An autonomous system designed for automatic detection and rating of film reviews.
Extraction and linguistic analysis of sentiments.

Grzegorz Dziczkowski [1,2] and Katarzyna Wegrzyn-Wolska [2]
[1]Ecole des Mines de Paris
35, rue Saint-Honore 77305 Fontainebleau, France
[2]Ecole Superieur d'Ingenieurs en Informatique et Genie des Telecommunicatiom (ESIGETEL)
1, rue de Port de Valvins 77-215 Avon-Fontainebleau Cedex, France
Email:{grzegorz.dziczkowski, katarzyna.wolska}@esigetel.fr

Abstract

This paper describes the functions of a system designed for the assessment of movie reviews. Such a system enables the automatic collection, evaluation and rating of film critics' opinions of movies. First the system searches and retrieves probable movie reviews from the Internet, especially those expressed by prolific reviewers. Subsequently the system carries out an evaluation and rating of those movie reviews. Finally the system automatically associates a numerical mark to each review, this is the objective of the system. This data constitutes the input to the cognitive engine. Our system uses three different methods for classifying opinions in critics' reviews. We introduce two new methods based on linguistic knowledge. Results are then compared with the overall statistical method using Bays classifier. The last step is to combine the results obtained in order to make the final assessment as accurately as possible.

1. Introduction and issue

With the growth of the Web, e-commerce has become very popular. A lot of websites offer on line sales and propose object ratings to their clients, for films for example. People like to check out other users' recommendations before making up their minds. Those profiles are very useful for the customers. The Recommender System was created (RS) in order to predict the potential choice of clients. RS allows people to make choices without any personal knowledge of the alternatives. Algorithms for suggestion are based on the experience and the opinion of other users. It is helpful to find recommendations from people who are familiar with the same problems, who have made similar choices in the past, whose perspective we value, or who are recognized experts [15].

RS provides correspondences between the users who have similar profiles. A new user has to create their own profile. The RS will suggest a new limited choice based on the similar taste of other users. The results of RS must not be tampered with for commercial reasons because this would make people distrustful. The effectness of this system depends on the data's quality and quantity. Our system supplies user profiles which are necessary for the algorithms of the cognitive engine. The main goal of the developed system is to collect a huge base of film reviews and automatically attribute marks which express the sentiments of the writer. Each review receivs a new mark and a user profile. The result of this treatment is the creation of a user profile database. Our system is based on the statistical and semantic representation of documents. Our work comprises the extraction and filtering of opinions from the text and the assignment of the mark to subjective sentences. The extraction and information filtering consists of the identification of quite precise information in a text in natural language and its representation in a structured form [13].

2. Related work

So far scientific research has not been able to automatically understand the written text. We should bear in mind however that these systems resulting from the work of automatic treatment of language carried out in the 80s made it possible to explore a generic approach of text comprehension. This meant that a large number of researchers started to describe natural languages in the same way as formal language. Maurice Gross [9] undertook with his team of the LADL (French Laboratory for Linguistics and Information Retrieval) the exhaustive examination of simple sentences in French, in order to have reliable and quantified data on which it would be possible to make rigorous scientific experiments. To exploit the linguistic knowledge an application called Unitex was created at LADL [14]. Unitex is an environment of enhancement used to build formalized descriptions of natural languages with all the coverage that this implies and apply them as texts of great size in real time. Unitex manages (in real time) texts of several mega-bytes for indexing according to morpho-syntactic criteria as well as searching for set phrases or semi-fixed phrases, and producing agreements and a statistical study of the results [11], [8].

Another way to detect an opinion automatically from the text is the use of a classifier. The statistical methods suppose that descriptions of the objects of the same class are divided by respecting a specific structure of the class. Learning methods based on an example are often used in information research on a large group of texts. Problems consist of constituting a representative corpus of the field in which we operate, and finding the rules or creating an operational model of this corpus. This model makes the system able to predict the correct behavior to adopt when a new candidate arrives for classification. Research in the of area opinion mining covers several topics such as the learning of semantic orientation of words, sentiment analysis of documents and analysis of opinions. Previous works closely related to our work include: document level sentiment classification (Turney [16], Pang, Lee [12], Dave, Lawrance [4]) and sentence level sentiment analysis (Riloff, Wiebe [18]).

The approach of Turney is presented in three steps. Firstly parts-of-speech are tagged, than pairs of consecutive words are extracted from reviews if their tags conform to given patterns. Next the semantic orientation (SO) of the extracted phrases is estimated using Pointwise mutual information (SO-PMI). At the end the average SO of all phrases is calculated.

The approach presented by Pang and Lee applied several machine learning techniques (like Naive Bayes NB or Support Vector Machine SVM) to classify movie reviews into positive or negative. First they detected subjective phrases and then the intensity of the polarity.

Dave and Lawrence in their approach add an initial selection of product features. After selecting a set of features and optionally smoothing their probabilities, they assign them scores and then place test documents in the set of positive reviews or negative reviews. When each term has a score, it's possible to add the scores of the words in an unknown document and use the sign of the total to determine a class. In the end the classification of the review using the sign is performed.

Another point of view is using learnt patterns presented by Rilloff and Wiebe. The approach is based on the use of a high precision classifier to identify subjective and objective sentences automatically. Then a set of patterns are learned from these sentences. Finally the learned patterns are used to extract more subjective and objective sentences.

3. Linguistic resources

Our approach is based on linguistic knowledge. In this section we present linguistic resources which are used in our methods. The linguistic resource used for the information retrieval and extraction are as follows: dictionaries, networks of recursive transitions (local grammar) and tables of lexicon-grammar.

The digital dictionaries employed by Unitex [14] describe both simple and complex words of a language. Dictionaries associate the word with a lemma and a series of grammatical, semantical and inflexional codes.

Grammar is a representation of linguistic phenomena by recursive transitions (RTN), this formalism is close to that of the finite state automaton. Many studies have highlighted the adequacy of automates on linguistic problems. A transducer is a graph with a finite number of states which shows entry sequences and associates sequences produced as an output. Generally a grammar represents sequences of words and produces linguistic information, for example information on the syntactic structure.

A local grammar [10] is an automaton representation of the linguistic structures which are difficult to formalize in lexicon-grammar tables or numeric dictionaries. The local grammars, represented in the forms of graphs, describe elements which concern the same syntactic or semantic fields. The linguistic descriptions grouped together in the form of local grammars are used for a large variety of automatic processes applied to the text. Thus various methods of lexical clarification were developed to implement grammatical constraints described before using this type of graph.

The corpora of text are represented by automates, in which each state corresponds to a lexical analysis. The linguistic phenomena are represented by local grammar, and are then translated into a finite state automat in order to be easily applied to the corpora of text.

Tables of lexicon-grammar are matrices that outline the properties of all the simple verbs which are described by syntactic properties. The lexicon-grammar tables supply the grammar of each element of the lexicon although each has almost unique behavior. With Unitex we can build grammar from such tables. The lexicon-grammar is a systematic description of the syntactic and semantic properties of the syntactic factors such us predicative verbs, nouns and adjectives. It is organized in groups of tables, which are associated with the syntactic category for example full verbs, verb supports, names, etc... A table corresponds to a particular syntactic construction and gathers all the words within this construction. Currently lexicon-grammar is especially developed for verbs and predicative phrases [15], [16].

4. General system architecture

The principle tasks of our system are: collecting the reviews from Internet, checking if the text found is a review, assigning a mark to the reviews and the presentation of the results. Our system is structured with a modular architecture organized in three main modules: collection of reviews, verification and notation of sentiments and data publication [Figure 1]. This paper is focused on the middle module shown in the figure below.

In order to assign a mark to the review we needed a group of characteristics which had already been evaluated - a learning base. We were able to find film reviews which had already been marked on various websites (e.g. IMDB, Amazon). We used that data (critics, users, marks) to create our learning base. We used a scale of marking from 1 to 5. We regrouped all the reviews by their mark. Thus we obtained 5 different groups of film reviews: a group of reviews with a score 1, 2 ... 5. [6]. Our research was limited to a base of reviews containing 200 000 inputs.

We developed and tested three different methods for assigning a mark to the reviews. These methods were based on different approaches to corpus classification. For each method we developed a classifier which separately assigned a mark. Finally we obtained three marks for each review, and those marks were not always the same. We used another classifier which correlated the three marks in order to obtain the final mark [5], [6], [7]. The final classifier only used the three marks so as not to repeat the characteristics which are used in previous classifications. In this way no single classifier is privileged. This is sufficient because we have already used all the characteristics in the previous classifiers. There is no need to repeat the characteristics in the final assignment of marks.

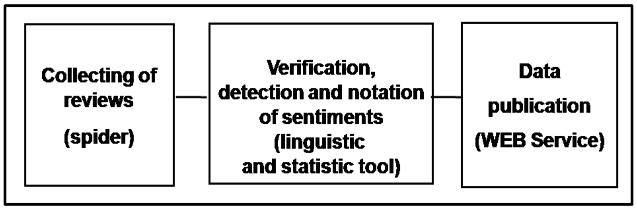
Collecting of reviews (spider)
Verification, detection and notation of sentiments (linguistic and statistic tool)
Data publication (WEB Service)

Figure 1. System architecture

We carried out tests of all classifiers for all groups of marks. The corpus of movie reviews used for the test contains 2264 sentences for a mark equal to 5, 1957 sentences for 4, 1308 sentences for 3, 1925 sentences for 2, and 1835 sentences for 1. The test corpus is the same for each classifier. At the end of each section describing classifiers we presented results using precision, recall and f-scores.

5. Classification and mark assignment

5.1 Verification, detection and notation of sentiments

Opinion mining is the most important task in our system. It is carried out by module: verification, detection and notation of sentiments [Figure 1]. The functional principles of this process (assignment of the mark to the reviews) are shown in figure 2.

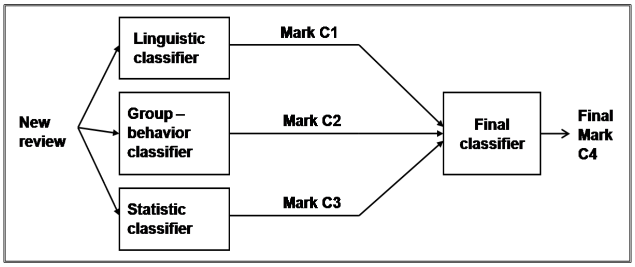

Figure 2. The process of mark assignment

For marking reviews we use three different approaches which are as follows:

- Linguistic classifier: For each sentence of reviews we assign a rule of grammar that expresses intensity of opinion.
- Group-behavior classifier: Statistical research on linguistic data to determine the behavior of reviews which have the same mark. The characteristics are for example: characteristic words, sentence length, corpus width, presence of negation, characteristic expressions, special punctuation. For the entire corpus of reviews we have calculated the distance between the characteristics of new reviews and the characteristics of the groups.
- Statistic classifier: Statistical research based on Bayes classifier, a categorizer of the probabilistic type founded on Bayes' theorem.

Finally the scores are combined with a neural network in order to obtain the best possible results. The final assignment is based entirely on the marks obtained from three classifiers.

5.2 Linguistic classifier

As we used the scale of marking from 1 to 5, we created a grammar in each group. This grammar is based on an analysis of the learning base, which contains about 2000 sentences for each mark group. For this part we used a linguistic treatment which requires lexicons and specialized grammar. The development of such resources is a long and tiresome task, which generally requires an expertise in the field and knowledge in data-processing linguistics such as the techniques of filtering, categorization of documents and extraction of information. Comprehension is seen as a transduction which transforms a linear structure, i.e. text (the linear structure) is transformed into an intermediate logico-conceptual representation, which is then used to draw conclusions. The semantic analysis aims to produce a structure representing as accurately as possible, a unit of the sentence, with its meanings and its complexity; then it has to integrate all structures into a single textual structure. Finally we obtain a logico-conceptual representation of the text [2], [10], [1]. Semantico-conceptual structures can be more or less broad, rich and complex and more or less ambiguous [5].

This part of the system was developed with Unitex application, the example of linguistic resources used is shown in figure 3. We use a linguistic analyzer Unitex to pre-treat, to lemmatize the words, to add synonyms, to detect negation, to add semantic classes to the words and lastly to build complex local grammars. Semantic classes are associated to the word and show the polarity and the intensity of the word. In order to associate semantic classes to the words we used a subjective word dictionary - General Inquirer Dictionary [1]. The General Inquirer is a mapping tool. It maps each text file with counts on dictionary-supplied categories.

The main purpose of linguistic classifier is the assigning of the mark in harmony with the sentiments contained in the review. The assignment of mark is carried out sentence by sentence. In order to create rules of grammar for each mark (in our case the mark from 1 to 5) the study of reviews from the learning base was carried out. In this way 5 grammars were created - one for each mark. Each grammar contains a lot of rules - local grammars. For each grammar more than 30 local grammars was created. In order to assign the mark to the new opinion, research is performed sentence by sentence so as to find the rule corresponding to the examined sentence. At the end of this treatment we obtained selected sentences of new reviews with corresponding rules. To obtain the final mark we calculated the average of marks corresponding to main grammars.

The construction of local grammars was done manually way by analyzing sentences from the reviews with the same mark associated. The local grammar can not be too general as this would make the results of the research too much ambiguous. If the local grammar is too specific and complex the application is uncertain because the quantity of silence increases significantly. The local grammars were created to detect the polarity and intensity of opinion in one sentence. Other classifiers used in our system perform the statistic classification. In linguisitic classifier sentiments detection is based local grammars forms. Other more statistical futures like typical words, typical expression, size of sentence, frequency of characteristic, word repetition, number of punctuation marks etc are not taken into account. Of course the typical words are in dictionaries with semantic classes and in local grammars, but the grammar is necessary for linguistic treatment.

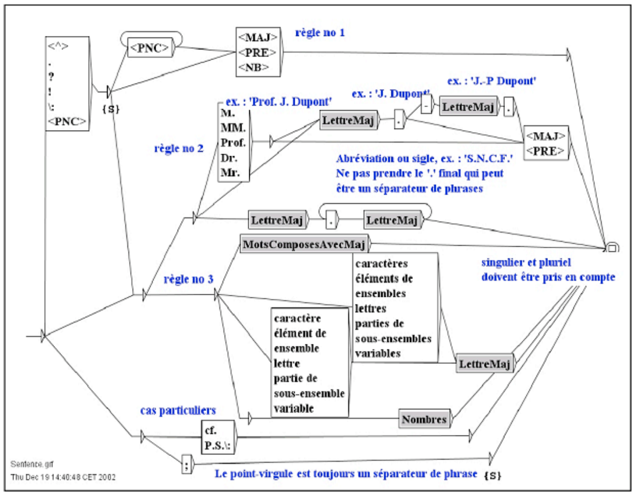

Figure 3. Linguistic resources

The creation of local grammar is a time-consuming task. The grammars used in our system were genereted in empiric way. We proceeded by adding a more complex level of linguistic analyzis, performing tests and then repeated the procedure. For each level we effected tests and calculated F-score. The final result of the rules of grammars was chosen to provide the best F-score. Unfortunately we can not be sure that our choice is the most coherent. We took into consideration that each classifier presented in our system should have its own futures. In spite of this method it's important to notice that the linguistic classifier gives the best results. Specifically we can see that the precision parameter is better than that which we obtained using other approaches. The results for linguistic classifier are shown in Table 1.

Table 1. Linguistic classifier results

	Precision	Recall	F-score
Class 5 *	72.4%	83.4%	76.5%
Class 4 *	70.8%	82.4%	76.1%
Class 3 *	67.8%	71.6%	69.6%
Class 2 *	62.5%	55.9%	59%
Class 1 *	76.3%	84.2%	80.1%

[1] http://www.wjh.harvard.edu/ inquirer/

5.3 Group-behavior classifier

In this section we present next classifier used to opinion notation. The general approach is based on checking whether the reviews with the same marks have common characteristics. Then we determine a behavior of reviews which have the same mark, so we determine a general behavior for each of 5 classes.

We have an enormous amount of assessed reviews, but in order to compare the methods we use the same learning base as for the previous classifier (200 reviews for each class). We gathered together all the reviews according to their mark. So we obtained 5 different groups of film reviews. Then, we tried to determine the future characteristics for each group. We defined all the parameters which could characterize the behavior of a group like:

- a characteristic word or expression,
- the sentence size,
- a review size,
- the frequency of repetition of several words,
- negation,
- the number of punctuation marks (!, ;), ?) and so on...

In this approach we present the statistical research on linguistic data. To determine group behavior we parse a large corpus of reviews with the same mark to find the characteristic futures. We assigned the semantic classes to our corpus word. Then we parsed the corpus to obtain statistical results. The results shown great differences between the characteristics of those groups. The creation of the behavior of groups enables us to determine to which group a new review may belong. For new reviews we calculate the distance between its characteristics and the characteristics of the groups.

We carried out tests of group-behavior classifier for all groups of marks. The corpus of movie reviews is the same as for the linguistic classifier.
The results are shown in Table 2.

Table 2. Group-behavior classier results

	Precision	Recall	F-score
Class 5 *	70.2%	71.4%	70.8%
Class 4 *	70.4%	72.4%	71.4%
Class 3 *	57.8%	62.6%	60.1%
Class 2 *	61.7%	57.9%	59.7%
Class 1 *	75.9%	78.3%	77.1%

5.4 Statistic classifier

In this section we present a general approach used in opinion mining. We present this method to compare the results from our approaches. The way of carrying out a classification is to find a characteristic of each class and to associate a function of belonging. Among the methods using this process we can quote decision trees, Bayes classifiers, method of SVM, etc. We used Naive Bayes classifier [3], [17]. In our research we used this classifier firstly to determine subjective and objective phrases and subsequently to assign a mark to the reviews. The general process nesessitates the preparation of learning bases for two classifiers: classifier of filtering phrases subjective / objective and classifier for assigning a mark. The intermediate steps are as follows:

- Pre-treatment
- Lemmatization
- Vectorization, calculating complete indexes
- Constitution of learning bases for each classifier
- Reducing the index dedicated to a classifier
- Adding synonyms
- Classification of texts

This method is generally used for text categorization, so we only present the results. We carried out tests of statistic classifier for all groups of mark. The corpus of movie reviews used in test is the same as for previous classifiers. The results are shown in Table 3.

Table 3. Statistic classifier

	Precision	Recall	F-score
Class 5 *	73.3%	67.7%	70.4%
Class 4 *	72.8%	60.4%	66%
Class 3 *	68.8%	50.4%	58.2%
Class 2 *	63.4%	44.4%	52.2%
Class 1 *	74.3%	64.9%	69.3%

6. Final assignment

So far, we have presented three different methods of automatically assessing a mark for reviews. Thus, we get three different assessments (one from each classifier). Ratings are not always the same. So another problem is the final evaluation of reviews. We need a final assessment, which will be forwarded to the Recommender System. We noticed that in the case of counting the final average results are worse than the results of the linguistic classifier, which gives the best results.

We also noticed that it often occurs that one classifier in specific situations gives better results, where as in other situations it may be another classifier. We give an example, frequently when the first classifier gives a score of 2 and the

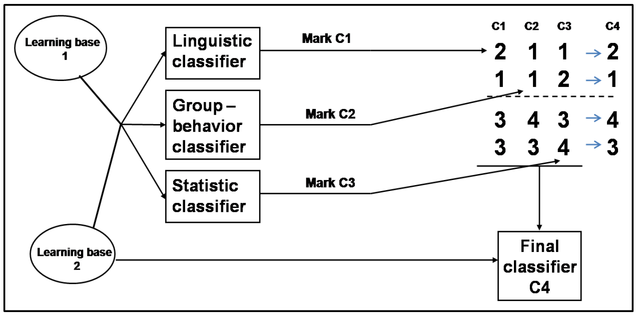

Figure 4. Final classier

two last classifiers scores equal 1, and the correct result is 2. Consequently, it is the first classifier, which is critical in this situation.

If, however, the two first classifiers give scores equal to 1, and the last score of 2, in this case the correct assessment is equal to 1. So in this case we notice that we should not count the final mark as the average in certain situations, because one classifier can be more influential. In the second example above the situation is similar, only in this situation the second classifier is influential with a mark equal to 4 when others give the mark of 3. We may notice many more examples of similar behavior. The examples described are shown in figure 4.

As the input to the final classifier we use marks from previous classifiers - marks from each classifier represented by probability of belonging to one of five classes of marks. For example the linguistic classifier assigns a mark in this way: the probability that a mark is equal to 5 is p=0.6, equal to 4 - p=0.2, equal to 3 - p=0.1 equal to 2 - p=0.1, equal to 1 - p=0. We used the neural network to determine the correlation of results. The use of neural networks is justified, because we have a very large database of reviews already assessed. It is easy to implement this data for a learning base. We use Multi-Layer Perceptrons MLP using backpropagation gradient algorithms. The process is shown on figure 5. We use:

- 15 input,
  3 classifiers give probability p_{ij} for each of 5 marks
  (i -classifier number, j -probability of mark for each class)
  Cl1 (5 - p_{15} , 4 - p_{14} , 3 - p_{13} , 2 - p_{12} , 1 - p_{11}),
  Cl2 (5 - p_{25} , 4 - p_{24} , 3 - p_{23} , 2 - p_{22} , 1 - p_{21}),
  Cl3 (5 - p_{15} , 4 - p_{14} , 3 - p_{13} , 2 - p_{12} , 1 - p_{11}),
- 3 layers,
- 1 output (final mark),
- new learning base of 200 reviews for each mark (1000 reviews in total).

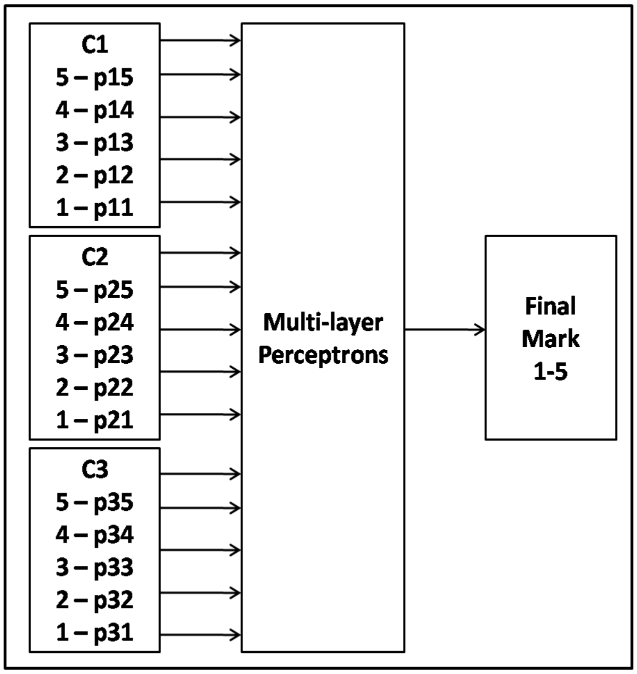

Figure 5. Multi-Layer Perceptrons

This way we improved the results which are better than results from the most accurately classifier - linguistic classifier.

7. Results

We noticed that we obtain better results with the linguistic classifier (section 4.1). The worst results were for the statistic Naive Bayes classifier. This proved the necessity of deep linguistic analyzis. We observed that the best results were obtained for the extreme opinion in each approach. It was easier to automatically mark and to judge the movies reviews with a mark equal to 1 or 5. This seems to be obvious, because extreme emotions are strongest. Moreover extreme reviews are more often longer so it favours the correct assessment. In spite of these improvements we made, we are still far from the ideal case. According to our results, and since it is necessary to start from the principle that more complex and complicated grammars are needed, we noticed that the linguistic classifier gives better results that the statistical or group-behaviour classifier.As we noticed that we have in several situations a more infuential classifier we improved our results again using neural networks (section 6). For this stage we based our approach only on the outputs from 3 classifiers previously described. We noticed that the results obtained either by calculating the average or based only on scores from each classifier in scale 1 to 5 were even worse than results form linguistic classifier. By implementation of neural networks for this stage and by taking into consideration each probability for each score for each classifier we improved our results for 3 to 7% depending on the class. The results are shown in figure 6.

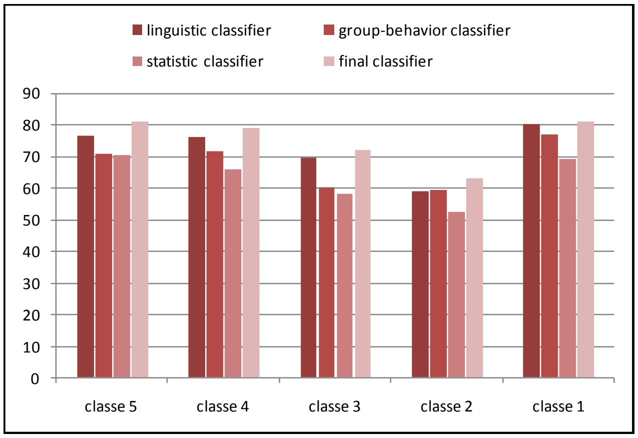

Figure 6. Results

8. Conclusions

The system presented carries out a collection of movies reviews and automatically assigns a mark to each review. This system is a support for RS. The goal of our work is to automate the whole system, particularly to assign a mark to individual user's reviews using sentiment detection knowledge. The system allows an automatic assignment of a mark. However, to increase the research on other fields it will be necessary to create a linguistic database and a new analysis of the different elements of the group's behavior.

We focused on the automatic search task for information in a corpus, more precisely on the linguistic analysis of sentiments. Our study for first classifier was made on the application "Unitex" since it's the tool that makes it possible to carry out a major search by using grammars, tables of lexicon-grammar and dictionaries. Our objective was to prepare the data and creation of complex local grammars. The second linguistic method is based on statistical researches on linguistic data to determine the behavior of reviews which have the same mark. We compared our results with a general statistical method using Naive Bayes classification.

We succeeded in the creation and in the integration of two linguistic approaches. This method made it possible to automatically assign a mark to the sentiments in movies reviews. The adjustment of the linguistic resources like the creation of the complex local grammars or the adaptation of the dictionaries was an important part of our work in improving the linguistic classifier. We obtained satisfying results, but it is necessary to specify that there remain several points to be improved. The solutions from the automatic information retrieval presented in this paper give an idea of the complexity of this field and highlight the need for making improvements. We also succeeded in the improvement of our results by using neural networks to combine the individual results.

References

[1] H. Alshawi. *The core language Engine*. MIT Press, 1992.

[2] H. Altai. The core language engine. In *ACL-MIT Press Series in Natural language Processing*. MIT Press, 1992.

[3] T. Cover. *Elements of Information Theory.* John Wiley, 1991.

[4] S. Dave, K. Lawrence and D. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW'03: Proceedings of the 12th international conference on World Wide Web*. ACM, 2003.

[5] G. Dziczkowski and K. Wegrzyn-Wolska. Graph based system purpose - built for automatic retrieval and extraction of the electronics data. In *Internet and Multimedia Systems and Applications*. ACTA Press, 2007.

[6] G. Dziczkowski and K. Wegrzyn-Wolska. Rcss - rating critics support system purpose built for movies recommendation. In *Advances in Intelligent Web Mastering*. Springer, 2007.

[7] G. Dziczkowski and K. Wegrzyn-Wolska. Tool of the intelligence economic: Recognition function of reviews critics. In *ICSOFT 2008 Proceedings*. INSTICC Press, 2008.

[8] B. Eriksson. Sentimen classification of movie reviews using linguistic parsing. In *Natural Language Processing*. CS 838, 2006.

[9] M. Gross. The construction of local grammars. In *Finite-State Language Processing*. MIT Press, 1997.

[10] H. Kamp. Evenements representations discursives et reference temporelle. In *Langages nb 64*, 1981.

[11] A. Kennedy and D. Inkpen. Sentimen classification of movie reviews using contextual valence shifters. In *Computational intelligence*. Blackwell Publishing LTD, 2006.

[12] B. Pang and L. Lee. Sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*, 2004.

[13] M. Panzienza. *Information extraction (a multidisciplinary approach to an emerging information technology)*. Springer Verlag (Lecture Notes in Computer Science), Heidelberg, 1997.

[14] S. Paumier. *De La reconnaissance de formes linquistique a l'analyse syntaxique.* These, Marne-la-Valee, 2003.

[15] L. Tarveen and W. Hill. Beyond recommender systems: helping people help each other. In *HCI in the millennium.* Addison-Wesley, 2001.

[16] P. Turney and M. Littman. Measuring praise and criticism: Inference of semantic orientation from association. In *ACM Transactionon Information Systems*. TOIS, 2003.

[17] Y. Wang, J. Hodges, and B. Tang. Classification of web documents using a naive bayes method. In *ICTAI Proceeding of the 15th IEEE International Conference on Tool with Artificial Intelligence*. IEEE Computer Society, 2003.

[18] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. In *Computational Linguistics*. MIT Press, 2004.