Paolo Curtoni and Luca Dini
(CELI France, SAS, Grenoble, France
http://www.celi-france.com
{maurel, curtoni, dini}@celi-france.com)

Abstract: We present three different methods to perform an automatic classification of French texts which include opinions. The first method is symbolic, the second statistic and the last, hybrid, is a combination of the first two. We will show how the combination facilitates to exploit the advantages of both methods, namely robustness of statistical machine learning and the possibility of a manual configuration which is given by the symbolic method that allows the use of real-life applications.

Key Words: sentiment analysis, opinion mining, text classification, symbolic and statistic methods
Category: H.3, H.4

1 Introduction

This article focuses on the classification of opinion texts in the French language. The target of the classification is the analysis of the feelings and emotions expressed in different types of texts such as discussion forums and newsgroups on the Internet where people exchange views and help each other. The texts are sources of updated and spontaneous information, they are unavoidable to acquire knowledge on consumers. Thus they are necessary to anticipate needs and expectations in order to try to improve the customer/supplier relationship. By analyzing these texts, suppliers of a product or service become more responsive to its customers. Indeed, customers can in turn be guided by the feelings and opinions of other customers on the product to which they are interested and enjoy a decision support (purchase or not the product, choose the product A instead of the product B, etc.). As evidenced by numerous works in socio- and psycho-linguistic (c.f. [Sproull and Kiesler, 1991]), the computer-mediated communication encourages the expression of emotions, feelings and opinions often controlled or repressed in the context of more traditional communication aimed at studying the consumer opinion (face-to-face interviews, closed surveys, investigations, etc.).

As part of its participation in the evaluation campaign DEFT'07 (c.f. section 5 for more details) CELI France has developed three methods to classify the texts of different corpora. The first is a symbolic method, which includes a system for extracting information adapted to the corpora. It is based on the rules of a syntactic and semantic analyzer. The second is a statistical method based

on machine learning techniques. The last, SYBILLE, is a hybrid method that combines the two previous techniques to increase the quality of the results. The sentiments of a document are extracted sentence by sentence, and afterwards an overall value is attributed to the entire message. This processing method makes it possible to extract a context information which is very accurate.

The following sections describe briefly the state of the art and the used corpora. We then focus on the three developed methods and give finally an evalution of their results.

1.1 State of the Art

Today the analysis of opinions focuses on the assignment of a polarity to subjective phrases (words and phrases that express opinions, emotions, feelings, etc.) to decide about the orientation of a document (c.f. [Turney, 2002, Wilson et al., 2004]) or on the positive/negative/neutral polarity of an opinion in a document (c.f. [Hatzivassiloglou and McKeown, 1997, Yu and Hatzivassiloglou, 2003, Kim and Hovy, 2004]). Work going beyond focused on the strength of an opinion of expression where each proposal in a phrase can have a neutral, low, medium or high ground (c.f. [Wilson et al., 2004]). Some grammatical categories were used for the analysis of feelings in [Bethard et al., 2004] where adjective phrases like *too rich* were used to extract opinions conveying feelings. [Bethard et al., 2004] use an assessment based on the sum of scores of adjectives and adverbs sorted manually, while [Chklovski, 2006] uses methods based on a model that reflects the degree of adverbial phrases such as *sometimes*, *much*, *somewhat* or *very strong*.

The approach that we have adopted for the classification of opinion texts is characterized by a mixed use of a symbolic rule-based technology and a statistical technology based on machine learning (c.f. [Dini, 2002, Dini and Mazzini, 2002, Maurel et al., 2007]). The symbolic technology analyzes the text sentence by sentence and extracts the relationships that convey feelings, while the statistical technology processes texts in a single step and assigns a global opinion at the whole text at the end. Note that, unlike other current approaches, the sentiment analysis technology developed by CELI France (SYBILLE) is not limited to a lexical analysis (i.e. identification and pondering of the positive and negative words), but extends to a syntactic and semantic analysis. The parsing is done through a robust surface analysis as described by [Aït-Mokhtar and Chanod, 1997, Basili et al., 1999, Aït-Mokhtar et al., 2001], and gives a result very close to that produced by dependency grammars.

1.2 Corpus

The texts are articulated as a structured flow of interactions, for example: question-answer, argument-opposite argument, comment-disagreement, etc. This

flow is arranged on a time dimension which requires a chronological processing of the thread. Unlike the corpus used by [Wilson et al., 2004] it is not necessary to identify the person to which a sentiment is associated, because in 95 % of the cases the analysed messages are in the first person. The corpora used are from two distinct sources and differ quite a lot in total size and typical size of the threads. We used the corpora of DEFT'07, to which we have added some own corpora collected on the internet. The texts of our corpora focus on tourism (in France and elsewhere in the world) and video games (criticisms and problems). They include on the one hand troubleshootings, but also advice on products and places to visit.

2 Symbolic Method

The symbolic method is based on a parsing of the text made by a functional and relational analyzer (c.f. the work on syntactic and semantic analysis of [Basili et al., 1999, Aït-Mokhtar et al., 2001, Dini, 2002, Dini and Mazzini, 2002, Dini and Segond, 2007]). This analyser processes texts sentence by sentence, and extracts for each sentence the syntactic relations. There are basic syntactic functional relations, such as modifier of a noun, of a verb, subject and object of a sentence, as well as more complex relations, such as coreference between two phrases within the same sentence. The user has the possibility to develop a grammar for a specific application and to add new rules to extract the relationships which he is interested in. Therefore he can change the extraction-rules (e.g. add rules for a new relationship), increase/decrease features on the words in the lexicon that act on the rules, remove certain parts of the processing, etc.

2.1 Grammar

The grammar was originally developed to extract sentiment relations as part of a project on tourism in France. It was then modified and improved for the participation in DEFT'07 (c.f. section 5, [Maurel et al., 2007]). The grammar is constituted by two parts: a first basic part (the "generic" grammar) applies to all the texts that contain feelings, and a second part for each different domain, depending on the subject of the corpus: tourism, video games, etc. The differences lie mainly in the implemented lexicons, as each domain has its own words and expressions. For instance the words related to speed (*lent*, *rapide*, etc.) have different polarities according to if they describe a printer or a holiday trip. Similarly, as shown in the sentences below, the adjectif *effrayant* is rather seen as positive in a novelistic description while it is perceived as negative in the insurance and tourism domain: *"Dans* Ghost*, les habitants du village sont vraiment effrayants!"* (*In* Ghost *the people of the village are really frightening!*)

or *"C'est effrayant de voir comment la côte est de plus en plus bétonnée."* (*It's frightening to see how much the coast is spoiled.*).

In general, a sentiment relation has two arguments: the first is the linguistic expression that conveys the feeling in question, the second is the cause or the object of the sentiment (if the cause is expressed in the sentence). This provides for the sentence *"J'aime beaucoup Grenoble."* (*I like Grenoble very much.*) the relation `SENTIMENT_POSITIF(aimer,Grenoble)`. The feature `POSITIF` of the relation, i.e. the value of it's class, indicates that it is a positive feeling and the cause is *Grenoble*. In a sentence like *"Je déteste!!"* (*I hate!!*) the relation will have only one argument: `SENTIMENT_NEGATIF(détester)`, insofar as the subject matter of sentiment is not expressed in the sentence. The goal of the grammar is to extract a maximum of information from the thread, in particular the positive and negative feelings, places and products. For this threads are analyzed sentence by sentence. Each sentence can have zero, one or more sentiment relations. It is possible to have relations of positive and negative feelings in a single sentence: *"En qualité d'impression, la Epson est meilleure, en texte comme en photo, malheureusement c'est aussi la plus chère."* (*In printing quality the Epson is better, in text and in photo, unfortunately it's also the most expensive.*).

2.2 Sentiment Lexicons

The analysis of the text is based on the words in the lexicon that have received specific features marking the positive or negative feelings. Most of the words are verbs (*aimer*, *apprécier*, *détester*, ...) and adjectives (*magnifique*, *superbe*, *insupportable*, ...), but also some common nouns (*plaisir*) and adverbs (*malheureusement*). For example, when a modifier of the noun is extracted (*paysage magnifique*) and the modifier (*magnifique*) bears the feature `sents`, the relation of sentiment (⇒ `SENTIMENT_POSITIF(magnifique,paysage)`) is then extracted between the noun and its modifier. There are of course more complex rules to extract relations of more complicated sentences.

The attribute of the relation (`positif` or `negatif`) of a sentiment will be reversed when a denial is present in the sentence. Where possible, the pronouns *qui* and *que* referring to an entity present elsewhere in the same sentence, will be replaced by this entity. Some interrogative common nouns and verbs have received the feature (`no-sents`) to prevent the extraction of relations. There are no real sentiments in *"Je cherche un bon hôtel."*, *"Bon voyage!"* ou *"Bonne journée!"* expressed by the author of the text, but rather wishes as they can be found primarily at the beginning or the end of messages. The size of the lexicon varies depending on the domain of application. The basic sentiment grammar contains about 250 words (nouns, verbs, adjectives, etc.) with the sentiment features (`positif` and `negatif`). We added approximately 150 words in the

domain of tourism at this basic lexicon, and about 250 words in the domain of video games.

2.3 Manual Annotation of Texts

The configuration of the generic grammar was made on the basis of a manual annotation work (using Protégé 3.2[1] software with the plugin Knowtator[2]) of forums from the domain of tourism. This corpus contains a hundred annotated threads. Each thread consists of messages from the forum, the length varies between ten and 55 posts per document, a message can contain only a sentence or several paragraphs. The annotation of this corpus is made according to the works of [Riloff et al., 2005, Riloff et al., 2006, Wiebe and Mihalcea, 2006].

The annotation includes informations of cause, intensity and the subject expressing the sentiment. In *"J'aime énormément Grenoble."* (*I love Grenoble very much.*) *aimer* conveys the feeling, *Grenoble* is the cause of the feeling, and *je* is the issuer of the sentiment. The adverb *énormément* expresses the intensity, the feeling here is more intense than in the sentence *"J'aime bien Grenoble."* (*I like Grenoble.*). The annotation for the *tourism* corpus does not only include the values *positive* and *negative* to classify the feelings, but is much more detailed (c.f. for example the work of [Mathieu, 2000, Mathieu, 2006]). The chosen annotation scheme is even finer and therefore the classification of feelings we propose allows a large number of modalities and goes beyond the simple positive-negative opposition. Indeed, we have taken the taxonomy of [Ogorek, 2005] which offers 33 different feelings (17 positive and 16 negative) to which we added the pseudo-sentiments as *bon-marché*, *conseil*, *cher* and *avertissement*, because in the field of tourism there are a lot of messages about the prices of services whose authors are happy (or not). The sentiments of Ogorek's taxonomy are ordered in groups as AMOUR-DÉSIR (*amour, envie, tendresse, désir*), JOIE (*enchanté, excité, heureux, joyeux*), TRISTESSE-DÉTRESSE (*découragé, bouleversé, démoralisé, triste*), COLÈRE-DÉGOÛT-MÉPRIS (*colère, mépris, désapprobation*), etc.

3 Statistical Method

We use a machine learning technique for the statistical method which is based on the work of [Pang et al., 2002, Pang and Lee, 2004, Pang and Lee, 2005]. We have adapted it to the French corpora and tested it on threads of the *tourism* corpus and on the texts of the corpora of DEFT'07. [Pang and Lee, 2004] propose two possible classifications, namely subjective/objective sentences and among the subjective sentences opinions in the opposition positive-negative. The method

[1] `http://protege.stanford.edu/`
[2] `http://bionlp.sourceforge.net/Knowtator/index.shtml`

is based on the removal of all objective phrases of the text and the classification is done only on the subjective part. This extract corresponds in their experiments to 60 % of the original text. We didn't adopt this approach because of the lack of a training corpus with distinct subjective and objective parts. We chose to separate subjective-objective text by using the symbolic method which allows us to get more nuanced results. The extracts can be seen as good summaries of the text on the feelings they express.

The statistical method is based on n-grams of characters. For French we chose $n = 12$. As for the symbolic method, a confidence index is attributed to the texts. It allows us to compare the result with the symbolic method to conclude the final outcome with the hybrid method. For the training of the texts, technologies like *support vector machines* (SVM) and *naive bayes* (NB) were used. The results are slightly better with NB, but this remains negligible. Experiments were made with a corpus of DEFT'07 which contains criticisms of books and films, taking only the first and/or the last sentence(s) of the message. This is based on the assumption that the judgement of the author in a book review or film can be found most often at the beginning or the end of the message, the middle being probably occupied by the summary of the book or film. The results of classification in positive or negative with this technique are better than by taking the message in its entirety. However, this technique was not adopted because it would not be repeatable on messages from areas other than movie critics and book, where there is not necessarily a summary in the middle of the message. The training of the statistic module is done only on the sentences of each thread, which were selected by the symbolic method containing sentiments.

4 Sybille, the Hybrid Method

The hybrid method is a combination of the two previous methods. It takes as input the output of the other two methods and calculates according to the confidence indices for each result, an average which will be translated into positive or negative. The exact weighting varies with several factors, including the accuracy of the manual annotation and the size of the corpus. The statistical method allows to make an initial mining in the texts to get positive and negative messages. Then the user who set up the grammar can modify and improve it for better results. The work takes the form of a cycle where results are constantly improving.

The analysis of the thread is done at the level of sentences and allows to improve the result by adding or removing words in the lexicons. This has the advantage of showing exactly what sentences of the document express a feeling. It is an approach that keeps the robustness of the machine learning of the statistical method and orients at the same time the base of the training on a manual

configuration of the symbolic method. This helps to correct significantly errors of machine learning and to integrate the specific project specifications, i.e. the particularities of each corpus (using glossaries and lexicons of different terms depending on the domain of application).

5 Evaluation

DEFT'07[3] (*le Défi Fouille de Texte*) was an evaluation campaign dealing with classification of opinion texts. Several research groups (university laboratories or private companies) participated to test and compare their classification systems on several corpora. In the initial phase each group received two thirds of each of the four different corpora which had as subjects criticism of films and books, tests of video games, proofreading of scientific articles and memos of parliamentary debates. For the first three corpora a note with three values (positive, neutral or negative) has been manually assigned to each text by the committee of the organizers, whereas a note with only two values (positive or negative) for the last corpus. After a certain period of time in which each group has developed their classification systems the third thirds of each corpus has been sent for tests. Each group submitted their results only a few days later. Ten teams participated in the 2007 edition, CELI France came in third place, the first two teams are from academic research labs.

The grammar of the parser has been set to meet the needs of the different DEFT'07 corpora, from the point of a lexical view but also to resist common spelling errors. The most important thing was to change the classification of the entire message, which may contain several sentiments with a single overall value, and in particular the introduction of the concept of an neutral sentiment. Our standard approach is that on the sentence-level feelings are positive or negative. It is not necessary to use neutral sentiments, insofar as the taxonomy used (c.f. section 2.3) allows to nuance enough.

6 Conclusion

In this article we presented how the automatic processing of natural language can improve the quality of a system for extracting sentiments. We have described a method with a symbolic grammar adapted to the text domain, and a method of machine learning. The evaluation of our classification system SYBILLE has shown that the combination of a statistical and a symbolic methods gives more accurate results than either method used separately. To go further we will improve the system and are now in the phase of extending it to deal with other domains such as hardware, environmental policies and fashion.

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

References

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

[Aït-Mokhtar et al., 2001] Aït-Mokhtar S., Chanod J.-P. and Roux C. "A multi-input dependency parser"; In *Proc. of IWPT'01*.

[Basili et al., 1999] Basili R., Pazienza M. T. and Zanzotto F. M. "Lexicalizing a shallow parser"; In *Proc. of TALN'99*.

[Bethard et al., 2004] Bethard S., Yu H., Thornton A., Hatzivassiloglou V. and Jurafsky D. "Automatic extraction of opinion propositions and their holders"; In *Proc. of AAAI'04*.

[Chklovski, 2006] Chklovski T. "Deriving quantitative overviews of free text assessments on the web"; In *Proc. of IUI'06*, 155–162.

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

[Dini and Mazzini, 2002] Dini L. and Mazzini G. "Opinion classification through information extraction"; In Zanasi A., Brebbia C. A., Ebecken N. F. F. and Melli P., Eds., *Data Mining III*, 299–310. WIT Press.

[Dini and Segond, 2007] Dini L. and Segond F. "La linguistique informatique au service des sentiments"; In *Revue de l'électricité et de l'électronique*, 66–77. Editions SEE.

[Hatzivassiloglou and McKeown, 1997] Hatzivassiloglou V. and McKeown K. R. "Predicting the semantic orientation of adjectives"; In *Proc. of ACL'97*, 174–181.

[Kim and Hovy, 2004] Kim S.-M. and Hovy E. "Determining the sentiment of opinions"; In *Proc. of COLING'04*, 1267–1373.

[Mathieu, 2000] Mathieu Y. Y. "Les verbes de sentiment. De l'analyse linguistique au traitement automatique"; CNRS Editions.

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

[Maurel et al., 2007] Maurel S., Curtoni P. and Dini L. "Classification d'opinions par méthodes symbolique, statistique et hybride"; In *Proc. of DEFT'07*, 111–117.

[Ogorek, 2005] Ogorek J. R. "Normative picture categorization: Defining affective space in response to pictorial stimuli"; In *Proc. of REU'05*.

[Pang and Lee, 2004] Pang B. and Lee L. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts"; In *Proc. of ACL'04*, 271–278.

[Pang and Lee, 2005] Pang B. and Lee L. "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales"; In *Proc. of ACL'05*, 115–124.

[Pang et al., 2002] Pang B., Lee L. and Vaithyanathan S. "Thumbs up? Sentiment classification using machine learning techniques"; In *Proc. of EMNLP'02*, 79–86.

[Riloff et al., 2006] Riloff E., Patwardhan S. and Wiebe J. "Feature subsumption for opinion analysis"; In *Proc. of EMNLP'06*, 440–448.

[Riloff et al., 2005] Riloff E., Wiebe J. and Phillips W. "Exploiting subjectivity classification to improve information extraction"; In *Proc. of AAAI'05*.

[Sproull and Kiesler, 1991] Sproull L. and Kiesler S. "Connections: New ways of working in the networked organization"; Cambridge: MIT Press.

[Turney, 2002] Turney P. D. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews"; In *Proc. of ACL'02*.

[Wiebe and Mihalcea, 2006] Wiebe J. and Mihalcea R. "Word sense and subjectivity"; In *Proc. of ACL'06*, 1065–1072.

[Wilson et al., 2004] Wilson T., Wiebe J. and Hwa R. "Just how mad are you? Finding strong and weak opinion clauses"; In *Proc. of AAAI'04*.

[Yu and Hatzivassiloglou, 2003] Yu H. and Hatzivassiloglou V. "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences"; In *Proc. of EMNLP'03*, 129–136.