Opinion Detection as a Topic Classification Problem

Juan-Manuel Torres-Moreno
Université d´Avignon et des Pays du Vaucluse
245 PUBLICATIONS **1,812** CITATIONS

Marc El-Bèze
Université d´Avignon et des Pays du Vaucluse
124 PUBLICATIONS **907** CITATIONS

Patrice Bellot
Aix-Marseille Université
234 PUBLICATIONS **1,704** CITATIONS

Chapter 9

Opinion Detection as a Topic Classification Problem

9.1. Introduction

In recent years the *classification of documents according to their opinion*[1] considered as a sub-task of document classification, has attracted a steadily growing interest from the Natural Language Processing (NLP) community. Various problems are resolved in document classification, including those which consist of determining the thematic of a document among a finite set of possible thematics. For example, in a corpus of journalistic documents, the task consists of classifying the thematics of texts as *politics, society, sports, arts*, and so on. The objective of opinion detection is to find out, for example, whether a positive or negative opinion is expressed in a text on a certain subject. From this perspective, positive and negative opinions can be considered as two classes which have to be attributed in the framework of the classical classification task. *A priori*, detection and classification of opinions might appear to be a simple task. For numerous reasons, the problem turns out to be rather complex and difficult to solve. An aggravating factor is that often only corpora of limited size and with an asymmetrical distribution of their classes are available.

However, the highly subjective nature of the documents (which may be, among other things, texts associated with products, music criticisms, cinema, political interventions, blogs, or discussion forums) adds to the difficulty of the task. This

Chapter written by Juan-Manuel TORRES-MORENO, Marc EL-BÈZE, Patrice BELLOT, and Fréderic BÉCHET.

[1] Also known as sentiment classification, sentiment analysis, or opinion mining.

particular feature calls for solutions other than those currently used in a classical classification task [PAN 02].

In a classification based on opinion, complex language phenomena come into play. They can be found at all levels: lexical, syntactic and semantic. But they are also pragmatic when real-world knowledge is required. For example, to determine the polarity of the phrase: *The author offers, in this book, another of the gems which people have got used to*, it is essential to know the author. What other books has he written? Or even: are his books any good? With this previous knowledge (which may or may not be present in the same document in which the opinion is expressed) the extent of the underlying semantic field should be expanded, so that we can increase the chance of determining whether the opinion expressed is positive or if a lot of irony masks a negative opinion.

Studies on opinion classification in scientific literature deal mostly with English corpora. We will also discuss studies based on French corpora realized by [TOR 07, TOR 09]. These corpora concern films, books, video game reviews, political debates, and scientific articles. For obvious reasons, we will focus here on approaches attempting to remain language-independent as far as possible.

Classifying a corpus to a predetermined set of classes and its corollary, profiling texts, is a significant problem in the domain of text search. The aim of classification is to automatically assign a class to a given text object as a function of a profile which will or will not be defined depending on the classification method. A wide range of applications exists. They go from the filtering of large corpora (in order to facilitate information retrieval or scientific and economic monitoring) to the classification by the genre of the text to adapt the linguistic processes to the feature of the corpus. The assignment of a class to a text also involves the assignment of a value which can serve as a criterion in a decision process. The classification of a text according to the opinion expressed is a relevant practical problem, notably in market studies. Nowadays there is a strong demand by some companies for the capability to automatically assess whether the company image in a press article is positive or negative. Hundreds of products are evaluated on the Internet by professionals or non-professional users on dedicated sites: what conclusive judgment can we derive from this collection of information provided by the consumer or the company which makes this product? Besides marketing another possible application concerns the articles of a collaborative encyclopedia on the Internet such as Wikipedia: does an article convey a favorable or an unfavorable judgement or is it rather neutral, in line with a fundamental principle of this free encyclopedia?

The core concepts that will be used are based on the adequate representation of the document and the numerical and probabilistic methods for their classification. The semantic orientation of an opinion (expressed as a word or a set of words, generally known as a term) will be approached by numerical values: either by using the *n*-grams occurrences, or by conditional probabilities, or a combination of the two. A value exceeding a empiric threshold can be considered as an indication that it belongs to a positive class, or a negative one otherwise. Intermediate values indicate the degree of these implications.

The first attempts to automatically classify adjectives according to their semantic content were realized by [HAT 97] using conjunctions between adjectives. [MAA 04] have used the semantic distance of WordNet [FEL 98] between a word to classify the *good* and *bad* terms. [TUR 03] describe an unsupervised classifier based on the content of an opinion. The classifier decides whether the type of a document is positive or negative based on the semantic orientation of the terms the document contains. This orientation is calculated by estimating the *Pointwise Mutual Information* between a term in question and a pair of primer words that are unambiguous representations of the positive or negative orientations, through a search on Web pages in order to estimate their values. This system correctly ranks 84% of the documents of a small corpus related to cars. The same method when applied to a test corpus of cinema reviews obtains a precision of 65%, which is a particularly difficult task.

In [TOR 07] the authors describe a study that relies on corpora which contain not just two polarities but three: positive, negative, and neutral. This complicates the task in two ways. Firstly, the distribution of class sizes is highly skewed. Secondly, the typical size of the corpus available is rather limited and the characteristics of the polarities are not evident. In an original manner, [TAK 05] have employed the spin *up/down* models to characterize the polarity of the words. They use up/down systems to find the semantic orientations of words: either positive or negative (desirable or undesirable) according to the primer words. The output is a list of words indicating their orientation estimated according to the average field approximation. The authors point out that their approach is equivalent to that of entropy maximization. Even though these results are average in quality, we cite them to show the diversity of methods which can be applied in order to resolve this difficult problem.

9.2. The TREC and TAC evaluation campaigns

The NIST[2] is at the origin of several evaluation campaigns of natural language processing systems working on opinions. From 2008 the DUC conferences (*D*ocument Understanding Conference) were continued as evaluation tracks TAC (*Text Analysis Conference*)[3] of NIST, i.e. machine research of texts relying on (TREC) opinions, precise questions (TAC, QA), opinion summaries (TAC *Summarization*). These evaluations are performed on the Blog06 corpus [MAC 06] composed of 3 million texts issued from 1,00,000 blogs. Expressed in the form of a set of questions relating to people, organizations, varied topics or objects, the goal is to detect all the opinions expressed in the blogs and provide the user with the most exhaustive view possible.

[2] National institute of standards and technology, http://www.nist.gov

[3] Text analysis conference, http://www.nist.gov/tac/

9.2.1. *Opinion detecttion by question–answering*

The objective of the *opinion QA track* of TAC is to precisely respond to questions that relate to an opinion expressed in the documents. For example, "*Who likes Trader Joe's?*" or "*Why do people like Trader Joe's?*" are two questions referring to the grocery shop chain "*Trader Joe's*" and aim at well-defined named entities (*rigid list*) as resonse or at difficult explicative responses (*squishy list*). In the latter case, the responses must contain the *nuggets* of information. Those are subdivided into essential information (*vital*) and exact but not essential information. The chains of responses generated by the systems can be in the following form: "*Trader Joes is your destination if you prefer Industrial wines (unlike Whole Foods)*"; "*Sure, we have our natural food stores, but they are expensive and don't have the variety that Trader Joe's has*".

In 2008, nine teams have participated in the TAC evaluation campaign. For the 90 series of *rigid* questions, the best system has obtained an average F-score of 0.156, where the F-score is a combination of ($\beta = 1$) precision (number of correct responses divided by the number of responses supplied) and recall (number of correct responses divided by the number of known correct responses). In comparison, the manual reference score was above 0.559. The scores of the other eight teams in 2008 ranged between 0.131 and 0.011.

For the series of *squishy list* questions, the evaluation uses the pyramidal method [NEN 04]. It starts from a list of *nuggets* defined by a first assessor and later enriched by nine other judges from the responses provided by the participants. If, for instance, the average F-score ($\beta = 3$) of the assessors is in the order of 0.5, then those from the systems in 2008 should be 0.17.

The best scores were obtained by the University of THU Tsinghua (China) [LI 08] and IIIT Hyderabad (India) [VAR 08]. The first, THU QUANTA is based on the QUANTA questions–responses system enriched by the use of a lexicon expressing *sentiments*. The authors have tried several lexica, such as Hownet, Wordnet, or Mpqa but without much success as these do not establish the polarity of a word in context (e.g. the adjective *big* can be positive or negative). They have preferred to use their own lexicon, with a more limited size but without having the same inconvenience. The rigid questions are subdivided into two classes, each adopting a specific strategy for their resolution. Firstly, there are the questions asking for the names of blogs or the authors of messages and secondly there are the more classical (factual) questions requiring the precise named entities to be found. The questions are additionally labelled with the sentiment they express (positive or negative). A similar labeling is applied to the documents found in the documentary research phase using the Lucene engine[4]. Lucene has two independent sets to process, i.e. the set of positive and the set of negative documents.

[4] Lucene is available at http://lucene.apache.org/

To extract the nuggets (*squishy* question) only documents that have a polarity similar to those of the questions are considered [LI 08]. The question analysis phase uses different NLP modules which perform morpho-syntactic and syntactic analyses, the resolution of anaphora, and the recognition of named entities. Opinion detection is effected by a labeler based on the resource of relationships PropBank[5]. PropBank together with the use of predefined patterns (e.g. "*reason for XXX*" or "*opinion about XXX*") associates the detected predictive forms with an opinion verb, opinion carrier, and opinion object. Labeling with a positive or negative sentiment is in turn realized by using an ad-hoc lexicon of positive or negative "opinion words". Depending on the type of question, responses are extracted as a function of analyses based, among other things, on the occurrence frequency of words from the query in the selected phrases and in the document title as well as the number of carrier words of an opinion either by using a BM25 similarity associated with a density measure of the words of the question in the passages of the candidate phrases (traditional approach in question–responses). The extraction of nuggets is in turn realized by combining a pattern-based approach (for the *why* questions, for example) or external knowledge (list of actors and films extracted from the IMDB[6], for example).

The second system that obtained the best performance in 2008 essentially employs numeric approaches [VAR 08]. The architecture of the question–response system remains classical: question categorization, document and passage search (from Lucene and a Bayesian ranker), extraction of candidate responses, and possible candidates ranking. The categorization of questions enables us to determine the type of response, so that we can look for the polarity of the question. Support vector machines (SVM) and Bayesian rankers are used to this effect. The documents found by Lucene are classified into two categories based on their polarity and matched corresponding to that of the question. Responses to *rigid* questions are extracted following a recognition of Named Entities by the SNER software (*Stanford Named Entity Recognition*).

To extract *squishy* answers, the authors have chosen approaches close to those which were previously used for the DUC evaluation of 2007: selection of the more informative phrase according to a combination of divergence scores between the phrases (Kullback-Leibler distance) and similarity to the question. In order to determine the polarity of the questions, the authors have used lists of positive and negative words and have established classification rules. In order to calculate the document polarity, they have decided to create two Bayesian rankers: the first to distinguish the opinion carrier phrases (no matter the polarity of opinions) and the second is for differentiating positive or negative opinions.

[5] http://verbs.colorado.edu/~mpalmer/projects/ace.html of the predicate-argument type. PropBank can be downloaded here (January 2012): http://verbs.colorado.edu/verb-index/pb/propbank-1.7.tar.gz.

[6] Internet movie database.

9.2.2. *Automatic summarization of opinions*

As a continuation of the task described in the previous section, NIST introduced a pilot evaluation titled *Opinion Summarization* in the domain of automatic summarization of TAC evaluation campaigns in 2008. The objective is to generate texts with a maximum length of 7,000 or 14,000 characters. They should summarize the opinions about certain precise topics found in the blogs. The topics are expressed in the form of a set of questions in natural language. For example, on the *Gregory Peck* topic, the questions in 2008 were as follows:

– *What reasons did people give for liking Gregory Peck's movies?*

– *What reasons did people give for not liking Gregory Peck's movies?*

Some of the extracts from the blogs which carry the opinions responding to questions asked are as follows:

– *I've always been a big Peck fan since I saw him in* To Kill a Mockingbird. *A* Charmed Life *is a fine biography of one my favorite actors.*

– *Gregory Peck can be seen playing his first character without absolutely any redeeming quality in this thriller* (Boys from Brazil).

– *After half an hour of the slow-paced antics of* Roman Holiday, *I voted to stop watching it, so I yoinked it from the DVD player, sealed up the disc in its return envelope*.

22 sets of questions (*topics*) are evaluated as a function of the readability of the summaries[7] (grammar, non-redundancy, structure, and coherence) as well as their informative richness by using the pyramidal method[8].

19 teams participated in this evaluation campaign in 2008. Among all systems, our systems [BOU 08] have pursued an automatic summarization approach by extraction, for certain ones, with a compression phase of phrases and post-processing to improve the readability of the final summary [DAN 08]. Some of the teams have used external resources such as WordNet[9] or Wikipedia[10]. As is often the case, numerical approaches were combined with symbolic approaches. [GEN 08] describes the system that has obtained the best score with respect to several criteria by employing deep syntactic analysis. The IIIT, Hyderabad [VAR 08] system applies methods similar to those described in the previous section for the *opinion QA track*. The main differences are at the ranking level so that we can determine the polarity of the phrases (SVM rather

[7] The nature of the corpus, text coming from blogs, has naturally increased the difficulty of obtaining highly readable texts.

[8] See http://www1.cs.columbia.edu/˜becky/DUC2006/2006-pyramid-guidelines.html.

[9] http://wordnet.princeton.edu.

[10] http://www.wikipedia.org.

than Bayesian), and in the new usage of the SentiWordNet lexicon[11] which assigns a polarity to each *synset* of WordNet (positive, negative, or neutral).

9.2.3. *The text mining challenge of opinion classification (DEFT (DÉfi Fouille de Textes))*

In July 2007, the third edition of the text mining challenge DEFT (DÉfi Fouille de Textes) was organized [AZÉ 05, AZÉ 06]. The text mining challenge in 2007[12] was motivated by the need to put in place text searching techniques enabling us to classify texts according to the opinions they express. Specifically, it consisted of classifying the texts of four corpora in French language according to the opinions formulated. The task proposed by DEFT'07 falls into the application domain of decision making. The data made available to the participants of the challenge consisted of four heterogeneous corpora:

to see/read: 3,460 film reviews, books, shows and comic books, and associated grades. The latter is due to the fact that many film or book review organizations[13] in addition to the commentary assign a grade in the form of an icon. A three level grading system was employed which resulted in three distinct classes: 0 (bad), 1 (medium), and 2 (good);

video games: 4,231 video game reviews with an analysis of different aspects of the game (graphics, playability, length, sound, scenario, etc.) and an overall summary of the game. As in the previous corpus, a three level grading system with the classes 0 (bad), 1 (medium), and 2 (good);

proofreading: reiews of 1,484 scientific articles which aid the decision making of the program committees of scientific conferences and give advice and make recommendations to the authors. Again a three level grading system is applied. Class 0 is assigned to proofreadings that recommend rejection of the article. Class 1 contains the proofreadings that recommend acceptance either demanding major modifications or the referral of the paper to a poster session. Class 2 comprises all articles accepted without or with minor modifications. In this corpus (like the following one) the names of persons have already been anonymized;

debates: 28,832 statements of Members of Parliament on law projects discussed in the national assembly. Each statement is accompanied by the Members' vote: 0 (in favor) or 1 (against).

[11] http://sentiwordnet.isti.cnr.it.

[12] http://deft07.limsi.fr/.

[13] For example see site http://www.avoir-alire.com.

The organizers have split the corpora into two parts: one part (approximately 60%) of the data has been given to the participants as learning data in order to develop their methods, the other part (approximately 40%) has been reserved for the test phase. The use of any data beyond those provided by the organizing committee was forbidden. This specifically excluded access to Websites or whatever other source of information.

The challenge aims at classifying each text according to the opinion expressed, i.e. positive, negative, or neutral in the case of three classes, and for or against in the binary case (parliamentary debates corpus). To test our methods and fine tune their parameters, the learning set of each corpus has been split into five subsets of approximatively the same size (number of texts to process).

9.2.3.1. *Integration of systems*

Nine decision systems (to be presented later) using classifiers with different text representations (n-grams, n-lemmas, *seeds*, and terms) have been combined with the objective to obtain *different recommendations* on the label of a text. Furthermore, the goal is not to optimize the stand-alone results of each classifier but to use them as tools in their default setting and to approach the optimum by the combination of their results.

1) LIA_SCT [BÉC 00] is a classifier based on semantic decision trees *SCT-Semantic Classifin Tree* [KUH 95]). The texts are represented as lemmas. BoosTexter [SCH 00] is a wide margin classifier based on the *boosting Adaboost* algorithm [FRE 96]. Four of our systems use the *BoosTexter* classifier:

2) LIA_BOOST_BASELINE: a document is representated by word trigrams;

3) LIA_BOOST_BASESEED: each document is represented by *seeds* weighted by the number of its occurrences, in uni-gram mode;

4) LIA_BOOST_SEED: each document is represented by the words and also by *seeds*;

5) LIA_BOOST_CHUNK: The *LIA-TAGG* tools cuts the document into a set of lemmatized syntagms. Each syntagm contains a *seed* and the previous and following syntagms are also chosen as the representation. The other syntagms are not part of the representation of the document. *BoosTexter* is applied in the trigram mode on this representation.

6) SVMTorch [COL 02] is a classifier based on the Support Vector Machines (SVM) [VAP 82, VAP 95]. The LIA_NATH_TORCH system is obtained with *SVMTorch* and the input vector is represented by the *seeds* lexicon.

7) Timble [DAE 04] is a classifier that implements several *Memory-Based Learning* (MBL) techniques. The LIA_TIMBLE system applies the *TiMBL* tool to the *seeds*.

8) Probabilistic modeling of a uni-lemma and family of words. The texts have been slightly filtered (in order to keep little intonations such as a passive voice, interrogation,

or exclamation forms), an aggregation process of composed words, and regrouped in the words of the same family (by using a dictionary with approximately 3,00,000 forms). Each document has been transformed into a bag of uni-lemmas. Then, the class to which document t belongs is calculated as:

$$P_t(w) \approx \prod_i \lambda_1 P_t(w_i) + \lambda_0 U_0 \qquad [9.1]$$

where $P(w)$ is the probability of belonging to word w, U_0 is a constant ...

9) Modeling according information theory. Here, we envisage going back to a classical information theory model, all the while looking to integrate some of the specificities of the problem. The formulation initially chosen is close to those employed during a previous DEFT [ELB 05]:

$$\widetilde{t} = \arg\max_t P(t) \times P(w|t) = \arg\max_t P(t) \times P_t(w) \qquad [9.2]$$

The labeling t can be based on values from a reduced cardinal set of two or three elements [0-1] or [0-2], *a priori* the problem seems to be simple, and the quantity of the data supplied is sufficient to learn the models well. Even if the specific vocabularies of different corpora are not large (between 9,000 different words for the smallest corpus and 50,000 for the largest), it remains that certain entries are under-represented. Also in line with what is normally done to calculate the value of the second term of equation [9.2] an n-lemma smoothing model has been opted for (n starting from 0 to 3):

$$P_t(w) \approx \prod_i \lambda_3 P_t(w_i|w_{i-2}w_{i-1}) + \lambda_2 P_t(w_i|w_{i-1}) + \lambda_1 P_t(w_i) + \lambda_0 U_0 \qquad [9.3]$$

The originality of the modeling used in DEFT'07 lies essentially in the discriminant aspects of the model. In the learning phase, the n-lemma counts are rescheduled in proportion with the discriminating power. This is estimated according to a point of view which complements the Gini impurity criteria:

$$G(w,h) \approx \sum_t P_t^2(t|w,h) \qquad [9.4]$$

This formula is employed to give more importance to the (w,h) events which appear only in one class ($G(w,h) = 1$) as compared to the other distributions. In the worst case, scenario $G(w,h)$ is $1/|T|$, if T is the class set.

Figure 9.1 illustrates the F-score performances of an incremental integration of the methods used. However, the order displayed has no impact on the final integration: it

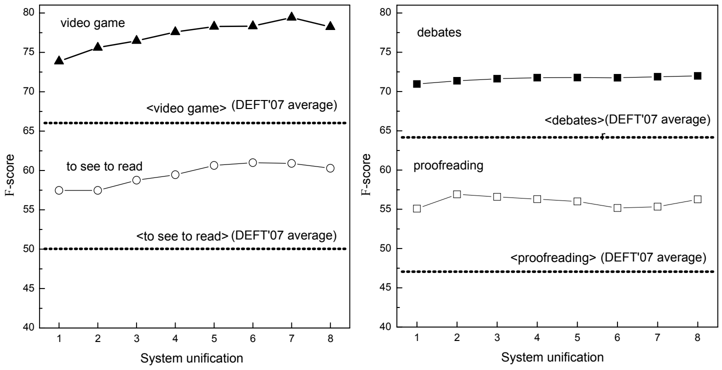


Figure 9.1. *F-score of unification following the 9 methods added. 1: BOOST_BASELINE; 2: 1 ∪ BOOST_BASESEED ∪ BOOST_SEED; 3: 2 ∪ Information theory; 4: 3 ∪ 1-gram; 5: 4 ∪ NATH_TORCH; 6: 5 ∪ TIMBLE; 7: 6 ∪ BOOST_CHUNK; 8: 7 ∪ SCT*

was only chosen to better illustrate the results. It can be seen that these results are above the average of results of the teams having participated in the DEFT '07 challenge, with all corpora combined.

In Figure 9.2, an F-score of the validation set (V) vs. the test set (T), on the four corpora is shown. A remarkable coincidence between the two is observed, which signifies the learning and validation strategies in the five subsets and the integration of several rankers worked well.

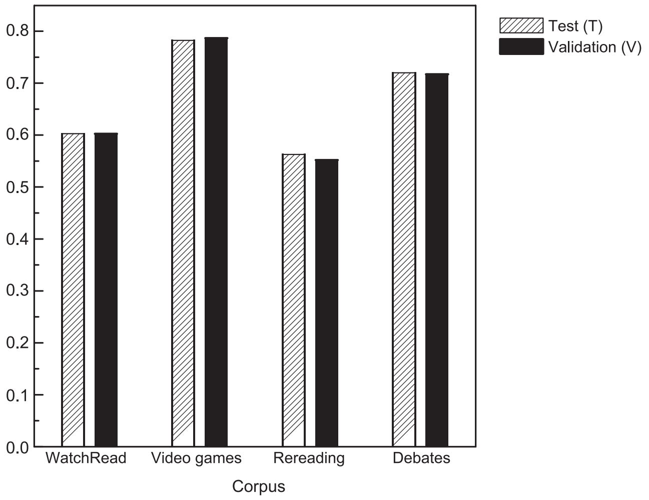


Figure 9.2. *Comparison of the F-score of the validation set (V) vs. the test set (T) for each of the corpora, obtained by the integrated system*

The following section is dedicated to a variation of the 8th model described in section 9.2.3.1. This variation has been implemented since 2007 in one of the systems developed by the LIA, and most often yields the best results.

9.3. Cosine weights – a second glance

The index most often employed in information retrieval to quantify the similarity of two documents, or a document and a query is the cosine similarity measure. Equation [9.5] is given as a reminder. This index is also used in text classification to estimate the similarity of document d and class c. The problems linked to the disproportion of the size of the document and the class will not be discussed due to lack of space:

$$\cos(d,c) = \frac{\sum_{i \in d \cap c} w_{i,d} \times w_{i,c}}{\sqrt{\sum_{i \in d} w_{i,d}^2 \times \sum_{i \in c} w^2{}_{i,c}}} \quad [9.5]$$

with:

$$w_{i,x} = T_{i,x} \times \log \frac{N}{F_i} \quad [9.6]$$

As shown in equation [9.6], the weights $w_{i,x}$ are generally obtained by calculating the product of two terms: the number of times (TF is written $T_{i,x}$ here), where the term i appears in the segment or the class x ($x = d$ or $x = c$), and an inverse function (IDF) of the number of (F_i) segments among the set of N segments, where the term i appears at least once. We propose to enrich the weight calculation formula by adding a discriminant factor D and by allowing for a more or less larger elasticity of each of the three factors by an increase to a variable power:

$$w_{i,d} = T_{i,d}^{\alpha} \times D_i^{\beta} \times \log^{\gamma} \frac{N}{F_i} \quad [9.7]$$

$$D_i = G_i = 1 - I_i = \sum_{i=1}^{k} P^2(j|i) \quad [9.8]$$

The three exponents found in equation [9.7] are also parameters. Their values are conveniently estimated on a development corpus. In the case where the data will lead to a β value of zero and if it also happens that $\alpha = \gamma = 1$, the classical TF.IDF which corresponds to equation [9.6] will be returned. This proposition therefore integrates, like in a particular case, the weight calculation generally employed in information retrieval.

During DEFT'07 [TOR 07], it has been proposed to retain as an estimation of the discrimination power associated with term i, the Gini purity factor which is easily derived, as demonstrated in equation [9.8], from the Gini impurity factor I_i by taking its complement to 1. When the cardinality k of the set of classes is greater than 2, it is often observed that among them there is a subset containing the classes presenting a strong collection among themselves, whereas they are distinctly different from the others. It can be easily identified that terms belonging to the two classes are connected to one another but not to others. If they do not completely identify a class, it is useful to measure at which point they reject one or even several. Their refuting power can be evaluated due to a relaxation of the Gini factor G'_i (see equation [9.9]). For this, research is conducted on the distribution P_r obtained by combining two classes. Among which the provided $(k - 1)k/2$ possibilities gives the best Gini index:

$$G'_i = \max_r G_i(r) = \max_r \sum_{j=1}^{k-1} P_r^2(j|i) \qquad [9.9]$$

$$D_i = \lambda \times G_i + (1 - \lambda) \times G'i \qquad [9.10]$$

In order to play with the two tables, namely the determination and refuting tables, it is evident that it is only interesting when we combine the two criteria. (see equation [9.10]).

We note that the addition of a third term in equation [9.7] can be understood as a way to modulate each apparition of a term i in the document d up to a fraction of the unit proportional to discrimination power D_i. It remains to be defined what we will call a term and more specifically which are the lexical units meriting to be retained.

9.4. Which components for a opinion vectors?

Whenever possible, it is preferable to integrate components corresponding to lexical units formed from clustered terms into the opinion vectors rather than isolated words. This choice comes from an analysis of the difficulty of the problem.

Among the focus points researched to make a correct decision during the categorization process, it is possible to envisage the size of the text, length of the phrases, percentage of the root word, and adjectives or all other grammatical categories. However, it is natural to think that the focus points are the most appropriate and are in fact the terms which compose the text, and, through their intermediary, the concepts to which they refer. However, these terms are mostly polysemic and ambivalent (this is the case for words such as *terrible, amateur, price, costly, rent, double, simple, first, last*, and so on).

Since there is no system that is performant enough to make a semantic disambiguation viable, it is easy to go back to an automatic process which through these discussion choices introduces noise in the data. This can possibly be done by maintaining all the viable sense assumptions associated with a probabilistic value. Another path has been proposed based on the following observation: it is most often in a close context where these ambiguities may be resolved. From this, encompassing an ambiguous word in a sequence sticking it to its close environment better identifies its use, and especially serves as a more precise indicator of the polarity.

9.4.1. *How to pass from words to terms?*

A preliminary stage, named pre-processing, enables us to pass from words to terms. Essentially, during this phase, it consists of splitting the text into lexical units, where a larger point of view is normally taken. Since instead of being content with segmenting it into constituent parts, the concatenation of contiguous words is performed to obtain types of *molecules* increasing the power to influence the decision toward one or the other opinions which are to be determined. While passing, misspelled words can be corrected, the inflected forms can be brought back to their roots, lemma, or stem. In order to produce so-called *discriminant clusterings*, or more briefly *clusterings*, it is recommended that these reductions should not be practiced systematically, but only if justified according to the discrimination power defined in [9.10]). These rules serve to overhaul the texts which will be automatically inferred from the corpus itself and are never written by a human.

The recommended method comes back to automatically inferring each rule according to the annotated learning corpus under the constraint that the result of its application leads to a better identification of the class supplied in the annotation. Owing to a set of experiments made on a given corpus (IMDB in this case[14]), in the rest of this chapter, the large comparative studies on the behavior of four possible methods among others to compose the words in terms will be presented:

1) *Classical collocations*: classically, different tests can be employed to calculate the collocations on a corpus (whether it is annotated or not): χ^2, the mutual information or the test of the likelihood relationship. This last criteria is employed by following the indications given by ([MAN 00]). We note that the calculation can be done on the learning, development, and test corpus, without introducing a bias, since the annotations are not used, contrary to the three following methods;

2) *Collocations by class*: the same criteria as those retained for the calculation of collocations is used to calculate, class-by-class, a set of collocations (in this case named **coloclasses** in the following) specific to each class. We note that the different sets are united in the following. In this case, as in the previous and the following two, the procedure employed is iterative. From this, it is apparent that this is not optimal.

[14] See section 9.5 “Experiences”.

It would be if we sought to determine in on pass the collocations no matter the number of their components;

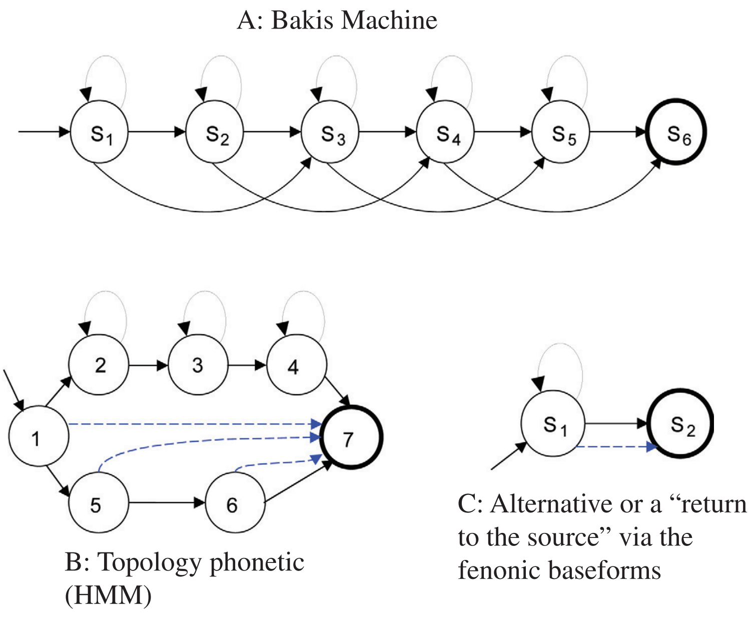
3) *Discriminant clusterings*: the same procedure as in method 2, with a slight difference: the coloclasses are only the product of intra-class characteristics, but they are not determined by the inter-class characteristics. The Gini criteria[15] fills this gap. It should also be noted that here the clusterings are calculated class-by-class on the learning corpus, and rendered in an overall set at the end of the first phase.

REMARKS 9.1.– By briefly explaining where the idea of producing clusterings has come from, certain steps are taken toward the potential refinement of the model. Some of these paths are specifically discussed in the examples given in the following section.

To model the production of acoustic vectors in the initial speech recognition systems, just as in the 1980s, the Bakis machines such as those represented in Figure 9.3 A, have been used. The fact that the machine must be trained by the word does not give access to very large vocabularies. The passage with a phonetic representation relieves this problem. At each automatically phoneticized word, it can correspond to a concatenation of the phonetic machines as illustrated in Figure 9.3 B, with more or less states, a certain number of full or empty transitions (discontinuous lines corresponding to the elided part of the phenomena) as well as the two paths differentiating fast flow and slow flow (presence of loops). In an attempt to get past this manner of proceeding which presumes, despite certain latitudes, a standardized pronunciation of the words, it has been envisaged, among other things, from real observations on the pre-learning corpora for each word or syllable automatically constructing a *fenonic baseform* machine dedicated to this base unit, and obtained by concatenating one or the other of the base cell chosen in a finite set of atoms as in Figure 9.3 C.

There is no question of confirming that there is a perfect transposition of the inference of acoustic machines to the acquisitions of clusterings from the annotated texts, but the analogy that can be established between the two domains can provide a glimpse of the improvements to come. First of all, we can note that a clustering is a concatenation in contiguous terms, which (in the sense D_i) is more discriminating than each of the terms composed of it. The empty transition model enabling jumping a state has not been incorporated in that which enables the modeling of ellipses. The presence of several possible paths to produce a sound on the graphic B, for example, leads to thinking that it would be just as appropriate to predict the development of alternative possibilities rather than, as done now, rewriting a word in another.

[15] The terms w_1 and w_2 are clustered if, and only if, w_1 and w_2 have been contiguously observed at least once in a sufficient number of examples and if $G(w_1, w_2)$ is larger than an empirically determined threshold.


A: Bakis Machine

B: Topology phonetic (HMM)

C: Alternative or a "return to the source" via the fenonic baseforms

Figure 9.3. *Markov machines*

The reader is invited to imagine all the other perspectives that can be traced from this parallel established between the two domains. A non-negligible point is insisted on to illustrate what is lost by applying the substitution and composition rules (it is coloclasses or clusterings) to produce a unique version of the text from which the categorization is done.

In note 42 in which Perreau has inserted some of the commentaries published in the satire Persius16 VI[16], the translator explains that three different senses can be attributed to the phrase *"Non adeo inquis exossatus ager juxta est"* according to the assumptions made on the punctuation and such that *adeo* is taken for an adverb or for a verb. Only two are retained. If it is read: "*Non adeo, inquis, exossatus ager; juxta est*", the meaning would be *"the heritage is not already in such a good state; as you wish"*; whereas if the reading is: *"Non adeo, inquis; exossatus ager juxta est"*, the passage can be interpreted as follows: *"I don't want, you say, any succession; Nearby, I have a well-cultivated field"*.

Even if it is not probable that these systems aim to process texts in Latin, this seems to be a good opportunity to choose an example in the language which no longer

[16] *Persian Satires* translated and commented by A. Perreau, Paris: C.-L.-F. Panckoucke, 1832.

comprises some insiders in order to give others the image what these machines are up against. Not only the texts that are submitted to them which are written in a language which is, for them, far from being alive, but also they would still have to make the distinction between two possible meanings of the same phrase. To come back to the problem, it is interesting to make the decision of ranking a text in class X because by linking it according to the rule (or automatons) of the class X with the discriminating parameters learned from this class, a better cosine was obtained than class Y by linking the text according to the rules of the other class Y.

9.5. Experiments

The experiments presented here have led to the IMDB[17]. These data are described in a detailed manner in two publications [PAN 04, PAN 08]. This corpus is particularly well balanced in the sense that it contains 1,000 negative criticisms and 1,000 positive criticisms on cultural products (mostly films). They have been split into 1,400 for learning (700+ and 700−) and 600 for test (300+ and 300−). The development corresponds to a cross-validation performed seven times on a seventh learning (of 200 criticisms).

The discriminant terms (Table 9.2) obtained from the two iterations are mostly "more speaking" than the discriminant words (Table 9.1).

POSITIVE CLASS			NEGATIVE CLASS		
22	1.0	magnificent	13	1.0	sucks
13	1.0	darker	12	1.0	vomit, justin
12	1.0	en	12	1.0	insulting, incoherence
11	1.0	seamles, ourselves	12	1.0	atrocious, 3,000
11	1.0	organized, lovingly	11	1.0	jolie
10	1.0	thematic, melancholy	10	1.0	uh, dud, degenerate
10	1.0	gattaca, gaining	9	1.0	shoddy, angelina
9	1.0	sullivan, steady	9	1.0	overwrougth
9	1.0	ideals, comforts	8	1.0	silverstone, missfire,liu
8	1.0	revolutionnary	8	1.0	shuttle, nutty
8	1.0	widescreen, perceive	8	1.0	horrid, haunted, coyote
8	1.0	juges, glorify	8	1.0	conceived, brukheimer
46	0.9	outstanding			...
		...	18	0.9	uninvolving
		...	29	0.9	ludicrous

Table 9.1. *Example of discriminating words*

[17] Corpus available at: http://www.cs.cornell.edu/People/pabo/movie-review-data/.

POSITIVE CLASS		NEGATIVE CLASS	
22	is-excellent	13	is-poor
13	well-worth	12	is-terrible
12	right-time, is-rare	12	stupidity-of
11	characters-this	12	one-of-worst, 's-as-if
11	son-'s, very-effective, terrific-as	11	worst-movie, this-mess
10	other-film, disney-animated	10	to-waste
10	only-problem-with, ed-harris	9	up-a, amounts-to, SSTAR-i
9	world-that, see-by	9	this-turkey, in-bad
9	reality-of, out-of-life	8	case-i, aren't-good
8	of-pace, of-finest	8	not-enough-to, mission-to-mars
8	mood-of	8	mess-that, (-scott, can't-save)
8	is-fantastic, great-thing	8	is-not-good, bad-dialog
46	i'm-not-say, characters-with		...

Table 9.2. *Examples of discriminant terms*

In Figure 9.4, it is possible to see how the number of composition rules of the terms evolve as a function of the number of iterations for each of the four assumptions discussed in section 9.4.1. As previously mentioned, this is done through the bias of rules of rewriting automatically inferred from the corpus. By playing on the thresholds discussed in note 15, a particular emphasis can be put on the rules of clustering incorporating a negation (some examples are given in Table 9.3).

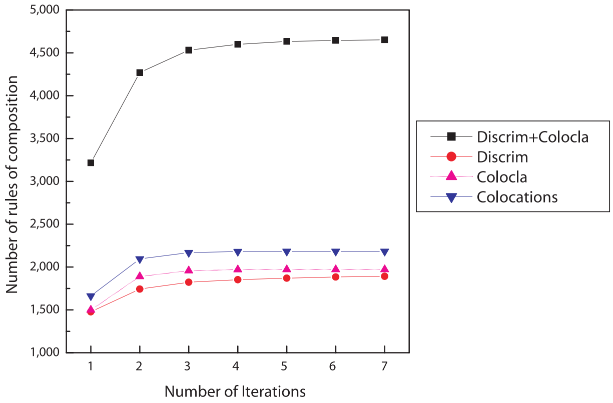


Figure 9.4. *Number of clustering rules automatically obtained from N iterations on the IMDB corpus*

doesn't-even	work
doesn't-go	anywhere
doesn't-have	clue
doesn't-help	either
doesn't-know	how
doesn't-really	matter

Table 9.3. *Example of IMDB clustering rules incorporating a negation*

Normalization and generalization of terms

Provided that the discriminant power of a term does not reduce, the coverage can be increased by replacing one (even many) of its components with another. These rewritings can be justified by suppression of the repetition or tripling of letters, the final suppression of certain words, or even the substitution of letter sequences. For these three cases, examples of rules obtained automatically from the IMDB corpus while maintaining the discrimination power from D to 1 are represented in (Table 9.4). The numbers which precede each rule are the apparition frequencies in the learning corpus after unification.

Repetition of letters			**Suppression of finals**			**Substitutions of letters**		
21	uhh	uh	15	insultingly	insulting	9	atheism	atheist
15	moll	mol	14	atrociously	atrocious	8	unnerve	unnerving
13	lasser	laser	10	sullivans	sullivan	8	deliberation	deliberate
11	ooooh oooh	ooh	10	departments	depart	7	perrier	perry
8	unerving	unnerving	9	wilderness	wilder	6	ideologies	ideology
7	traveller	traveler	9	perceived	perceive	6	ideological	ideology
7	tatoo	tattoo	8	ineffectuality	ineffectual	6	homophobic	homophobe
7	surveilance	surveillance	8	divinely	divine	5	intolerant	intolerance
7	kauffman	kaufman	7	enchantment	enchant	5	homophobia	homophobe
7	cassanova	casanova	7	barreness	barren			

Table 9.4. *Example of repetition, suppression of finals and substitution of letters*

Some of the rules can be discussed and a certain noise can be introduced but it is the price we have to pay to escape supervision of the process (which is translated by writing, rereading, and correction of rules).

9.5.1. *Performance, analysis, and visualization of the results on the IMDB corpus*

Since there are only two balanced classes in each of the sub-corpus of the IMDB set and since the system never abstains from answering, here the precision is equivalent to recall and to the F-score.

excellent-acting	excellent-performance
excellent-movie	excellent-film
lack-of-chuckle	few-chuckle
save-private-ryan-is	save-private-ryan
unstructured	poorly-construct

Table 9.5. *Examples of IMDB rewriting rules*

9.5.1.1. *IMDB performance*

The results obtained are equivalent in the test and development, and are the best of the cases at the same level as those published in [PAN 04]. As expected, the collocations calculated in a general manner give poorer results. The coloclasses and clusterings give equivalent results, but it is the union of the two sets which achieves better performance. The first iteration enables a leap of the results. After that, at best stagnation is observed, at worst oscillation or decrease. Finally, the comparison between the scores observed at the first iteration and the following demonstrate the gain due to the composition and rewriting rules.

Figure 9.6 shows that performance does not improve after the sixth iteration due to the good reason that there are no composition or rewriting rules which add to one or the other of the four assumptions retained (see Figures 9.4 and 9.5).

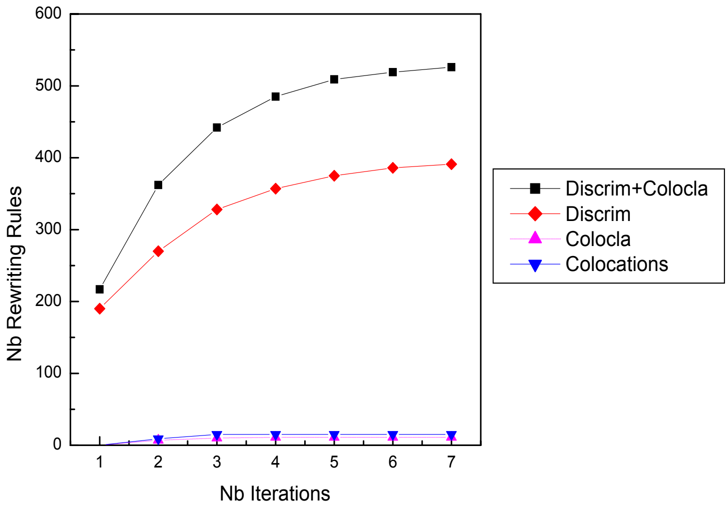


Figure 9.5. *Number of rewriting rules automatically obtained from N iterations on the IMDB corpus*

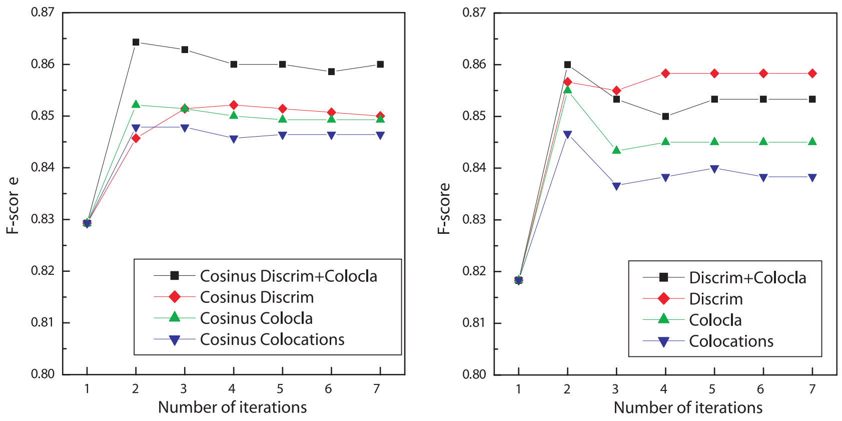


Figure 9.6. *IMDB development performances*

9.5.1.2. *Presentation and analysis of a 857_17527 IMDB example*

```
<DOCUMENT id="857_17527">
<NOTE valeur="neg"/>
<TEXT>
Claire Danes, Giovanni Iibisi, and, Omar Epps make a likable
trio of protagonists, but they’re just about the only palatable
elements of the mod squad, a lame-brained big-screen version
of the 70s TV show. the story has all the originality of a
block of wood (well, it would if you could decipher it), the
characters are all blank slates, and Scott Silver’s perfunctory
action sequences are as cliched as they come. by sheer force
of talent, the three actors wring marginal enjoyment from the
proceedings whenever they’re on screen, but the mod squad is just
a second-rate action picture with a first-rate cast.
</TEXT>
<LIA_TAGG>
<s> ZTRM <s>
claire NNP claire
danes NNPS danes
, YPFAI ,
giovanni <UNK> giovanni
...
```

Example 857_17527 after 7 iterations:

```
HYP = NEG 0.52002 HYP = POS 0.47998 REF = NEG
```

DEB claire danes, giovanni ribisi, omar epps make likable trio of protagonist but ’re just about-only palatable element of mod squad lame brain big-screen version-of 70 tv-show story have-to-do-with all originality of block of wood (well it-would if-you could decipher it) characters-be blank slate scott silver ’s perfunctory action-sequences be as cliched as come-by sheer force of talent three actors wring marginal enjoyment from proceedings whenever ’re-on screen but mod squad is-just second rate action picture with first-rate cast FIN.

It is always helpful to explain to the user the reasons that have driven the system toward making such and such a decision. This is classically done by framing, in a more or less subtle manner, each term with a color describing the class to which it contributes the most. However, this can also be done by creating an analogous table to Table 9.6 for recapitulating the 10 or 20 terms of the text contributing the most to the choice of each class.

TOP TEN NEGATIVE			**TOP TEN POSITIVE**		
marginal	0.29	0.29	come-by	0.0289	0.0289
about-only	0.0157	0.0446	picture	0.021	0.0499
action-sequences	0.0143	0.0589	version	0.0107	0.0606
perfunctory	0.0143	0.0732	first-rate	0.0058	0.0664
proceedings	0.0104	0.0836	likable	0.0054	0.0719
characters-be	0.0093	0.0929	enjoyment	0.0051	0.077
originality	0.0086	0.1015	palatable	0.0045	0.0815
characters	0.008	0.1095	element	0.0033	0.0848
is-just	0.0061	0.1156	version-of	0.0031	0.0879
sequences	0.0052	0.1208	it-would	0.0021	0.0900

Table 9.6. *Terms (under column 1) having the highest contribution (under column 2) to orient the decision toward one or the other two classes, below column 3 is a combination of the problems*

9.6. Extracting opinions from speech: automatic analysis of phone polls

Initial sections of this chapter have processed the expression of opinions in written texts. In this section, the problem of extracting opinions from oral messages will be presented, particularly the automatic analysis of opinion polls. The main difficulty of this type of corpus is the processing of spontaneous speech, typical for oral expression of opinions, for which Automatic Speech Recognition (ASR) nowadays are still imperfect.

Several studies have been carried out to identify positive or negative emotions in vocal messages collected in dialog situations. Acoustic parameters such as the fundamental frequency, energy of the signal, or even the values of the formatives have been used in conjunction with linguistic parameters [LEE 05]. A unification of linguistic parameters and acoustic parameters is also proposed [LIT 06] to predict the emotions attached to a vocal message. The selection of an optimal set of acoustic parameters linked to emotions is discussed in [NEI 06].

The application described in this section differs from the previous studies in that it focusses on detecting opinions that have not necessarily been expressed by emotional speech. It is for this reason we restrict our study to linguistic parameters. After describing the corpus used, two opinion detection methods will be used, one based on labeling methods performed on the ASR transcription modules; the other method consists of directly integrating this detection stage in the ASR process in order to increase the sturdiness of the detection.

9.6.1. *France Télécom opinion investigation corpus*

People are invited by a short message to call a free number to express their satisfaction of the customer service that they recently telephoned. By entering this number, the vocal message invites them to leave a message:

"*[...] You have recently contacted our customer service. We wish to ensure that you are satisfied with the help received during your phone call. You can leave your answer after the beep. Do not hesitate to share all your comments and suggestions for our service, since this will help us to improve it. We wish to thank you for your help and are always available to you. Leave your message after the beep*".

Since the messages were originally recorded by an automatic operator, no natural feature has facilitated automatic processing: no recommendation on the mode of elocution, open question, and even an incitation to leave comments. Thus, the collected messages are *realistic* and have varying length (from tens of words to hundreds of words). For this study, 1,779 messages collected over a period of 3 months, have been manually transcribed at the level of words, opinions, and markers (disfluency indications and discursive markers).

The corpus was divided into three sub-corpora Around 50% of the phrases make up the learning corpus, 33% account for the development corpus, and 17% the test corpus. The application's lexicon comprises about 4,500 words.

Analysis of user satisfaction is performed by the poll analysis team according to 3 dimensions: the quality of the reception (denoted as *reception*), the waiting time to access the service (denoted as *wait*), and the efficiency of the service (denoted as *efficiency*). This last dimension is most represented in the corpus. It concerns both the evaluation of the responses with respect to the expectation of the user (in other words, has the problem been solved?). But it also refers to the quality of the information provided. Each subjective expression can receive two polarities: *positive* and *negative*. There are thus a total of six labels to characterize the subjective expression of the corpus.

In manual transcription, in each message, these expressions are indicated by a label. The corpus is in segments, each with one or more specific opinion. The aim of the automatic comprehension module is to find these segments and label them with one of the six labels. An example of a message with reference labels is given in Table 9.7.

> “*yes Mr SURNAME NAME I phoned the customer service yeah* **<seg label=reception,pos>** *I was very well received***</seg>** *some* **<seg label=efficiency,pos>** *good information* **</seg>** *except that* **<seg label=efficiency,neg>** *it still does not work* **</seg>** *therefore I don't know if I made a bad correction or there is a problem otherwise* **<seg label=efficiency,pos label=reception,pos>** *the reception and advice was very appropriate* **</seg>** *even though* **<seg label=efficiency,neg>** *no positive result was obtained* **</seg>** *thank you goodbye*”

Table 9.7. *Message example of the France Télécom opinion corpus with several opinions with the segment markers*

Nb concept by message	Distribution (% corpus)	Average size (nb words)
0	19.2	61.0
1	51.3	40.3
2 and more	29.5	60.8

Table 9.8. *Distribution of the messages in the France Télécom opinion corpus as a function of the number of concepts expressed*

Table 9.8 shows that the average size of a message does not increase as a function of the number of concepts that are expressed. It is interesting to note that more words are required to express none of the concepts researched than the number of words to express one. This is explained, by the fact that the speaker can express himself *off topic* just as much as on the cause of the problem or on his personal situation as on his feelings on the customer service and also, by the fact that a same segment of the message can support several criteria.

Concerning the average number of words necessary to express a concept, the concepts describing a negative polarity often require more words than the concepts describing a positive polarity. Thus: “*very good reception super very good thank you*” will be labeled *satSerAcc satisfied by the reception service* whereas “*I asked to remove the option but since I have still received numerous call I would like you to cancel them*” would be labeled *insatSerEff* for *unsatisfied by the service efficiency*.

One of the problems that these messages point out is that an identical concept can be seen several times in a message with opposite opinions. This is the case when the person is not entirely satisfied (e.g. satisfied by the customer service but not by the results) or that a temporal notion enters its discourse. An example of this type of message is given in Table 9.7.

9.6.2. *Automatic recognition of spontaneous speech in opinion corpora*

The transcription of spontaneous speech remains one of the challenges with which the ASR methods struggle even now. At the acoustic level, this speech will be characterized by hesitation, pauses, in the interior of words or even truncated words; at the linguistic level, the statements will often be truncated, glazed with auto correction and discursive markers difficult to predict by the statistical language models employed. Table 9.9 presents the transcription of a verbal message issued from this corpus which, without being part of the most difficult messages to process, illustrates well the type of difficulty shown.

> "*oh hello so its XX i don't know if you know who i am so in relation to the satisfaction of my personal satisfaction in relation to your service i would say that overall you have a very good customer service which knows how to listen not that i have much to it is a very very good if not in relation to what you have just put in place is a very good idea since there is a good response rate it's true that i'm obliged to recall several times and still someone takes the time to respond because when i'm told to give ideas I had nothing in my mind so i was forced to hang up and think about what i was going to say to you i think that that is the collective point or it is the negative point otherwise i am very satisfied overall otherwise there is one thing that i noted i have two accounts with you i find this a little inconvenient to not be able to access both at the same time when i call customer service so i find that a little disappointing that i am forced to spend more because its not the same person that deals with my folder it would have been good to regroup both folders so when i call i can accesses both folders separately otherwise i am grateful for the pleasantness and amiability of your customer team they are very nice and listen well and i wold like to thank you goodbye good day good evening*"

Table 9.9. *Message example of the France Télécom opinion corpus*

This message is an example of spontaneous speech or *unprepared speech*. The numerous retakes and continuous correction in it illustrate the definition, and give rise to enormous transcription difficulties for the ASR systems. Even by assuming that the ASR models are sufficiently sturdy to transcript these messages into words, this type of output is not necessarily usable through the systems used for comprehension. In fact, in a spontaneous statement, the transcription of words represents only 1D of the message. It misses information linked to the signal of the speech itself: prosody, expressivity, and quality of the voice. Without this information, the message becomes difficult to understand, and it is not adapted to an in depth technical analysis, irrespective of the syntax or semantic.

In the absence of a viable representation of information linked to signal, the default analysis consists of looking into the statement for *nuggets* of information which characterize and respond to the task aimed for. For example, in the message of Table 9.9, the *nuggets* linked to the detection task are as follows:

– “overall i am quite satisfied”;

– “a very good customer service who knows how to listen”;

– “it is very very good”;

– “it is the negative point”;

– “overall i’m very satisfied”;

– “I find that a little inconvenient”;

– “I find that a little disappointing”;

– “I thank you for your pleasantness and amiability”;

– “your customer service providers are very very nice and they listen well”;

– “I thank you”.

It is from these elements that a characterization of the message can be made, for example through automatic classification methods using the segments initially detected. If these segments can be relatively well modeled as being in number in the messages of learning corpus, it is not the same with other parts of the message, characterized by a large variation, and badly accounted by the statistic models due to this variation.

In fact, from the degree of freedom left to the user in the statement of their message, there is a large dispersion observed in the distribution of the word frequencies. This is even more the case in the message portions where the users share the origin of their problem which can in itself be quite varied. Once the proper nouns are filtered, the learning corpus in its set contains 2,981 different words for a total of 51,056 occurrences. Nearly half of the words appear only once in the learning corpus and the lexical restriction to words whose occurrence frequency is larger than or equal to 2, leads to a lexicon of 1,564 words for rate of words outside the vocabulary equal to 2.8%.

One of the first bigram language models was led on a learning corpus. The error rate on the words obtained on the test corpus is very significant, mainly due to digression and commentaries present in the messages. This is illustrated in Table 9.10. As can be seen the results deteriorate strongly when the size of the messages increases.

% word-error		> 20	> 30	> 40	> 50	> 60
	≤ 20	≤ 30	≤ 40	≤ 50	≤ 60	
average number of words	8.8	16.7	32.9	38.6	53.9	54.6

Table 9.10. *Correlation between the error rate on the words and the length of the messages (expressed in words)*

9.6.2.1. *Segmentation of the opinion support messages*

The two strategies were developed to extract and classify the subjective expression of vocal messages: one relies on the manual or automatic transcription of the messages;

the other is integrated in the speech decoding process. These two strategies allow for the dissociation of the errors due to a bad word transcription of the opinion detection errors.

9.6.2.2. *Segmentation of messages with conditional random fields*

For the processing of transcriptions, a segmenter based on Conditional Random Fields (CRF) has been developed. CRF's have been used successfully in numerous labeling tasks such as morphosyntactic labeling or the detection of named entities. The main advantage of CRF's in relation to generative models such as the Hidden Markov Model (HMM) is the possibility of using a set of observations of a sequence to predict a label. It is thus not the only issue which constrains the attribution of a label to an observation but potentially all the previous and following observations. In this case, the learning corpus is formatted in a manner to associate a label to each word indicating if it is part of the opinion expression or if it belongs to an empty segment. During the analysis of a new message, the labels are asked by the CRF serve to find the segment boundaries. The labeler developed is based on the *CRF++* tool[18].

9.6.2.3. *Language models specific to opinion expressions*

In [CAM 06], a thematic language model was introduced to solve the problem of poor modeling of off-subject messages in relation to the task. The idea is to explicitly model only the opinion carrying messages. For this, a sub-corpus was extracted for each label which regroups the set of segments associated with each label in the initial learning corpus. A language sub-model is then estimated for each label according to the associated sub-label. Furthermore, a bigram type encompassing model on the labels has been estimated to model the sequences between the different opinion segments. The portions which correspond to none of the opinion expressions are themselves modeled by a context loop phenomena, without *a priori* constraints on the sequence phenomena. Finally, a supplementary sub-model was estimated for the segments which correspond to politeness formulas, often found at the start and end of the message. In fact, these segments show a strong regularity and their modeling avoids larger deviation of the decoding in the loop phenomena model. The set is compiled of a unique model presented in Figure 9.8.

9.6.2.4. *Classification opinion*

To assign an opinion label to a segment containing a subjective expression, a classification method named *AdaBoost* [SCH 00] is used for each dimension, the segments labeled manually from the learning corpus. The simple rankers correspond to regular expressions automatically constructed on the corpus words. The complete method is presented in [CAM 09]. The ranker also gives a confidence score, and this score is used in the strategy presented in Figure 9.7 (β threshold).

[18] Toolkit CRF++: http://www/chasen.org/ taku/software/CRF++/

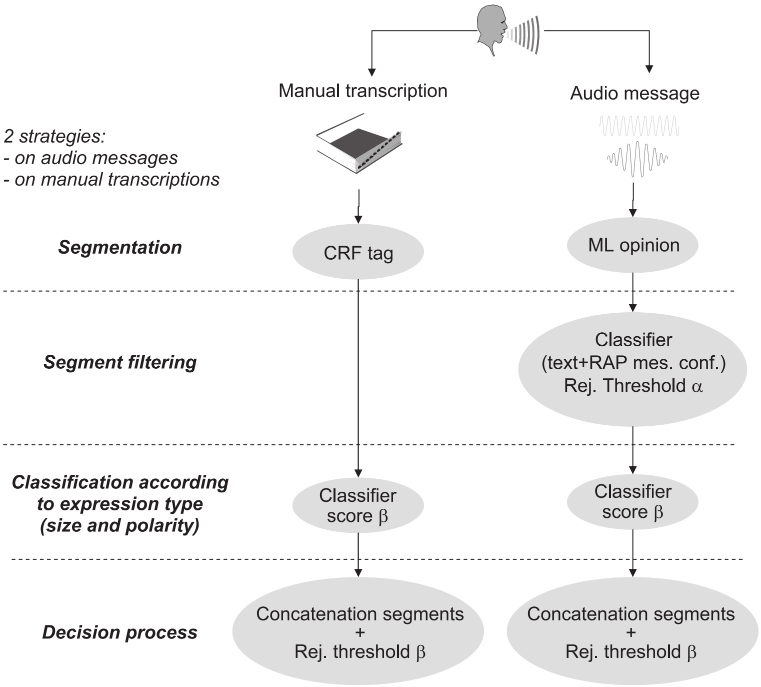


Figure 9.7. *Subjective expression detection and classification strategies for manual transcription and audio messages*

9.6.3. *Evaluation*

The two message segmentation strategies have been evaluated in terms of precision and recall on the opinion detection task on the vocal statement corpus. Four lines are compared in Figure 9.9:

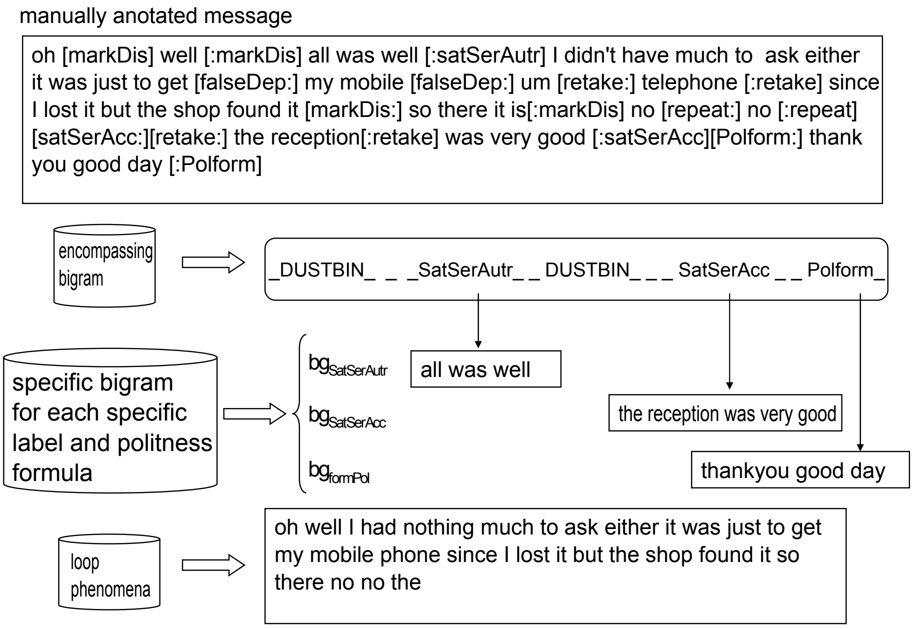
– manual transcription + manual segmentation;

– manual transcription + CRF segmentation;

– auto transcription + CRF segmentation;

– integrated method: transcription + segmentation (opinion language model).

As we can see, the results are strongly deteriorated when going from manual transcription to automatic transcriptions. If the CRF's give comparative results to manual transcriptions, they are strongly deteriorated by the high error rate of the transcriptions which is inevitable on this type of corpus. The detection method of the opinion integrated in the ASR process due to the opinion language models reinforces the sturdiness of the detection system.



manually anotated message

oh [markDis] well [:markDis] all was well [:satSerAutr] I didn't have much to ask either it was just to get [falseDep:] my mobile [falseDep:] um [retake:] telephone [:retake] since I lost it but the shop found it [markDis:] so there it is[:markDis] no [repeat:] no [:repeat] [satSerAcc:][retake:] the reception[:retake] was very good [:satSerAcc][Polform:] thank you good day [:Polform]

encompassing bigram
DUSTBIN _ _SatSerAutr_ _ DUSTBIN_ _ _ SatSerAcc _ _ Polform_
specific bigram for each specific label and politness formula
$bg_{SatSerAutr}$
$bg_{SatSerAcc}$
$bg_{formPol}$
all was well
the reception was very good
thankyou good day
loop phenomena

oh well I had nothing much to ask either it was just to get my mobile phone since I lost it but the shop found it so there no no the

Figure 9.8. *Example of decoding with three types of thematic language models used*

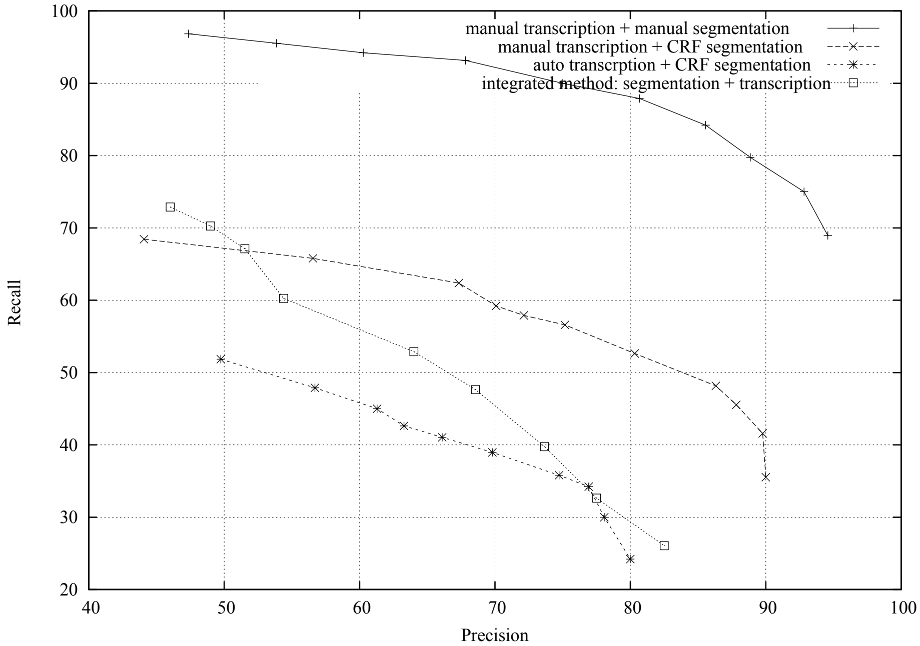


Figure 9.9. *Comparison of the segmentation methods for opinion detections*

9.7. Conclusion

In this chapter, it has been shown how it is possible to consider the detection of opinions as a thematic classification problem. Employing identical methods in both cases, should facilitate the conjoined processing of the two tasks.

After having skimmed over different evaluation campaigns, we have focused our discussions on the methods and strategies employed by the LIA during its participation in DEFT 2007. After having discussed a unification strategy, the method specifically focused on the introduction of discriminant criteria on a similarity index, and the choice of the components in a vectorial representation of the documents was discussed.

For each of these approaches, experiments have been reported in order to show that the performance of these systems are situated at the best level and enable the user to understand why such types of decision were taken.

Finally, the specificities related to orally expressed opinions during investigations performed by France Télécom with respect to their clients were studied.

9.8. Bibliography

[AZÉ 05] AZÉ J., ROCHE M., "Présentation de l'atelier DEFT '05", in *Proceedings of TALN 2005 - Atelier DEFT '05*, vol. 2, p. 99-111, 2005.

[AZÉ 06] AZÉ J., HEITZ T., MELA A., MEZAOUR A.-D., PEINL P., ROCHE M., "Préparation de DEFT '06 (DÉfi fouille de textes)", in *Proceedings of Atelier DEFT '06*, vol. 2, 2006.

[BOU 08] BOUDIN F., EL-BÈZE M., TORRES-MORENO J.-M., "The LIA update summarization systems at TAC-2008", in *Proceedings of the Text Analysis Conference 2008*, Gaithersburg, USA, 2008.

[BÉC 00] BÉCHET F., NASR A., GENET F., "Tagging unknown proper names using decision trees", in *38th Annual Meeting of the Association for Computational Linguistics*, *Hong Kong*, *China*, p. 77-84, 2000.

[CAM 06] CAMELIN N., DAMNATI G., BECHET F., MORI R.D., "Opinion mining in a telephone survey corpus", in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, USA, p. 1041-1044, 2006.

[CAM 09] CAMELIN N., BÉCHET F., DAMNATI G., MORI R.D., "Opinion analysis of spoken surveys", 2009.

[COL 02] COLLOBERT R., BENGIO S., MARIÉTHOZ J., Torch: a modular machine learning software library, in Technical Report IDIAP-RR02-46, IDIAP, 2002.

[DAE 04] DAELEMANS W., ZAVREL J., VAN DER SLOOT K., VAN DEN BOSCH A., TiMBL: Tilburg Memory Based Learner, Version 5.1, Reference Guide, Report, ILK Research Group Technical Report Series, 2004.

[DAN 08] DANG H., OWCZARZAK K., “Overview of the TAC 2008 update summarization task”, in *Proceedings of the Text Analysis Conference (TAC)*, 2008.

[ELB 05] EL-BÈZE M., TORRES-MORENO J.-M., BÉCHET F., “Peut-on rendre automatiquement à César ce qui lui appartient? Application au jeu du Chirand-Mitterrac”, in *TALN 2005 – Atelier DEFT ’05*, vol. 2, p. 125-134, 6-10 June 2005.

[FEL 98] FELLBAUM C. (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[FRE 96] FREUND Y., SCHAPIRE R.E., “Experiments with a new boosting algorithm”, in *Thirteenth International Conference on Machine Learning*, p. 148-156, 1996.

[GEN 08] GENEST P., LAPALME G., NERIMA L., WEHRLI E., “A symbolic summarizer for the update task of TAC 2008”, in *Proceedings of the First Text Analysis Conference*, Gaithersburg, Maryland, USA, 2008.

[HAT 97] HATZIVASSILOGLOU V., MCKEOWN K.R., “Predicting the semantic orientation of adjectives”, in *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, Association for Computational Linguistics, p. 174-181, 1997.

[KUH 95] KUHN R., DE MORI R., “The application of semantic classification trees to natural language understanding”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, p. 449-460, 1995.

[LAF 01] LAFFERTY J., MCCALLUM A., PEREIRA F., “Conditional random fields: probabilistic models for segmenting and labeling sequence data”, in *Proceedings of 18th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, p. 282-289, 2001.

[LEE 05] LEE C., NARAYANAN S., “Toward detecting emotions in spoken dialogs”, *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, p. 293-303, 2005.

[LI 08] LI F., ZHENG Z., YANG T., BU F., GE R., ZHU X., ZHANG X., HUANG M., “THU QUANTA at TAC 2008 QA and RTE track”, in *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, BC, Canada, 2008.

[LIT 06] LITMAN D., ROSÉ C., FORBES-RILEY K., VANLEHN K., BHEMBE D., SILLIMAN S., “Spoken versus typed human and computer dialog tutoring”, *International Journal of Artificial Intelligence in Education*, vol. 16, no. 2, p. 145-170, IOS Press, 2006.

[MAA 04] MAARTEN J.K., MARX M., MOKKEN R.J., RIJKE M.D., “Using wordnet to measure semantic orientations of adjectives”, in *National Institute for*, p. 1115-1118, 2004.

[MAC 06] MACDONALD C., OUNIS I., *The TREC Blogs06 Collection: Creating and Analysing a Blog Test Collection*, Report, Department of Computer Science, University of Glasgow Tech Report TR-2006-224, Glasgow, 2006.

[MAN 00] MANNING C.D., SCHÜTZE H., *Foundations of Statistical Natural Language Processing*, The MIT Press, 2000.

[NEI 06] NEIBERG D., ELENIUS K., LASKOWSKI K., “Emotion recognition in spontaneous speech using GMMs”, in *Ninth International Conference on Spoken Language Processing*, ISCA, p. 809-812, 2006.

[NEN 04] NENKOVA A., PASSONNEAU R., "Evaluating content selection in summarization: the pyramid method", in *Proceedings of HLT-NAACL*, vol. 2004, 2004.

[PAN 02] PANG B., LEE L. VAITHYANATHAN S., "Thumbs up? sentiment classification using machine learning techniques", in *Empirical Methods in Natural Language Processing*, p. 79-86, 2002.

[PAN 04] PANG B., LEE L., "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts", in *Proceedings of the 42nd ACL*, p. 271-278, 2004.

[PAN 08] PANG B., LEE L., "Opinion mining and sentiment analysis", in *Foundations and Trends in Information Retrieval 2(1-2)*, p. 1-135, 2008.

[SCH 00] SCHAPIRE R.E., SINGER Y., "BoosTexter: a boosting-based system for text categorization", *Machine Learning*, vol. 39, p. 135-168, 2000.

[TAK 05] TAKAMURA H., INUI T., MANABU O., "Extracting semantic orientations of words using spin model", in *ACL '05*, p. 133-140, 2005.

[TOR 07] TORRES-MORENO J.-M., EL-BÈZE M., BÉCHET F., CAMELIN N., "Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne? Application au défi DEFT 2007", in *AFIA/DEFT '07*, p. 119-133, 4-6 July 2007.

[TOR 09] TORRES-MORENO J.-M., EL-BÈZE M., BÉCHET F., CAMELIN N., "Fusion probabiliste pour la classification d'opinions", in *DEFT '09*, p. 10-25, 22 June 2009.

[TUR 03] TURNEY P., LITTMAN M., "Measuring praise and criticism: inference of semantic orientation from association", *ACM Transactions on Information Systems*, vol. 21, p. 315-346, 2003.

[VAP 82] VAPNIK V.N., *Estimation of Dependences Based on Empirical Data*, Springer-Verlag Inc., New York, USA, 1982.

[VAP 95] VAPNIK V.N., *The Nature of Statistical Learning Theory*, Springer-Verlag Inc., New York, USA, 1995.

[VAR 08] VARMA V., KRISHNA S., GARAPATI H., REDDY K., PINGALI P., GANESH S., GOPISETTY H., BYSANI P., SARVABHOTLA K., REDDY V. *et al.*, "Iiit hyderabad at TAC 2008", in *Text Analysis Conference*, USA, 2008.