Opinion Detection by Combining Machine Learning & Linguistic Tools

Olena ZUBARYEVA
University of Neuchâtel
rue Emile-Argand 11
2009 Neuchâtel
+41 32 718 2741
Olena.Zubaryeva@unine.ch

Jacques SAVOY
University of Neuchâtel
rue Emile-Argand 11
2009 Neuchâtel
+41 32 718 1375
Jacques.Savoy@unine.ch

ABSTRACT
This paper presents our work in the Multilingual Opinion Analysis Task (MOAT) done during the NTCIR-8 evaluation campaign. We suggested a probabilistic model derived from Muller's method [1] that allows us to determine and weight terms (isolated words, bigram of words, noun phrases, etc.) belonging to a given category (or subset of the corpus) compared to the rest of the corpus. Based on these terms and their weights, we have adopted the logistic regression method in order to define the most probable category for each input sentence. Our participation was strongly motivated by the objective to suggest an approach on the polarity subtask of the MOAT with a minimal linguistic component with a possibility to have its performance improved by natural language specific tools. Thus, for the English language, we have adopted a combination of both machine learning approach (Z score and logistic regression) and a polarity dictionary (linguistic component). For the traditional Chinese and Japanese languages however, our current system is limited to a machine learning scheme.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval–Information Filtering.

General Terms
Algorithms, Experimentation, Human Factors, Languages, Theory.

Keywords
Opinion detection, polarity classification, word distribution, opinionated IR, logistic regression.

1. INTRODUCTION
Users searching information on the Web can look for either factual or opinionated documents, e.g., review on a given movie or book. Given the power and effectiveness of the search engines, users can easily search for factual information. The retrieval of opinionated documents is by far more complex, since it requires not only the retrieval of the pertinent information items, but the detection and classification of the opinion in them. Most current search engines work with factual information. However, with the growth of the user activity of adding content in blogs, online forums, Internet platforms, etc., the task of opinion detection has become popular in the research community. With the multitude of languages represented on the web, it is important to develop a system that would be adaptable for different languages to detect opinionated documents on the one hand, and on the other to be able to detect their polarity (positive, negative or neutral). This task is important in many areas of Natural Language Processing (NLP) [2], [3], [4], [5] from consumer information, product reviews or question/answering (Q/A).

The NTCIR-8 MOAT (Multilingual Opinion Analysis Task) defined sentences as the information items. This year there are 6 subtasks in total: five conventional subtasks including the opinion detection, relevance of the sentence, if the opinion was detected its opinion holder/holders and target, and polarity; and cross-lingual subtask [6]. We consider the latter subtasks to be more challenging and requiring more sophisticated NLP tools depending more heavily on the underlying natural language. We participated in the conventional subtasks of the detection of the opinion, relevance and polarity. Since it is our second time participating in MOAT, based on the results of the last year we set a clear goal to improve the performance of our system on the principal subtask of detecting the opinion, focusing at the second step on the polarity of the sentence. The second goal remained the same – adaptation and testing of the approach on different languages. We want to promote an effective search system in which the linguistic component could be both clearly identified. Thus, to achieve this goal, we have participated in the English, traditional Chinese and Japanese language tracks.

The remainder of the paper is organized in the following way. Section 2 presents related work while Section 3 exposes our approach to determine the opinion, relevance and polarity of the sentences. We present the results and their evaluation in Section 4. Finally, the future work and conclusions are given in Sections 5 and 6 respectively.

2. RELATED WORK
We approach opinion detection task as a classification task with two classes initially: opinionated and factual information. For this purpose we designed an automatic retrieval and classification scheme that will be able to first to retrieve short information items (e.g., sentences, short paragraphs) according to a submitted query. In the second stage, the system classifies them according to their opinionated content as factual (no opinion), positive, negative and neutral (presenting mixed opinions). The focus in our participation in the NTCIR-8 was to propose a general approach that can be easily deployed for different natural languages.

We must first recognize that classifying short information items into positive, negative and neutral opinion categories is a difficult task, due to the fact that the semantic differences between the category neutral and the two others could be small leading to complex problems when designing and implementing an effective discrimination function. Moreover, the distinction between positive or negative could be denoted by a small element in the underlying text (e.g., a simple "not"). Finally, the distinction between neutral and either positive or negative could sometimes be questionable for a human being, as well as evaluating whether or not a given sentence (or short paragraph) conveys an opinion is not.

When viewing an opinion-finding task as a classification task (after retrieving the relevant items), it is usually considered a supervised learning problem where a statistical model performs a learning task by analyzing a pool of labeled documents. Two questions must be solved [7], namely defining an effective classification algorithm [8], and determining pertinent features that might effectively discriminate between opinionated and factual sentences / paragraphs.

From this perspective, during the two last TREC opinion-finding tasks [9], [10] and last NTCIR workshops [11], [12], a series of suggestions surfaced. Based on the English grammar, Levin defined different verb categories (characterize, declare, conjecture, admire, judge, assess, say, complain, advise) and their features (a verb corresponding to a given category occurring in the analyzed information item) that may be pertinent as a classification feature [13] (another example is given in [14]). However, words such as these cannot always work correctly as clues, for example with the word "said" in the two sentences "The iPhone price is expensive, said Ann" and "The iPhone price is 600 $, said Ann." Both sentences contain the clue word "said" but only the first one contains an opinion on the target product.

We might also mention OpinionFinder [15], a more complex system that performs subjectivity analyses to identify opinions as well as sentiments and other private states (speculations, dreams, etc.). This system is based on various classical computational linguistics components (tokenization, part-of-speech (POS) tagging [16], [17] as well as classification tools. For example, a naive Bayes classifier [8] is used to distinguish between subjective and objective sentences. A rule-based system is included to identify both speech events ("said," "according to") and direct subjective expressions ("is happy," "fears") within a given sentence. Of course such learning system requires both a training set and a deeper knowledge of a given natural language (morphological components, syntactic analyses, semantic thesauri).

The lack of enough training data for all these learning-based sub-systems is clearly a drawback, although not all groups participating in the pilot NTCIR-6 opinion analysis task encountered this same problem. Moreover, it is difficult to objectively establish when a complex learning system has enough training data (and to objectively measure the amount of training data needed in a complex ML model).

3. OUR OPINION-DETECTION APPROACH

Our system is based on two components, namely the extraction of useful features (isolated words in this study) to allow an effective classification, and second a classification scheme [8]. Our system uses word forms (tokens) to perform sentence identification within the two classes. As shown by Kilgarriff [19], the selection of words (or in general features) in an effort to characterize a particular category compared to another one is a difficult task, in which various statistical measures [20], [21], [4] have been analyzed and criticized. The selection and weighting of words is explained in Section 3.1 while Section 3.2 exposes the main aspects of our classification scheme based on logistic regression [22].

3.1 Features Extraction

In order to determine the features that can help distinguishing between factual and opinionated documents in one hand, and on the other between the polarities of the sentences, we have selected the tokens. The goal is therefore to design a method capable of selecting terms that clearly belong to one type of polarity compared to the other possibilities. Various authors have suggested formulas that could meet this objective under the condition that we use words and their frequencies or distributions [4], [19], [20], [21]. These suggested approaches are usually based on a contingency table (see Table 1).

Table 1. Example of a contingency table

	S	**C-**	**C = S ∪ C-**
ω	*a*	*b*	*a+b*
not ω	*c*	*d*	*c+d*
	a+c	*b+d*	*n=a+b+c+d*

In this table, the letter *a* represents the number of occurrences (tokens) of the word ω in the document set S (corresponding to a subset of the larger corpus C). The letter *b* denotes the number of tokens of the same word ω in the rest of the corpus (denoted C-) while *a+b* is the total number of occurrences in the entire corpus (denoted C). Similarly, *a+c* indicates the total number of tokens in S. The entire corpus C corresponds to the union of the subset S and C- (C = S∪C-) that contains *n* tokens (*n* = *a+b+c+d*).

Based on the MLE (Maximum Likelihood Estimation) principle the values shown in a contingency table could be used to estimate various probabilities. For example we might calculate the probability of the occurrence of the word ω in the entire corpus C as Prob(ω) = (*a+b*)/*n* or the probability of finding in C a word belonging to the set S as Prob(S) = (*a+c*)/*n*.

Now to define the discrimination power a term ω, we suggest deriving a weight attached to it according to Muller's method [1], [23]. We assume that the distribution of the number of tokens of the word ω follows a binomial distribution [24] with the parameters *p* and *n′*. The parameter *p* represented the probability of drawing the word ω (or Prob(ω)) and could be estimated as (*a+b*)/*n*. If we repeat this drawing *n′* = *a+c* times, we will have an expected number of occurrences of the word ω included in the subset S as Prob(ω)·*n′*. On the other hand, Table 1 gives also the number of observed occurrence of the word ω in S, and this value is denoted by *a*. A large difference between *a* and the product Prob(ω)·*n′* is clearly an indication that the presence of *a* occurrences of the term ω is not due by chance but corresponds to an intrinsic characteristic of the set S compared to the set C-.

In order to obtain a clear rule, we suggest computing the Z score attached to each word ω. If the mean of a binomial distribution is Prob(ω)·n', its variance is n'·Prob(ω)·(1-Prob(ω)). These two elements are needed to compute the standard score as described in Equation 1.

$$Z\ \ score(\omega) = \frac{a - n' \cdot \text{Prob}(\omega)}{\sqrt{n' \cdot \text{Prob}(\omega) \cdot (1 - \text{Prob}(\omega))}} \quad (1)$$

Using the MOAT-NTCIR 6 English corpus as an example, Table 2 indicates that the word "said" occurs 561 in opinionated sentences and 241 in the rest of the corpus composed of factual sentences (for a total of 802 tokens). The opinionated part contains 69,885 tokens, representing around 55.8% of the total number (125,226 tokens). Clearly, we encountered more often the word "said" in the opinionated sentences (561 times) that the simple proportion (441 = 55% of 802). The Z score for this term is equal to 5.34, indicating clearly an overuse of this term in the opinionated sentences.

Table 2. Example with the word "said" in the opinionated and the whole English corpus

	opinionated	**rest**	
"said"	561	241	802
- "said"	69,324	55,100	124,424
	69,885	55,341	125,226

As a decision rule we consider the words having a Z score between -δ and δ as terms belonging to a common vocabulary, as compared to the reference corpus (as for example "will," "with," "many," "friend," or "forced" in our example). This threshold may vary from one application to another, and we used δ = 1 in a related study [23] while 2 was used in our previous participation in NTCIR-7 [25]. A word having a Z score > δ would be considered as overused (e.g., "that," "should," "must," "not," or "government" in MOAT-NTCIR 6 English corpus), while a Z score < -δ would be interpreted as an underused term (e.g., "police," "cell," "year," "died," or "according"). In the current study, we have fixed δ = 2 because it corresponds to the limit of the standard normal distribution, allowing us to only find 5% of the observations (around 2.5% less than -2 and 2.5% greater than 2). As shown in Figure 1, the difference between our arbitrary limit of 2 (drawn in solid line) and the limits delimiting the 2.5% of the observations (dotted line) are rather close.

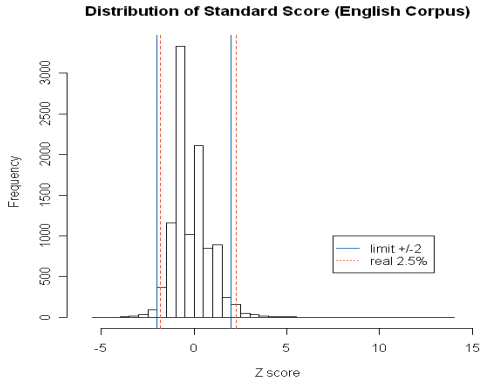

Figure 1. Distribution of the Z score (MOAT-NTCIR 6 English corpus, opinionated).

Based on a training sample, we were able to compute the Z score for different words and retain only those having a large or small Z score value. Such a procedure is repeated for all classification categories (e.g., positive, negative and neutral in the current context). It is worth mentioning that such a general scheme may work with isolated words (as applied here) or *n*-grams (that could be a sequence of either characters or words), punctuations or other symbols (numbers, dollar signs), syntactic patterns (e.g., verb-adjective) or other features (presence of proper names, hyperlinks, etc.).

3.2 Our classification Model

When our system needs to determine the polarity of a sentence, we first represent this sentence as a sequence of words (after stopword removal and applying a light stemmer for the English language [26]). For each word, we can then retrieve the Z scores for each category. If all Z scores for all words are judged as belonging to the general vocabulary, our classification procedure selects the default category. If not, we may increase the weight associated with the corresponding category (e.g., for the positive class if the underlying term is overused in this category).

Such a simple additive process could be viewed as a first classification scheme, selecting the class having the highest score after enumerating all words occurring in a sentence. For this model, we can define three variables, namely *SumPos* indicating the sum of the Z score of terms overused in positive class (i.e., Z score > 2) and appearing in the input sentence. Similarly, we can define *SumNeg*, and *SumNeutral* for the other two classes.

As additional explanatory variables, we may also take account of the number of terms that tends to be overused in positive opinionated sentences (i.e. Z score > 2), a variable called *#PosOver*. Inversely, we may count the number of terms that are underused in positive opinionated sentence (*#PosUnder*). Similarly, we can define the variables *#NegOver, #NegUnder, #NeuOver, #NeuUnder,* but for their respective categories, namely negative opinionated sentences and neutral.

From these statistics we may estimate a polarity score for each sentence according to the following formulae:

$$Pos\_score = \frac{\#PosOver}{\#PosOver + \#PosUnder}$$
$$Neg\_score = \frac{\#NegOver}{\#NegOver + \#NegUnder} \quad (2)$$
$$Neutral\_score = \frac{\#NeuOver}{\#NeuOver + \#NeuUnder}$$

Such scores are not directly related to a probability and to obtain such an estimate, we can use the logistic regression method. In this case, we obtain a probability denoted π(x) given by a set of explanatory variables using the following estimate:

$$\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}} \quad (3)$$

where β_i are the coefficients obtained from the fitting and x_i are the variables, and k is the number of variables. These coefficients reflect the relative importance of each explanatory variable in the final score.

For each sentence, we can compute the π(x) corresponding to the three possible categories and the final decision is simply to classify the sentence according to the max π(x) value. This approach takes account of the fact that some explanatory variables may have more importance (according to our training set) than other in assigning the correct category. However, we must recognize that the length of the underlying sentence is not directly taken into account in this first model. Our underling assumption is that all sentences have a similar number of indexing tokens.

4. EVALUATION

In order to evaluate the capability of an automatic system to retrieve and classify correctly different information items, we may impose that the answers are a ranked list and then evaluate the system's performance according to classical IR measures such as MAP. This approach was adopted during the last Blog tracks at TREC [9], [10]. As another approach we may evaluate the classification performance based on a set-based approach, judging the system's capability to identify the different categories. The traditional evaluation measures based on sets (precision, recall, F-measure) can then be applied. This choice was made for the NTCIR workshops [11], [12] and explained in the current workshop [6].

On the other hand, we have assumed until now that words can be extracted from a sentence in order to define the needed features used to determine if the underlying information item conveys an opinion or not. Working with the Japanese or Chinese languages this assumption does no longer hold and we need to determine indexing units by either applying an automating segmentation approach (based either on a morphological (e.g., CSeg&Tag) or a statistical method [27] or considering *n*-gram indexing approach (unigram, bigram or both unigram and bigram). Finally we may also consider a combination of both *n*-gram and word-based indexing strategies [27], [28].

4.1 Traditional Chinese Language

We participated in the traditional Chinese language task and were able to submit one run based on our first classification model. Based on our past IR experiments [28], we have selected a combined unigram & bigram indexing scheme for each sentence in this language.

Table 3. MOAT evaluation for the traditional Chinese opinion analysis

Subtask	Precision	Recall	F-measure
Relevance	86.2	48.25	61.87
Opinion	52.37	48.47	50.34
Polarity	47.01	23.27	31.13

In order to classify each input sentence, we have used our logistic model based on Equations 2 and 3. The result obtained with our official run is depicted in Table 3. In this case, we used only a learning scheme (logistic regression) without any additional linguistic information about this language. Furthermore, we mainly focus our effort on detecting opinions and not really on the polarity of them.

4.2 Japanese Language

With the Japanese language we submitted a single run based, as for the Chinese language, on our first classification model. Based on our past experiment [28], we have selected a bigram indexing scheme as a way to represent each sentence for this language.

Table 4. MOAT evaluation for the Japanese opinion analysis

Subtask	Precision	Recall	F-measure
Relevance	48.18	28.61	35.9
Opinion	63.3	28.56	39.36
Polarity	42.8	8.95	14.8

In order to determine the category of each input sentence, we have used our logistic model (see Eq. 2 and 3) without any additional linguistic knowledge about the Japanese language. The result obtained with our official run is depicted in Table 4.

4.3 English Language

For the English language task we were able to send two runs. The second run is based on the same classification model used for both the traditional Chinese and Japanese languages. As features for this run, we used isolated words after elimination of stopwords (e.g., “the,” “was,” or “in”), and applying a light stemmer (to remove the final '-s') [26]. As an additional feature, we have used the bigram of words (e.g., “North Korea,” “health care”) to hopefully improve the representation of each sentence. The classification scheme thus takes account for both the isolated word and also for the bigrams of words.

In order to improve the system performance, we did the classification in two stages. First, the system classified the sentences in two categories: opinionated and factual. Then, within the sentences classified as opinionated, the polarity detection was performed in a second stage.

In summary, this second run is based only on statistical features (words & bigrams) without considering any additional semantically-related tool.

On the other hand, the first run is therefore more complex and will include a linguistic component. In fact, the suggested statistical approach behind the second run is not flawless when applied to the natural language with its many exceptions and ambiguities. In order to minimize these errors, we explored the use of specific natural processing tools. For this purpose, SentiWordNet [29] give us a polarity score for each word in the English language. Our idea is thus to combine those scores with the system's scores for each sentence.

Using the scores given by the SentiWordNet [29] dictionary, we sum those scores if the word belongs to the opinionated category in the underlying sentence. The not opinionated score of a given sentence is computed in the same way with the difference that it is divided by the number of words in the sentence. Thus, if opinionated score is more than not opinionated one, there is an opinion, otherwise not. As an example, let's take an opinionated sentence with negative polarity from the NTCIR-7 campaign: "With Tokyo's economy declining about 3 percent this year, this seems unlikely."

Table 5. SentiWordNet positive and negative scores for each word in the example sentence

#	Word	SentiWordNet
		PosScore / NegScore
1	Tokyo	0.0 / 0.0
2	economy	0.125 / 0.25
3	declining	0.0 / 0.0
4	about	0.375 / 0.0
5	percent	0.125 / 0.0
6	this	0.0 / 0.0
7	year	0.0 / 0.0
8	seem	0.0 / 0.0
9	unlikely	0.0 / 0.625

The values *PosScore* and *NegScore* are the positivity and negativity scores assigned by the SentiWordNet [29]. The objectivity score is obtained in the following way: *ObjScore* = 1 – (*PosScore* + *NegScore*) [29]. Looking at the scores for the individual tokens in the example sentence the SentiWordNet opinionated score will be the sum of *PosScores* and *NegScores* for each token. The not opinionated score will be a sum of objectivity score for each token, divided by the number of words in the sentence. Thus, we have an opinionated score of 0.125 + 0.25 + 0.375 + 0.125 + 0.625 = 1.5 and not opinionated score of (1 + 0.625 + 1 + 0.625 + 0.875 + 1 + 1 + 1 + 0.375)/9 = 0.833 for our example sentence. This technique is favoring the opinionated score and is a heuristic approach that intuitively takes account of the rationalization that there are more not opinionated words than opinionated in the sentence. The presence of opinionated word weighs more than the presence of the not opinionated one. This approach seems to give good results in practice. Finally, the opinionated and not opinionated scores obtained from the SentiWordNet [29] are normalized and summed with our system's opinionated and not opinionated scores for the sentence.

For the calculation of the polarity score in Run 1, if our system classified the sentence as not opinionated but with the addition of the SentiWordNet [29] the sentence is classified as opinionated, we take the category with the highest sum of the Z scores for polarity attribution.

Table 6. MOAT evaluation for the three models used with the English corpus

Subtask	Runs	Prec.	Recall	F-measure
Relevance	Run 1	83.68	32.74	47.07
	Run 2	84.39	36.01	50.48
Opinion	Run 1	29.44	62.84	40.1
	Run 2	19.32	81.79	31.26
Polarity	Run 1	50.29	29.58	37.25
	Run 2	48.35	37.8	42.43

The results given in the Table 6 show, that we quite improved the precision for the relevance detection subtask. The Run 2 gives low precision for the opinion subtask with however a high recall value. Thus, in comparison to Run 1, we can see that the use of the SentiWordNet [29] improved precision but lowered recall, nevertheless, allowing us to achieve a quite high F-measure in comparison to other teams. Overall, it is possible to see a general improvement in relation to the previous year, even though this time the language specific techniques, like query expansion, were not used. It seems that with the growth of training data (NTCIR-6 and NTCIR-7 corpora) as well as the use of two-step classification with the bigram of words improves the system's performance. In order to evaluate some of the reasons of failure when doing opinion classification, we looked closely and analyzed the system's decision on a sample example from the NTCIR collection in the next section.

4.3.1 Failure Analysis

Several experiments were conducted on the NTCIR-6 and NTCIR-7 MOAT corpora that help to clarify some reasons why our method fails to make correct classification. As one of the corpora peculiarities pertinent to our classification system's performance, we determined that a great number of words occur 1 to 4 times in the collection. With such low frequencies of occurrence, they do not carry reliable information to help the classification procedure. As an example let's take the following neutral in polarity sentence: "Half of the job is psychiatry." If we eliminate the stop words, we end up with three words: "half", "job" and "psychiatry". The term "psychiatry" is a *hapax* term, meaning that it occurs only once in the collection, therefore we have no Z score for it. For the other two terms we have the following scores: "half" with -1.83 and "job" with 0.16. The Z score for the term "half" shows us that this term is overused in the not opinionated part of the corpora. It's absolute value being bigger than the Z score of the term "job", the system will classify the sentence in not opinionated category. As you can see, due to low frequencies of occurrences of lots of term in the collection, when calculating the Z score for the sentence, we can end up in a situation where we have score only for several terms, even in long sentences.

5. FUTURE WORK

Our system is based on the statistical method (Z score) to identify those terms that adequately characterize subsets of the corpus belonging to positive, negative, neutral or non-opinionated subsets. In this selection, we focused only on the statistical aspect (distribution difference) of words and bigrams of words for English, bigrams of ideograms for the Japanese, or both unigram and bigram of ideograms for the traditional Chinese language. As it is demonstrated in the English subtask, the use of the language specific techniques (polarity dictionary in our case) may help the detection of opinion in a sentence. In the future we intend to explore natural language specific tools for Japanese and Chinese.

In further research we could also consider punctuations (e.g., quotation marks (“”), question marks (?), exclamation points (!), etc.) as well as other symbols (e.g. $, mm, mainly associated with facts) to distinguish between factual and opinionated documents. The most useful terms would also then be added to the query to improve the rank of opinionated documents. As another approach, we could use the evaluation of co-occurrence terms of pronouns “I” and “you” mainly with verbs (e.g., “believe,” “feel,” “think,” “hate”) in order to boost the rank of retrieved items.

Other indicators, such as the presence /occurrence of proper names and their frequency or distribution might help us classify a document as being opinionated or not. The presence /occurrence of adjectives and adverbs, together with their superlative (e.g., best, most) or comparative (e.g., greater, more) forms could also be useful hints regarding the presence of opinionated versus factual information.

6. CONCLUSION

In our second participation in a MOAT task, we have suggested a general method to define and weight isolated words in order to build a set of useful features able to classify sentences into different categories. Our classification scheme is based on the Z score [23] for terms (words, bigrams of words for the English languages, unigram and bigrams for the Chinese or Japanese language) and a logistic regression method [22]. The goal is to build a statistical, language-independent tool for opinion polarity detection that could be later enhanced with the use of specifically adapted to the particular natural language text processing tools. Given the results and performance of our system last year we tried to improve the procedure and we can see that there is a clear improvement in the precision measure. This is probably due to the two-step classification approach that we applied to differentiate between opinionated and factual, and then within the polarities, as well as the use of the SentiWordNet [29].

We are quite content with our performance for the opinion subtask in English, compared to other results. However, there is still a great work to be done to improve the performance. It is important to note that this year the training set included the collection from previous NTCIR-6 and NTCIR-7 MOAT corpora, therefore, also influencing the system performance. Given the failure analysis conducted for the English corpus, we arrived at some conclusions that could improve and fasten the system performance. Namely, we could eliminate *hapax* terms, words with low frequency.

The performance of the suggested model may hopefully be enhanced with the use of language-independent (e.g., noun phrase, punctuation, word categories) and natural language specific tools (SentiWordNet [29], list of vocabularies and expressions, etc.).

7. ACKNOWLEDGMENTS

The authors would like to thank the NTCIR-8 task organizers for their efforts in developing test-corpus in opinionated IR. This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

8. REFERENCES

[1] Muller, C. 1992. Principe et méthodes de statistique lexicale. Champion, Paris.

[2] Mitkov, R. (Ed.) 2003. *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford.

[3] Nugues, P. M. 2006. *An Introduction to Language Processing with Perl and Prolog*. Springer- Verlag, Berlin.

[4] Manning, C. D., Schütze, H. 2000. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

[5] Indurkhya N., Damereau F.J., (Eds). 2010. *Handbook of Natural Language Processing*. 2nd Ed., Chapman & Hall/CRC, Boca Raton.

[6] Seki, Y., Ku, L.W., Sun, L., Chen, H.H., & Kando, N. 2010. Overview of multilingual opinion analysis task at NTCIR-8. *Proceedings NTCIR-8*, NII publication (National Institute of Informatics), Tokyo, 2010, to appear.

[7] Sebastiani, F. 2002. Machine learning in automatic text categorization. *ACM Computing Survey*, 14(1), 1-27.

[8] Witten, I.A., Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Ed., Morgan Kaufmann, San Francisco (CA).

[9] Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., Soboroff, I. 2007. Overview of the TREC-2006 blog track. *Proceedings TREC-2006*, NIST Publication #500-272, 17-32.

[10] Macdonald, C., Ounis, I., Soboroff, I. 2008. Overview of the TREC-2007 blog track. *Proceedings TREC-2007*, NIST Publication #500-274, 1-13.

[11] Seki, Y., Evans, D.K., Ku, L.W., Chen, H.H., Kando, N., & Lin, C.Y. 2007. Overview of opinion analysis pilot task at NTCIR-6. *Proceedings NTCIR-6*, NII publication (National Institute of Informatics), 265-278.

[12] Seki, Y., Evans, D.K., Ku, L-W., Sun, L., Chen, H.-H. & Noriko, K. 2008. Overview of multilingual opinion analysis task at NTCIR-7. *Proceedings NTCIR-7*, NII publication (National Institute of Informatics), 185-203.

[13] Bloom, K., Stein, S., Argamon, S. 2007. Appraisal extraction for news opinion analysis at NTCIR-6. *Proceedings NTCIR-6*, NII publication (National Institute of Informatics), 279-289.

[14] Sokolova, M. & Lapalme, G. 2008. Verbs speak loud: Verb categories in learning polarity and strength of opinions. *Proceedings of the 20th Canadian Conference on Artificial Intelligence*, 320–331.

[15] Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S. 2005. OpinionFinder: A system for subjectivity analysis. *Proceedings HLT/EMNLP*, Vancouver (BC), 34-35.

[16] Marcus, M.P., Santorini, B., Marcinkiewicz, M.A. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313-330.

[17] Toutanova, K., Manning, C. 2000. Enriching the knowledge sources used in a maximum entropy Part-of-Speech tagging. *Proceedings EMNLP / VLC-2000*, 63-70.

[18] Gandrabur, S. Foster G. & Lapalme, G. 2006. Confidence estimation for NLP applications. *ACM Transactions on Speech and Language Processing*, 3(3):1–29.

[19] Kilgarriff, A. 2001. Comparing corpora. International *Journal of Corpus Linguistics*, 6(1):97-133.

[20] Church, K.W., Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22-29.

[21] Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61-74.

[22] Hosmer D., Leneshow S., 2000. *Applied Logistic Regression*. Wiley Interscience, New York.

[23] Savoy, J. 2010. Lexical analysis of US political speeches. *Journal of Quantitative Linguistics*, to appear.

[24] Baayen, H.R. 2001. Word Frequency Distributions. Kluwer Academic Press, Dordrecht, NL.

[25] Zubaryeva, O., Savoy, J. 2008. Opinion and polarity detection within Far-East languages in NTCIR-7. *Proceedings NTCIR-7*, NII publication (National Institute of Informatics), Tokyo, 2008, 318-323.

[26] Harman, D. 1991. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), 7-15.

[27] Murata, M., Ma, Q., Isahara, H. 2003. Applying multiple characteristics and techniques to obtain high levels of performance in information retrieval. *Proceedings of NTCIR-3*, NII publication (National Institute of Informatics).

[28] Savoy, J. 2005. Comparative study of monolingual and multilingual search models for use with Asian languages. *ACM Transactions on Asian Languages Information Processing*, 4:163-189.

[29] Esuli, A. & Sebastiani, F. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. *Proceedings LREC-06*, Lisbon, 417-422.