USING SENTIMENT-ANALYSIS FOR TEXT INFORMATION EXTRACTION

Solov'ev A. N. (a.solovyev@i-teco.ru),
Antonova A. Ju. (a.antonova@i-teco.ru),
Pazel'skaia A. G. (Pazelskaya@i-teco.ru)

«I-Teco», Moscow, Russia

This article aims to demonstrate relationship between sentiment-analysis and information extraction from text. To illustrate the idea we propose a sentiment-based summarization prototype instrument. The performance of the method in comparison with 3 another summarization methods is tested on a small corpus of mass media and blog posts.

Key words: sentiment-analysis, semantics, emotional component of meaning, information, automatic text summarization, information extraction

1. Introduction

Automatic document sense extraction is considered to be among crucial problems of natural language processing. Usually a document contains numerous noun phrases. The extraction of the most characteristic NPs appears to be a difficult task partly solved by statistical methods or with the help of thesauri for a certain domain ([Lukashevich 2011][1]).

Blog texts procession poses even more difficult problems at all levels: lexical, syntactic and stylistic. Commonly a blog post consists of the post itself and a hierarchy of comments. Identifying the antecedent of a comment and the part of the post commented on is often problematic even for a human. To do this, one needs first to identify the topic(s) of the post (or a parent comment), the topic of the comments and look for matches between them.

In this case semantic analysis based on tokenization followed by morphological and syntactic analysis turns to be just helpless because of lexical, syntactic and stylistity specifics of the blogosphere. Polythematics as a typical peculiarity of blogs (unlike news) adds even more confusion to an automatic processing task.

Automatic text summarization aims to extract key information from texts, therefore it can be considered as an approach to text understanding.

Currently automatic summarization widely uses approaches from information theory (e.g. self-information, entropy [Gusev et al. 2005], Bayes models, Markov

[1] Chapter 22: Socio-political thesaurus and automatic annotation, pp. 414–446

chains and nets [Fung et al, 2003] etc.). Despite of the growth of sophistication of algorithms, getting a high-quality summary of a randomly chosen text is still something to be desired. Preliminarily constructed thesaurus of a specified domain can partly solve the problem, but it involves new troubles such as processing speed decrease with the increase of hierarchy complexity and thesaurus volume, as well as domain-specificity, homonymy etc. ([Barzilay R., Elhadad M. 1997]).

In our article we are discussing the semantics of a message focusing on its emotional component. We postulate that any news message or blog post always contains an emotional element. In other words, our starting point is an emotive function of communication as opposed to informational aspect. Our claim is that communicative fragments of a message with highest emotional charge are most important for the speaker and, therefore, most informative for the receiver. (For more details about communication functions see [Iakobson 1975, Iarceva et al. 1990].)

Thus, our approach can be applied is most applicable to the communication sphere that exploits emotive function (first of all, to blogosphere and mass-media). Obviously, good result is not expected where informative function dominates (science publications, law acts, instructions etc.). Still the emotive component is characteristic for texts of most styles and genres. It is almost always present in news messages, social network texts and, certainly, in verbal communication[2]. Hence, we propose the following hypothesis: the information retrieval task can be resolved via sentiment analysis of every sentence in the message and its sentiment force evaluation.

To verify our hypothesis, we used sentiment-based method of automatic text summarization. Our method was further compared to two other summarization methods (based on self-information and on TF/DF measure) and also to the performance of TextAnalyst system (Microsystems).

The structure of the rest of the article is as follows. Section 2 presents sentiment analysis as an instrument for text summarization. After basic concepts of sentiment analysis (2.1) a very brief overview of the method of calculating sentiment degree of a sentence is given (2.2). As in our investigation we used a program for sentiment analysis, some details of its functionality are described in 2.3 and 2.4. Section 3 focuses on the experimental verification of the aforementioned hypothesis. The conditions of the experiment are given in 3.1. As in our experiment we compared our method with three other ones, those are described in 3.3 after a very brief overview of summarization methods (3.2). The results of the experiment are presented in Section 4 and the interpretation and further perspectives conclude the article (5).

[2] “A man, using expressive features to indicate his angry or ironic attitude, conveys ostensible information [...]” (Jakobsón R. Language in Literature. Ed. Krystyna Pomorska and Stephen Rudy. Cambridge (Massachusetts), Belknap Press, 1987, p.67)

2. Sentiment and its use for text summarization

2.1. What is sentiment

We have already discussed the concept of lexical sentiment and the approach to its automatic detection in our previous article [Pazel'skaia, Solov'ev 2011]. We define lexical sentiment as either lexeme-level or communication fragment-level emotional component. The sentiment of a whole sentence is calculated on the basis of lexical (given in a dictionary) sentiment of its words and communicative fragments[3], considering linkage rules. Sentiment is always related to a special object of sentiment. Besides, taking into account the sentiment subject (the author of the statement) brings us to the task of opinion mining[4]. The sentiment score (the force of the sentiment component) is another valuable parameter for analysis. Though in the general case, emotional space can be described as multi-dimensional, we examine one-dimensional sentiment space, the sentiment being measured over "positive-negative" scale together with the force of the positive or negative sentiment expressed — sentiment score.

2.2. Calculating the sentiment

The polarity score is evaluated using contribution of several factors and is calculated within each sentence. The key factor is those adjectives, adverbs and collocations (with either neutral or ambiguous sentiment) that strengthen the emotional component. The sentiment score for these fragments is defined in sentiment dictionaries and has three levels (neutral, strong, very strong). Tonal adjectives, nouns and auxiliary parts of speech have neutral sentiment score. This is a technical simplification, imposed by the fact that these tokens combine with emotion-strengthening components to form sentiment chains. This operation together with word chains sentiment recalculation sometimes causes difficulties, therefore it is easier to reserve lexically defined sentiments strength for adverbs, tonally neutral adjectives, nonverbal collocations, and for predicative elements, namely, for verbs and verb collocations.

Sentence-level sentiment is calculated as a sum of sentiment strength of joined word chains in the sentence. Suppose that after building sentiment chains a sentence is represented as "subject chain" — "predicative chain"— "object chain". Then the system calculates sentiment value (positive/negative) and polarity score for each chain.

[3] Communicative fragments are speech fragments of different length that are stored in the speaker's memory as fixed segments of speaker's language experience that he (she) uses uttering or interpreting expressions [Gasparov, 1996]. Usually a communicative fragment is longer than a word and shorter than a sentence.

[4] Demo systems for sentiment analysis and opinion mining developed in CJSC "I-Teco" can be found at http://x-file.su/tm/ (sentiment analysis) and http://x-file.su/ds/ (opinion mining).

After that a special set of rules defines sentiment of the whole sentence, and the sentence polarity score is calculated as a sum of polarity scores of the three chains. Therefore, when calculating polarity score, we do not look at its value, these two parameters are independent.

Text-level sentiment takes into account the number of sentences with expressed sentiment in the text and their polarity score. It should be mentioned that even sentiment bearing sentences may have neutral (or zero-valued) polarity score if sentiment is weakly expressed, i. e. there are only lexis with neutral sentiment strength without any sentiment amplifiers.

2.3. System output

For readability purposes, in the system output the sentences where the sentiment score is higher than neutral are displayed with bigger font, the font size being proportional to the sentiment score. Depending on the sentiment score and sentiment value, a sentence can bear one of the following seven sentiment characteristics: weak negative, medium negative, strong negative, neutral, weak positive, medium positive and strong positive.

Examples (in italics):
В таких условиях состояние здоровья мамы не улучшается, а ухудшается.
(In such circumstances, mum's health is not improving, but is getting even worse.)– mixed mean negative and strong negative sentiment;
Кроме того, *он планирует запретить пропаганду культа личности*, пишет газета “Коммерсантъ”.
(Besides, he plans to prohibit the cult of personality propaganda, writes “Kommersant”.) — neutral and strong positive sentiment (italics);
Независимость автономии на данный момент признали 127 государств.
(127 countries have already recognized independence of the autonomy.)– weak positive sentiment;
Он хорошо освоил программирование, ***но так и не научился писать стихи.***
(He became keen in programming, but haven't learnt to write poems.)– weak positive (italics) and weak negative (bold italics) sentiment.

2.4. Sentiment-score in text summarization

Our automatic sentiment analysis system has an option of extracting non-neutral sentences only. The filter uses absolute values of sentiment score. If a sentence includes sub-sentences with different sentiment score, larger score is chosen. We used this filter function to get an emotive summary of a document. This gave us a small corpus of texts and their summaries which was used to check the hypothesis stated in the Introduction.

3. Experimental verification of the hypothesis

Our experiment was as follows. From each document those sentences were extracted that had absolute sentiment value over zero. The summaries obtained in this way were compared to those received with pure statistical methods (see Section 3.2). Each summary of every document was evaluated by a number of respondents.

3.1. Material and respondents

For current research we took originals of news texts and blog posts (source texts) and also their summaries, i. e. sequences of all non-neutral sentences from these texts. Every source text selected for the experiment meets the following criteria:

or current research we took originals of news texts and blogs and also their summaries, i. e. all chains of non-neutral sentences from these texts. Every selected text answers the criteria:

1) not very short (more than five sentences) and not very long (not more than 2000 symbols);
2) no thematic restrictions were posed for news texts (our sample includes texts on politics, incidents, sports, science, culture etc.), as sources we chose news portals www.rbc.ru, www.lenta.ru, www.rian.ru, www.utro.ru;
3) blog texts were picked from blog corpus belonging to CJSC "I-Teco" (http://www.i-teco.ru), none of the texts were news reports, advertisements or announcements.

Finally 25 random (with restrictions mentioned) official news texts and 25 blogs posts were taken. The number of texts was restricted by the experiment time limitation (2–3 hours for each respondent, in order to not to exhaust the experiment participants).

19 volunteers (age 20–60) participated in the experiment. None of them was experienced in analyzing summaries.

Our respondents were asked to read the source text and evaluate each summary according to 2 scales, each having three values:

1) summary quality, that we for convenience called precision[5] (good (contains all relevant information) / normal (contains only part of the relevant information) / bad (noise, all the relevant information is lost));
2) redundancy (no redundancy / little redundancy / very redundant).

The redundancy score (though sometimes correlated with the precision score) was taken to provide better evaluation of those summaries that apart from key

[5] In current research, precision is an averaged psycholinguistic score of respondents' subjective estimation of summaries. It appears impossible to estimate precision objectively as no tools for message semantics estimation exist.

sentences contain extraneous information. The participants were not asked to evaluate logical consistency, as it would bring us far from the goal of our investigation.

The respondents were unaware of the summary extracting methods we used. All texts were printed with the same type and color. Titles of source texts were erased. Pauses were allowed while completing the task.

3.2. Experiment preparation: automatic summarization of the documents

Though many different approaches to text summarization were presented in the literature, most of them can be divided into 2 groups. First is so-called quasi-summarization, where summaries consist of sentences extracted from the source document without any modifications. The second group algorithms, summary generating ones, generate a new text that somehow interprets the source one. Our experiment was held within the first approach (sentence extraction; for more detailed overviews see [Han, Mani 2000; Das, Martins, 2007; Ganapathiraju 2002]) Summarization algorithms can be further divided into two types: 1) ones that use statistical information on words and collocations distribution and 2) those based on linguistic information i. e. exploiting tokenization and syntax information together with dictionaries and thesauri.

As linguistic rules are typically domain-dependent and genre-dependent, it appears to be difficult to use them when working with texts belonging (like in our case) to different domains and genres, especially with blog posts. Therefore, only statistical algorithms were taken to be compared with our system. These are 1) an algorithm based on TF/DF; 2) an algorithm exploiting self-information measure. The TF/DF measure is widely used in summarization systems, e. g. in HMM (see [Fung 2003]). Second method was used to demonstrate the difference between the concept of self-information that was brought to linguistics from information theory [Shannon 1963] and the semantic notion of information. We also used for comparison output of TextAnalyst NLP system, a product of Microsystems company.

3.2.1. Automatic summarization using self-information[6] measure

This algorithm can be briefly described as follows. First in our mixed corpus (blogs post and mass-media, about 100 million words) all the personal names were replaced with a special tag. Second the corpus was statistically processed with free distributed SRILM-tools[7]. Then for each unigram of the language model a logarithm of its probability (i. e. its amount of self-information) was calculated. The possibility to come across a new word was also considered (Katz's back-off model). Finally from each source text the system extracted 1/3 of its sentences (with maximum value

[6] Self-information is a measure of the information content associated with the event with certain possibility. For an independent event X with possibility P Hartley formula was used: $I(X) = - \log_2 P(x)$.

[7] Available at http://www.speech.sri.com/projects/srilm

of self-information averaged and normalized by the number of unigrams in a sentence). This extraction was taken as a summary.

3.2.2. Automatic summarization using TF/DF measure

The TF/DF based algorithm [Salton 1988] on step 1 evaluated weights of terms (unigrams, bigrams and trigrams) in each sentence (every n-gram was considered as a separate term). We cut off rare n-grams and took only 50 % most frequent unigrams, 15 % bigrams and 5 % trigrams. Weights for news and blogs were evaluated separately. On step 2, by intersecting weighted term-document matrix with all terms in a document, normalized sum of terms for every sentence was calculated. Then (step 3) weights were normalized by the sentence length. The summary included 1/3 of document sentences with the highest normalized term weight.

3.2.3. Automatic summarization by TextAnalyst

To compare to summaries extracted by our system, we used commercial program TextAnalyst[8] that provides a summarization module. The system uses neural network model for inner text representation and ranks sentences using weights for previously identified key terms. With this instrument we collected summaries for news texts (about 80 % texts were successfully processed). As the program failed with most of the documents from social networks, TextAnalyst was excluded from comparative analysis of blog summaries.

3.2.4. Experiment conditions

Neither statistical algorithms nor the sentiment-based one had the advantage of using syntactical analysis. The weight of terms from the first sentence was not increased, although it could have been quite reasonable for news texts. No special heuristics for maximizing summarization accuracy were used (such as headlines contribution — technique that often turns to be absolutely useless when dealing with blogs).

Finding the best algorithm was not our aim. Our purpose was to show that sentiment-based method possesses its own competitive advantages and can be used either separately or within some complex method. So we emphasize that (except for TextAnalyst system) each method is not a ready to use one and can be equally improved. Thus, the three summarization methods for news texts and 2 methods for blogs were used in our experiment to compare with our sentiment-based method.

4. Results

Having received the results from our respondents, we compared the scores for each method. Tab.1 below presents this comparison.

[8] Free version at http://www.analyst.ru/index.php?lang=eng&dir=content/products/&id=ta

Table 1. Average precision and redundancy values and their standard deviation for the summarization methods. For the three-point scale where 1 is "bad" and 3 is "good"

	News		**Blogs**		**News**		**Blogs**	
Method	**Precision**	**Stand. deviation**	**Precision**	**Stand. deviation**	**Redundancy**	**Stand. deviation**	**Redundancy**	**Stand. deviation**
Self-information	1.69	0.22	1.71	0.24	1.88	0.33	1.94	0.36
TextAnalyst	1.79	0.21	–	–	**2.41**	0.23	–	–
Sentiment	**2.38**	0.25	**2.07**	0.27	2.08	0.28	**2.24**	0.25
TF/DF	2,04	0,23	1,91	0,20	1,99	0,25	1,74	0,27

Below the results are illustrated with histograms (see Figure 1).

The examination of processing results for both blogs and mass-media texts showed that the precision value was a bit better for the summarizing method extracting most expressive sentences from texts. (Although the advantage 2.07 against 1.97 (for blogs) can be considered as statistically insignificant, we would remind that our main purpose was not to prove our method highest accuracy compared with another methodologies.)

The widely used methodology based on TF/DF measure showed the second best result for precision test. Next is TextAnalyst that also had the best value for redundancy scale (on news texts).

The worst scores for both news and blogs categories were obtained by the methodology of extracting sentences with highest self-information values. Such comparatively low result had been anticipated. In information theory, the less probable event is considered to be most informative. Probabilistic concepts from information theory can be successfully applied for extracting statistically significant lexical, grammatical or syntactical templates (e. g. in collocation extraction, document flow frequency analysis etc.). Nevertheless, the result of application self-information measure to semantic analysis is very poor because the concept of "information" in mathematical theory of communication differs from the one in natural language theory.

The average precision value for mass-media is better than for blogosphere (2.04 against 1.89 respectively). Blog discourse structure is much more complicated, due to indirectly expressed sense (that often demands context understanding and further including of additional knowledge), polythematic character of a single post, low narration consistency.

The minimum redundancy value is shown by mass-media summaries by TextAnalyst system. Sentiment-based methodology has the second rank. Both statistics-based summaries have relatively low and quite similar redundancy because those algorithms always extracted 1/3 of source text.

We scored the degree of unanimity of our respondents with non-parametric Kendall's coefficient of concordance. This statistic was chosen because it makes no assumptions regarding the nature of probability distribution. We received the following results. Mass-media texts were scored with coefficient of W=0.49 for precision and W=0.29 for redundancy. The coefficient for blogs is lower W= 0.36 for precision and W=0.22 for redundancy. The redundancy criterion was scored with higher variability level. We suppose that this happened because respondents interpreted the criterion differently.

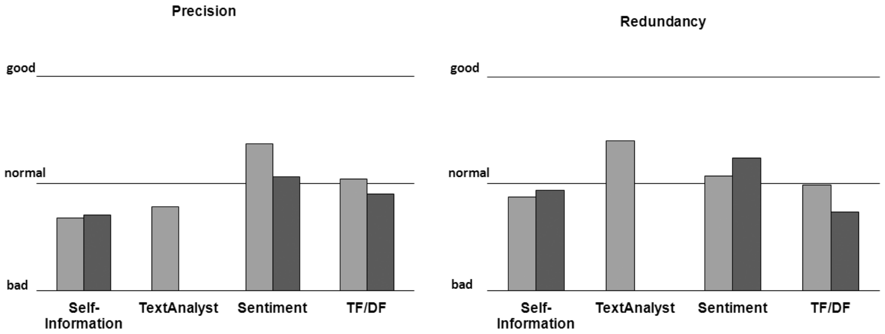

Figure 1. Comparison of four summarization methods. The left picture shows precision and the right one redundancy criterion. Light-grey boxes are used for media and dark-grey for blog texts

5. Conclusion and perspectives

This paper presents the following results. We applied sentiment analysis method to the task of summarization. This approach proved to be competitive with other previously developed summarization methodologies. The specific of our method lies in its ability to get quite a good summary without prior statistical processing of large document collections. Nevertheless, the contribution of statistical features will improve the summarizing accuracy. Application of coreference links and anaphora resolution would also significantly raise the quality of summaries.

Another contribution of our research is the proof (though indirect) of independence between document semantic characteristics and the amount of self-information from the mathematical theory of communication.

A possible application of the proposed method is the "text emotional temperature" estimation in its dynamic as well as comparing sentiment force for different sources commenting the same event. The theoretical issue of our research is that at least for some types of texts on emotional function of communication is inseparable from information function, emotive and informative peaks of text coinciding. In general case it might be untrue, but still this generalization holds for those texts people deal with in their everyday life. Thus, emotionality is the essential characteristic of interpersonal communication.

Acknowledgement

We thank all the respondents who took part in our research and we would like to acknowledge the help of Denis Mazur who had designed the architecture of sentiment-analysis instrument.

References

1. *Barzilay R., Elhadad M.* 1997. Using Lexical Chains for Text Summarization. — ACL/EACL Workshop Intelligent Scalable Text Summarization.- Madrid.Das D, Martins A. F.T. A Survey on Automatic Text Summarization.Literature Survey for the Language and Statistics II course at CMU, November, 2007
2. *Braslavski P., Kolychev I.* eXtragon: an Experimental System for Automatic QueryBiased Summarization of Web documents [eXtragon: eksperimental'naia sistema dlia avtomaticheskogo referirovaniia veb-documentov] ROMIP-2005, St-Petersburg, Russia, 2005, pp. 40–53.
3. *Das, D., Martins, A. F. T.*: "A Survey on Automatic Text Summarization"; Literature Survey for the Language and Statistics II course at CMU (2007)
4. *Fung P., Ngai G., Cheung C.-S.,* Combining optimal clustering and hidden Markov models for extractive summarization, in: ACL Workshop on Multilingual Summarization and Question Answering, Association for Computational Linguistics, Morristown, NJ, USA, 2003, pp. 21–28
5. *Ganapathiraju M. K.* "Relevance of Cluster size in MMR based Summarizer: A Report 11-742: Self-paced lab in Information Retrieval". November 26, 2002.
6. *Gasparov B. M.* (1996) Aizyk, pamiat', obraz. Lingvistika iazykovogo sushchestvovaviia [Language, memory, image. Linguistics language of existence]. Moscow, Novoe literaturnoe obozrenie [New literary review].
7. *Gusev V. D., Miroshnichsenko L. A., Salomatina N. V.* Thematic Analysis and Quasi-Abstracting of Text Using Scan Statistics [Tematicheskii analiz I kvasireferirovanie teksta s ispol'zovaniem skaniruiushikh statistic]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005"]. Zvenigorod, 2005, pp. 121–125.
8. *Hahn U., Mani I.* The Challenges of Automatic Summarization, IEEE Computer, November 2000, pp. 29–36.
9. *Iakobson R. O.* (1975) Lingvistika i pojetika: Strukturalizm: «za» i «protiv». [Linguistics and Poetics: structuralism "pro" and "contra"]. Moscow.
10. *Iarceva V. N., Klimov G. A., Zhuravljov V. K.* Sovetskoe jazykoznanie // Lingvisticheskij jenciklopedicheskij slovar' [Soviet Linguistics // Linguistic encyclopedic dictionary]. Jarceva (ed) — Moscow: Soviet encyclopedia, 1990.
11. *Lukashevich N.* Tezaurusy v zadachakh informatzionnogo poiska [Thesauri in Information Retrieval Tasks]. Moscow, MSU, 2011, pp.265–271 and pp.414–446.
12. *Pazel'skaia A., Solov'ev A.* A Method of Sentiment analysis of Russian Text [Metod opredeleniia emotzii v tekstah na russkom iazyke]. Komp'iuternaia Lingvistika

i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, pp. 510–523.

13. *Salton G. and Buckley C.* (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24 (5), pp. 513–523.
14. *Shannon C. E., Weaver W.* The Mathematical Theory of Communication. Univ. of Illinois Press, 1963.