Investigation in Statistical Language-Independent Approaches for Opinion Detection in English, Chinese and Japanese

Olena Zubaryeva
Institute of Informatics
University of Neuchâtel
Emile-Argand, 11, 2009 Switzerland
`olena.zubaryeva@unine.ch`

Jacques Savoy
Institute of Informatics
University of Neuchâtel
Emile-Argand, 11, 2009 Switzerland
`jacques.savoy@unine.ch`

Abstract

In this paper we present a new statistical approach to opinion detection and its' evaluation on the English, Chinese and Japanese corpora. Besides, the proposed method is compared with three baselines, namely Naïve Bayes classifier, a language model and an approach based on significant collocations. These models being language independent are improved with the use of language-dependent technique on the example of the English corpus. We show that our method almost always gives better performance compared to the considered baselines.

1 Introduction

The task of opinion mining has received attention from the research community and industry lately. The main reasons for extensive research in the area are the growth of user needs and companies' desire to analyze and exploit the user-generated content on the Web in the form of blogs and discussions. Thus, users want to search for opinions on various topics from products that they want to buy to opinions about events and well-known persons. A lot of businesses are interested in how their services are perceived by their customers. Therefore, the detection of subjectivity in the searched information may add the additional value to the interpretation of the results and their relevancy to the searched topic. The growth of user activity on the Web gives substantial amounts of data for these purposes.

In the context of globalization the possibility to provide search of opinionated information in different natural languages might be of great interest to organizations and communities around the world. Our goal is to design a fully automatic system capable of working in a language-independent manner. In order to compare our approach on different languages we chose English, traditional Chinese and Japanese corpora. As a further possibility to improve the effectiveness of the language independent methods we also consider the additional application of language dependent techniques specific to the particular natural language.

The related work in opinion detection is presented in Section 2. We describe our approach in detail with the three other baselines in Section 3. The fourth section describes language specific approach used for the English corpus. In Section 5 we present the evaluation of the three models using the NTCIR-6 and NTCIR-7 MOAT English, Chinese and Japanese test collections (Seki *et al.*, 2008). The main findings of our study and future research possibilities are discussed in the last sections.

2 Related Work

The focus of our work is to propose a general approach that can be easily deployed for different natural languages. This task of opinion detection is important in many areas of NLP such as question/answering, information retrieval, docu-

ment classification and summarization, and information filtering. There are numerous challenges when trying to solve the task of opinion detection. Some of them include the fact that the distinction between opinionated and factual could be denoted by a single word in the underlying text (e.g., "The iPhone price is $600." vs. "The iPhone price is high."). Most importantly evaluating whether or not a given sentence conveys an opinion could be questionable when judged by different people. Further, the opinion classification can be done on different levels, from documents to clauses in the sentence.

We consider the opinion detection task on a sentence level. After retrieving the relevant sentences using any IR system we automatically classify a sentence according to two classes: opinionated and not opinionated (factual). When viewing an opinion-finding task as a classification task, it is usually considered as a supervised learning problem where a statistical model performs a learning task by analyzing a pool of labeled sentences. Two questions must therefore be solved, namely defining an effective classification algorithm and determining pertinent features that might effectively discriminate between opinionated and factual sentences. From this perspective, during the last TREC opinion-finding task (Macdonald *et al.*, 2008) and the last NTCIR-7 workshop (Seki *et al.*, 2008), a series of suggestions surfaced.

As the language-dependent approach various teams proposed using Levin defined verb categories (namely, 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 (Bloom *et al.*, 2007). However, words such as these cannot always work correctly as clues, for example with the word "said" in the two sentences "There were crowds and crowds of people at the concert, said Ann" and "There were more than 10,000 people at the concert, said Ann." Both sentences contain the clue word "said" but only the first one contains an opinion on the target product. Turney (2002) suggested comparing the frequency of phrase co-occurrences with words predetermined by the sentiment lexicon. Specific to the opinion detection in Chinese language Ku *et al.* (2006) propose a dictionary-based approach for extraction and summarization. For the Japanese language in the last NTCIR-6 and NTCIR-7 workshops the opinion finding methods included the use of supervised machine learning approaches with specific selection of certain parts-of-speech (POS) and sentence parts in the form of *n*-gram features to improve performance.

There has been a trend in applying language models for opinion detection task (Lavrenko, Croft, 2001). Pang & Lee (2004) propose the use of language models for sentiment analysis task and subjectivity extraction. Usually, language models are trained on the labeled data and as an output they give probabilities of classified tokens belonging to the class. Eguchi & Lavrenko (2006) propose the use of probabilistic language models for ranking the results not only by sentiment but also by the topic relevancy.

As an alternative other teams during the last TREC and NTCIR evaluation campaigns have suggested variations of Naïve Bayes classifier, language models and SVM, along with the use of such heuristics as word order, punctuation, sentence length, etc.

We might also mention OpinionFinder (Wilson *et al.*, 2005), 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 (Toutanova & Manning, 2000) as well as classification tools. For example, a Naïve Bayes classifier (Witten & Frank, 2005) 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 thesaurus).

The lack of enough training data for the learning-based systems is clearly a drawback. 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 Language Independent Approaches

In this section we propose our statistical approach for opinion detection as well as the description of the Naïve Bayes and language model (LM) baselines.

3.1 Logistic Model

Our system is based on two components: the extraction and weighting of useful features (limited to isolated words in this study) to allow an effective classification, and a classification scheme. First, we present the feature extraction approach in the Section 3.1.1. Next, we discuss our classification model. Sections 3.2 and 3.3 describe the chosen baselines.

3.1.1 Features Extraction

In order to determine the features that can help distinguishing between factual and opinionated documents, we have selected the tokens. As shown by Kilgarriff (2001), the selection of words (or in general features) in an effort to characterize a particular category is a difficult task. The goal is therefore to design a method capable of selecting terms that clearly belong to one of the classes. The approaches that use words and their frequencies or distributions are usually based on a contingency table (see Table 1).

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

Table 1. Example of a contingency table.

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 in the current study). 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 with C=C-∪S). 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- 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 $Pr(\omega) = (a+b)/n$ or the probability of finding in C a word belonging to the set S as $Pr(S) = (a+c)/n$.

Now to define the discrimination power a term ω, we suggest defining a weight attached to it according to Muller's method (Muller, 1992). We assume that the distribution of the number of tokens of the word ω follows a binomial distribution with the parameters p and n'. The parameter p represented the probability of drawing a word ω also denoted in the corpus C (or $Pr(\omega)$) and could be estimated as $(a+b)/n$. If we repeat this drawing $n' = a+c$ times, we will have an estimate of the number of word ω included in the subset S by $Pr(\omega)·n'$. On the other hand, Table 1 gives also the number of observations of the word ω in S, and this value is denoted by a. A large difference between a and the product $Pr(\omega)·n'$ 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 subset S compared to the subset 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 $Pr(\omega)·n'$, its variance is $n'·Pr(\omega)·(1-Pr(\omega))$. These two elements are needed to compute the standard score as described in Equation 1.

$$Zscore(\omega) = \frac{a - n`·Pr(\omega)}{\sqrt{n`·Pr(\omega)·(1-Pr(\omega))}} \quad (1)$$

As a decision rule we consider the words having a Z score between -2 and 2 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 was chosen arbitrary. A word having a Z score > 2 would be considered as overused (e.g., "that," "should," "must," "not," or "government" in MOAT NTCIR-6 English corpus), while a Z score < -2 would be interpreted as an underused term (e.g., "police," "cell," "year," "died," or "according"). The arbitrary threshold limit of 2 corresponds to the limit of the standard normal distribution, allowing us to find around 5% of the observa-

tions (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.
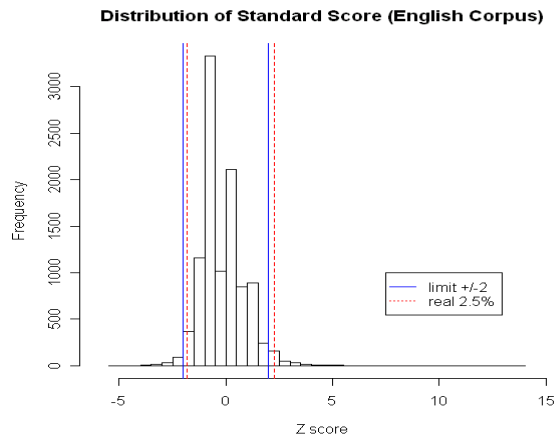

Figure 1. Distribution of the Z score (MOAT NTCIR-6 English corpus, opinionated).

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 (opinionated and factual). It is worth mentioning that such a general scheme may work with isolated words (as applied here) or *n*-gram (that could be a sequence of either characters or words), as well as with punctuations or other symbols (numbers, dollar signs), syntactic patterns (e.g., verb-adjective in comparative or superlative forms) or other features (presence of proper names, hyperlinks, etc.)

3.1.2 Classification Model

When our system needs to determine the opinionatedness of a sentence, we first represent this sentence as a set of words. 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 opinionated 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. This approach assumes that the word order does not have any impact. We also assume that each sentence has a similar length.

For this model, we can define two variables, namely *SumOP* indicating the sum of the Z score of terms overused in opinionated class (i.e. Z score > 2) and appearing in the input sentence. Similarly, we can define *SumNOOP* for the other class. However, a large *SumOP* value can be obtained by a single word or by a set of two (or more) words. Thus, it could be useful to consider also the number of words (features) that are overused (or underused) in a sentence. Therefore, we can define *#OpOver* indicated the number of terms in the evaluated sentence that tends to be overused in opinionated documents (i.e. Z score > 2) while *#OpUnder* indicated the number of terms that tends to be underused in the class of opinionated documents (i.e. Z score < -2). Similarly, we can define the variables *#NoopOver, #NoopUnder,* but for the non-opinionated category.

With these additional explanatory variables, we can compute the corresponding subjectivity score for each sentence as follows:

$$Op\_score = \frac{\#OpOver}{\#OpOver + \#OpUnder}$$

$$Noop\_score = \frac{\#NoopOver}{\#NoopOver + \#NoopUnder} \quad (2)$$

As a better way to combine different judgments we suggest following Le Calvé & Savoy (2000) and normalize the scores using the logistic regression. The logistic transformation $\pi(x)$ given by each logistic regression model is defined as:

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

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

For each sentence, we can compute the $\pi(x)$ corresponding to the two possible categories and the final decision is simply to classify the sentence according to the max $\pi(x)$ value. This approach takes account of the fact that some explanatory variables may have more importance than other in assigning the correct category.

3.2 Naïve Bayes

For comparison with our logistic model we chose three baselines: Naïve Bayes and language model and finding significant collocations. Despite its simplicity Naïve Bayes classifier tends to perform relatively well for various text categorization problems (Witten, Frank, 2005). In accordance with our approach, we used word tokens as classification features for the English corpora. For the Chinese and Japanese languages overlapping bigram approach was used (Savoy, 2005). The training method estimates the relative frequency of the probability that the chosen feature belongs to a specific category using add-one smoothing technique.

3.3 Language Model (LM)

As a second baseline we use the classification based on the language model using overlapping n-gram sequences (n was set to 8) as suggested by Pang & Lee (2004, 2005) for the English language. Using the overlapping 4-gram sequence for the word "company", we obtain: "comp", "ompa", "mpan", etc. For the Chinese and Japanese corpora bigram approach was applied. As in Naïve Bayes, the language model gives the probability of the sentence belonging to a specific class. Working with relatively large n allows a lot of word tokens to be processed as is, at least for the English language.

3.4 Significant Collocations (SC)

Another promising approach among the supervised learning schemes is the use of collocations of two or more words or features (Manning & Schütze, 2000). This method allows classification of instances based on significant collocations learned from the labeled data. Some examples of the frequent collocations in the corpora would be "in the", "of the". The idea of the method is to find significant collocations (SC) that occur more in the opinionated corpus than in the non-opinionated one. In order to do so the model returns the collocations of two words for the English language based on the degree to which their counts in the opinionated corpus exceed their expected counts in the not opinionated one. As an example for the English opinionated corpus the following collocations were found: "are worried", "pleaded guilty", "eager to", "expressed hope". Clearly, overlooking the list of new found collocations it is possible to judge their relevancy. However, it is not clear how to use this method with the Chinese and Japanese texts, since these languages do not have white space or other usual delimiters as in English. In order to solve the problem of feature selection we chose bigram indexing on the Chinese and Japanese corpora and searched for significant new collocations of bigrams.

4   Language Dependent Approach

As the language dependent technique to improve the obtained classification results we suggest the use of SentiWordNet for the English language (Esuli & Sebastiani, 2006). Since the vocabulary of words in SentiWordNet is quite limited it is not always clear how to combine the objectivity scores.

The SentiWordNet score was computed in the following way: to define the opinionated score of the sentence the sum of scores representing that the word belongs to opinionated category for each word in the sentence is calculated. The not opinionated score of the 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. This is a heuristic approach that intuitively takes account of the rationalization that there are more not opinionated words than opinionated in the sentence. At the same time the presence of opinionated word weighs more than the presence of the not opinionated ones. Especially, this approach seems to give good result.

5   Experiments

The experiment was carried out on the NTCIR-6 and NTCIR-7 English news corpora using 10-fold cross-validation model on a lenient evaluation standard as described in Seki et al. We do not question the construction and structure of opinions in this data set, since those questions were addressed at the NTCIR workshops. Using the Chinese and Japanese corpora we can verify the quality of the suggested language-independent approaches.

5.1 Feature Selection & Evaluation in English

For the evaluation of sentences in English, the assumption of isolated words (bag-of-words) previously stemmed was used by our system. The corpora are comprised of more than 13,400 sentences, 4,859 (36.3%) of which are opinionated. As the evaluation metrics precision, recall and F_1-measure were used based on gold standard evaluation provided by NTCIR workshops (Seki *et al.*, 2008). The precision and recall are weighted equally in our experiment but it should be recognized that based on the system's needs and focus there could be more accent on precision or recall.

Model	Precision	Recall	F_1-measure
Logistic model	0.583	**0.508**	**0.543**
Naïve Bayes	0.415	0.364	0.388
LM	0.350	0.339	0.343
SC	**0.979**	0.360	0.527

Table 2. Evaluation results of 10-fold cross-validation on NTCIR-6 and NTCIR-7 English corpora.

Comparing the results in Table 2 to the baselines of the Naïve Bayes classifier and LM evaluated on the same training and testing sets, we see that logistic model outperforms the baselines. In our opinion, this is due to the use of more explanatory variables that better discriminate between opinionated and factual sentences.

The use of language dependent techniques on the other hand might further improve the results. Especially, this seems promising observing the results when using the SentiWordNet on the English corpus. In Table 3 one can see that the first three models show improvement. Specifically, the precision of the logistic model increased from 0.583 to 0.766 (by 31.4%).

Model	Precision	Recall	F_1-measure
Logistic model	0.766	**0.488**	**0.597**
Naïve Bayes	0.667	0.486	0.562
LM	0.611	0.474	0.534
SC	**0.979**	0.420	0.588

Table 3. Evaluation results of 10-fold cross-validation on NTCIR-6 and NTCIR-7 English corpora with SentiWordNet.

When considering the F_1-measure, the impact of the language-dependent approach shows 9% of improvement, from 0.543 to 0.597.

The way that we incorporated the scores provided by SentiWordNet was done with the help of linear combination and normalization of scores for each of the models.

5.2 Feature Selection & Evaluation in Chinese

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 Chinese language this assumption does no longer hold. Therefore, we need to determine indexing units by either applying an automating segmentation approach (based on either a morphological (e.g., CSeg&Tag) or a statistical method (Murata & Isahara, 2003)) or considering *n*-gram indexing approach (unigram or bigram, for example). Finally we may also consider a combination of both *n*-gram and word-based indexing strategies.

Based on the work of Savoy, 2005 we experimented with overlapping bigram and trigram indexing schemes for Chinese. The experimental results show that bigram indexing outperforms trigram on all three considered statistical methods. Therefore, as features for Chinese we used overlapping bigrams.

The NTCIR-6 and NTCIR-7 Chinese corpora consisted of more than 14,507 sentences, 9960 (68.7%) of which are opinionated. The results of all three statistical models performed on the Chinese corpora are presented in Table 4.

Model	Precision	Recall	F_1-measure
Logistic model	**0.943**	**0.730**	**0.823**
Naïve Bayes	0.729	0.538	0.619
LM	0.581	0.634	0.606
SC	0.313	0.898	0.464

Table 4. Evaluation results of 10-fold cross-validation on NTCIR-6 and NTCIR-7 Chinese corpora.

From the results in Table 4 we clearly see that our approach gives better performance and confirms the results presented in Tables 2 and 3. The significant improvement in scores could be due to the fact that Chinese corpus contains more opinionated sentences in relevance to not opinionated once. Thus, the training set for opinionated classi-

43

fication was much larger compared to the English language. This proves the relevance of more training data for the learning-based systems. But the direct comparison with the results on the English corpus is not possible.

5.3 Feature Selection & Evaluation in Japanese

As with the Chinese language we face the same challenges in feature definition for the Japanese language. After experimenting with bigram and trigram we chose bigram strategy for indexing and feature selection.

The NTCIR-6 and NTCIR-7 Japanese corpora consisted of more than 11,100 sentences with 4,622 opinionated sentences (representing 41.6% of the corpus). The results of the statistical models are shown in Table 5.

Model	Precision	Recall	F-measure
Logistic model	0.527	0.761	0.623
Naïve Bayes	0.565	0.570	0.567
LM	0.657	0.667	0.662
SC	**0.663**	**0.856**	**0.747**

Table 5. Evaluation results of 10-fold cross-validation on NTCIR-6 and NTCIR-7 Japanese corpora.

From the results we can see that the significant collocations model outperforms the others. This could be due to the fewer number of opinionated sentences compared to the Chinese or English corpora. This tends to indicate the necessity of an extensive training data for the logistic model in order to provide reliable opinion estimates.

6   Future Work and Conclusion

In this paper we presented our language-independent approach based on using Z scores and the logistic model to identify those terms that adequately characterize subsets of the corpus belonging to opinionated or non-opinionated classes. In this selection, we focused only on the statistical aspect (distribution difference) of words or bigrams. Our approach was compared to the three baselines, namely Naïve Bayes classifier, language model and an approach based on finding significant collocations. We have also demonstrated on the English corpora how we can use the language dependent techniques to identify the possibility of opinion ex-

pressed in the sentences that otherwise were classified as not opinionated by the system. The use of SentiWordNet (Esuli & Sebastiani, 2006) in combination with our methods yields better results for the English language.

This study was limited to isolated words in English corpus but in further research we could easily consider longer word sequences to include both noun and verb phrases. 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") using part of speech tagging techniques in order to boost the rank of retrieved items.

Using freely available POS taggers, we could take POS information into account (Toutanova & Mannning, 2004) and hopefully develop a better classifier. For example, the presence of proper names and their frequency or distribution might help us classify a document as being opinionated or not. The presence 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.

Acknowledgments

References

Bloom, K., Stein, S., & Argamon, S. 2007. *Appraisal extraction for news opinion analysis at NTCIR-6.* Proceedings NTCIR-6, NII, Tokyo, pp. 279-289.

Eguchi, K., Lavrenko, V. 2006. *Sentiment retrieval using generative models.* Proceedings of EMNLP, Sydney, pp. 345-354.

Esuli, A., Sebastiani, F. 2006. *SentiWordNet: A publicly available lexical resource for opinion mining.* Proceedings LREC'06, Genoa.

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

Ku, L.-W., Liang, Y.-T., Chen, H.-H. 2006. *Opinion extraction, summarization and tracking in news and blog corpora.* Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, pp. 100-107.

Lavrenko, V., Croft, W.B. 2001. *Relevance-based language models.* SIGIR, New Orleans, LA, pp. 120-127.

Le Calvé, A., Savoy, J. 2000. *Database merging strategy based on logistic regression.* Information Processing & Management, 36(3):341-359.

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

Manning, C. D., Schütze, H. 2000. *Foundations of Statistical Natural Language Processing.* MIT Press.

Muller, C. 1992. *Principes et méthodes de statistique lexicale.* Champion, Paris.

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, Tokyo.

Pang, B., Lee, L. 2004. *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.* Proceedings of ACL, Barcelona, pp. 271-278.

Pang, B., Lee, L. 2005. *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales.* In Proceedings of the Association for Computational Linguistics (ACL), pp. 115-124.

Savoy, J. 2005. *Comparative study of monolingual search models for use with asian languages.* ACM Transactions on Asian Language Information Processing, 4(2):163-189.

Seki, Y., Evans, D. K., Ku, L.-W., Sun, L., Chen, H.-H., & Kando, N. 2008. *Overview of multilingual opinion analysis task at NTCIR-7.* Proceedings NTCIR-7*, NII, Tokyo, pp. 185-203.

Toutanova, K., & Manning, C. 2000. *Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagging.* Proceedings EMNLP / VLC-2000, Hong Kong, pp. 63-70.

Turney, P. 2002. *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews.* Proceedings of the ACL, Philadelphia (PA), pp. 417-424.

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), pp. 34-35.

Witten, I.A., & Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, San Francisco (CA).