ИСПОЛЬЗОВАНИЕ СЕМАНТИЧЕСКИХ ФИЛЬТРОВ В ЗАДАЧЕ КЛАССИФИКАЦИИ ОТЗЫВОВ О КНИГАХ

Фролов А. В. (anton_frolov@rco.ru),
Поляков П. Ю. (pavel@rco.ru),
Плешко В. В. (vp@rco.ru)

ООО «ЭР СИ О», Москва, Россия

В данной работе исследуется метод использования семантических фильтров в качестве классификационных признаков для решения задач классификации отзывов о книгах на 2 (положительный, отрицательный) и 3 (положительный, отрицательный и нейтральный) класса. Кроме того, проанализированы основные ошибки и подводные камни, которые могут встречаться в задачах подобного рода.

Ключевые слова: анализ мнений, определение тональности, автоматическая классификация, машинное обучение, извлечение классификационных признаков, метод опорных векторов, регрессия

USING SEMANTIC FILTERS IN APPLICATION TO BOOK REVIEWS SENTIMENT ANALYSIS

Frolov A. V. (anton_frolov@rco.ru),
Polyakov P. Yu. (pavel@rco.ru),
Pleshko V. V. (vp@rco.ru)

RCO LLC, Moscow, Russia

The paper studies the use of fact semantic filters in application to sentiment analysis of book reviews. The tasks were to divide book reviews into 2 classes (positive, negative) or into 3 classes (positive, negative, and neutral). The main machine learning pitfalls concerning sentiment analysis were classified and analyzed.

Key words: opinion mining, sentiment analysis, document categorization, machine learning, classification feature extraction, support vector machine, regression, two-class classifier, multi-class classifier

Introduction

The classification problem of goods reviews is very important today. This fact is supported by increased popularity of commercial resources offering services for monitoring social networks and blogs (i.e. [7]). However, until recently there were no public collections in Russian language that could be used to test research methods. New ROMIP tracks devoted to classification of books, films and digital cameras reviews, are to fill this gap.

This paper studies methods for solving the book reviews classification problem, involving 2 classes (positive, negative) and 3 classes (positive, negative, neutral), within the framework of ROMIP 2012 [3].

Problem specification

The participants were offered a training collection, composed of blog users reviews of books of different genres (24,160 reviews in total). Each review was graded on a decimal scale. It was decided to participate in the following tracks: classification of book reviews into 2 classes (positive, negative) and 3 classes (positive, negative, neutral). In the former case the task was to divide reviews into positive and negative, in the latter — into positive, neutral (the review mentions both positive and negative features) and negative.

Generalizing facts with semantic filters

It was decided to improve the linguistic approach based on fact extraction which was presented in [6] and demonstrated good results on the last year track. Therefore, we analyzed last year results and tested a hypothesis that the training collection was too small to ensure that individual facts have high enough frequencies to be used as good classification features. A possible solution for this problem is an application of semantic filters that allow combining several facts into one class.

Recall, that fact extraction is performed by the means of semantic templates. Semantic template is a directed graph with certain restrictions applied to its vertices. The restrictions can be applied to part of speech, name, semantic type, syntactic connections, etc. (see Fig. 1). Fact extraction is performed by finding a subgraph of a sentence syntax tree which is isomorphic to the template (with all restrictions applied).

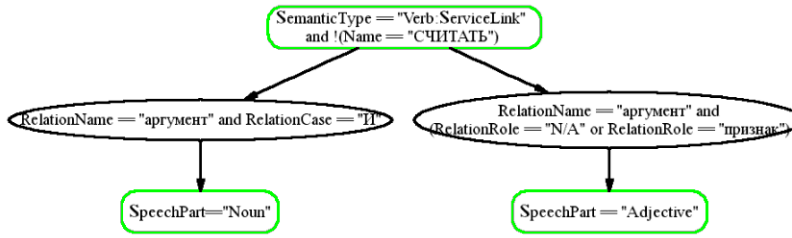

Fig. 1. Semantic template for detecting book review tonality

Moreover, facts can be generalized by the use of special dictionaries (so-called filters), containing synonyms for positive, negative and neutral appraisals. The main flaw of this approach is the necessity of manual selection of terms for the filters. It makes filter generation a labor-intensive task that requires help of a linguistic expert. On the other hand, the expert may form only the most basic vocabulary and all the additional terms can be added by an automated system. It was decided to rely on this method.

The idea was implemented as follows: we took filters used in last year track and expanded them by terms that the system was able to find independently. For more detailed explanation of how fact extraction and filters application works see [6].

New vocabulary was constructed as follows:

The training collection was processed by a system tuned to fact extraction. Then, the collection was classified by using the obtained facts as the only classification features. It was decided to use Naïve Bayes classifier with Poisson function as the PDF for words [5]. The system considered the profiles for each class individually and used filled frames slots to form the word lists for the filters vocabulary. Then, the lists were filtered against a frequency threshold and merged with the existing ones. Also, for better quality we used the vocabulary published by ROMIP organizers [2]. If ROMIP vocabulary contained a fact slot, the slot's weight was multiplied by 10.

Table 1. Filter example

Subject	Quality Verb	Quality Emotion	Quality Adjective
КОНЕЦ КНИГИ	УБИТЬ	ЖДАТЬ	УМОПОМРАЧИТЕЛЬНЫЙ
КОНЦОВКА	ИДТИ	УБИТЬ	ОПТИМИСТИЧНЫЙ
ФИНАЛ	РАСТЯГИВАТЬСЯ	НЕ ЖДАТЬ	ДУРАЦКИЙ
РАЗВЯЗКА	СДЕЛАТЬ	ИДТИ	ЗАКРЫТЫЙ
ХЭППИЕНД	НЕ ПОНРАВИТЬСЯ	РАСТЯГИВАТЬСЯ	УТОМИТЕЛЬНЕЙШИЙ
ХЭППИ ЭНД		НЕ ПОНРАВИТЬСЯ	СКУЧНЫЙ
ХЭППИ		НАЗВАТЬ	ПЕЧАЛЬНЫЙ
ХЭППИ-ЭНД			

Table 1 shows an example of an automatically filled filter, which combines several facts into one class: “negative review concerning book ending”. In this case, facts with four slots (subject, quality verb, quality emotion, quality adjective) will be merged if the contents of their corresponding slots belong to the same filter.

Despite rare errors (in example — term “оптимистичный” has been included in a filter for the negative class) most of the vocabulary is adequate. Furthermore, the quality of selected terms can be improved by increasing representativeness and size of the training sample.

Thus, we obtained fact classes for identifying tonality of reviews concerning characters, language, storyline, and author evaluations.

Classification methods

To obtain a good training set two of our experts independently evaluated the collection and marked reviews as being mostly positive, mostly negative or having both positive and negative features. Every expert evaluated about 4,000 reviews with most of them been marked as positive. The experts agreement equals $r \sim 0.8$, where r is Pearson’s correlation coefficient. Two approaches were used for the experiment.

In the former approach the classifier was trained using blog users’ evaluations and these evaluations were used to build a linear regression model (SVM-Light implementation, see [4]). Then, this model was used to compute weights of documents from the training collection and to determine thresholds for relating documents to corresponding classes so that the difference between the system’s partitioning and the experts’ partitioning is minimized (F-measure was used as an utility function).

In the latter approach the classifier was based on the training set, formed by the experts. Following classification methods were used:

- Linear classifier with the learning stage been conducted for each class independently (SVM-Light implementation, see [4]). In the case when the same document is classified as being a member of several classes, we select the class where the document has the greatest weight.
- Linear classifier, with the learning stage been conducted independently for 2 classes (positive and negative), that was used to classify documents into 3 classes (SVM-Light implementation, see [4]). In this case, a document is marked as being a member of the neutral class if the classifier considers it as being a member of both negative and positive classes.

Results

This paper studies the results of 4 runs devoted to classification into 2 classes and 4 runs devoted to classification into 3 classes. The runs are parameterized with the classifier’s type:

- SVM: support vector machines method with “one against all” partitioning
- Regression: linear regression model

and classification features sets:

- Base: classification features are lemmas (single words) and themes (word-combinations)
- Hybrid: fact classes are used in addition to Base features

We used F1-measure as a primary evaluation metric [1]. Additionally, for convenience, recall, precision and accuracy are also present in the tables.

Table 2. Runs for 2 classes

	P-macro	R-macro	F-macro	Accuracy
Base SVM	0.676425	0.620273	0.647133	0.86046
Hybrid SVM	0.577041	0.552521	0.564515	0.82945
Base Regression	0.627363	0.627363	0.627363	0.82945
Hybrid Regression	0.605004	0.634454	0.619379	0.79845

The data given in Table 2 indicates that the Base SVM classifier demonstrates the best result. The explanation for this fact is given in the next section of this paper. Also, it is evident that in the case of the hybrid model, the regression based classifier shows better result than SVM.

Table 3. Runs for 2 classes (detailed)

	P-pos	R-pos	F-pos	P-neg	R-neg	F-neg
Base SVM	0.898305	0.946428	0.921739	0.454545	0.294117	0.357143
Hybrid SVM	0.881355	0.928571	0.904348	0.272727	0.176471	0.214286
Base Regression	0.901785	0.901786	0.901786	0.352941	0.352941	0.352941
Hybrid Regression	0.905660	0.857143	0.880734	0.304348	0.411765	0.350000

Table 3 indicates that correct identification of negative reviews was the most difficult task for the classifier. The complexity of this task can be explained by the following factors:

1. Most of reviews in both test and training collection were positive: 112 (positive) vs 17 (negative).
2. The size of the test sample was small: as little as 129 documents. This, in addition to factor 1, leads to the result being statistically biased.
3. A significant part (8 out of 17) of negative reviews did not contain explicit negative opinions. Such reviews were correctly identified as neutral under classification into 3 classes.

It is worth mentioning, that the test collection was evaluated out by only one expert which results into increased bias in the final result.

The classification into 3 classes demonstrates completely different picture: SVM performs better than the regression model.

Table 4. Runs for 3 classes

	P-macro	R-macro	F-macro	Accuracy
Base SVM	0.544343	0.554074	0.549165	0.697674
Hybrid SVM	0.450879	0.467037	0.458816	0.666666
Base Regression	0.354825	0.333703	0.343940	0.542636
Hybrid Regression	0.354826	0.333704	0.343941	0.542636

Table 5. Runs for 3 classes (neutral class)

	P-neu	R-neu	F-neu
Base SVM	0.891566	0.74	0.808743
Hybrid SVM	0.870588	0.74	0.800000
Base Regression	0.857142	0.72	0.782608
Hybrid Regression	0.864864	0.64	0.735632

Table 6. Runs for 3 classes (negative, positive)

	P-pos	R-pos	F-pos	P-neg	R-neg	F-neg
Base SVM	0.341463	0.70	0.459016	0.400000	0.222222	0.2857140
Hybrid SVM	0.282051	0.55	0.372881	0.200000	0.111111	0.1428571
Base Regression	0.147058	0.25	0.185185	0.090909	0.111111	0.1000000
Hybrid Regression	0.116279	0.25	0.158730	0.083333	0.111111	0.0952380

Results analysis

The result was strongly affected by several properties of the test collection. Namely: collection's small size (twice as small as the last year collection) and strong odds towards neutral reviews (positive, in case of binary classification). Additionally, negative reviews are biased: about half of them are devoted to the same book, namely, "Angels and demons" by Dan Brown.

The agreement between our expert and ROMIP expert equals $r = 0.78$

As it is possible to see from the tables, negative reviews posed the main problem for the classifier. We analyzed and classified errors, made by the system. They can be divided into following categories:

The author mostly retells the storyline. In this case, the text may contain enough noise terms for the classifier to make an error.

Despite the author speaks about book's positive features, the final evaluation is negative, e.g: "Сюжет есть. И интрига присутствует. А вот то, как разворачиваются действия — не вдохновляет ни коим образом." As a result, positive terms overweight negative terms only due to their number. The methods employing fact extraction are particularly vulnerable to errors of this kind. The reason is that it is much more difficult to gather enough statistics for facts than for lemmas.

Although the author mentions positive reviews by other people, his/her own evaluation is negative, e. g. "С сожалением сообщаю: не для моих мозгов. Говорят, книга очень хорошая. Промолчу."

The presented system used a semantic filter rather than a regular stop-words list, i. e. all numerals and auxiliary words were filtered out. This method demonstrated good results in classification of reviews from Imho-net. However, current track contains blog posts rather than ordinary reviews. Blog posts vocabulary contains significantly more noise terms that cannot be filtered out by semantic filters solely.

It is worth mentioning, that we used "one against all" partitioning and chose class where the document had the greatest weight. As a result, many incorrectly classified documents had negative weight for both classes. In classification into 3 classes the system correctly identified such reviews as being neutral.

Table 7. Comparison of last year and this year results

	Expert 1 F-macro	Expert 2 F-macro
New hybrid SVM	0.503129181	0.500560892
Old hybrid SVM	0.467705308	0.484938518
Base regression	0.490300000	0.499800000

It is evident, that the classifier that employs fact extraction demonstrates worse results than the basic one. We suspected that the reason is that the bias of the collection. To prove it we conducted experiments with the last year collection. It turned out that the new classifier demonstrated improvement in classification into 3 classes in comparison to hybrid system and even regression method [6] that was the leader among all the systems participated in the last year track. It follows that the new classifier performs better than the old ones, provided the collection is not biased.

Possible improvements

The above mentioned problems can be solved by changing the set of classification features. First of all, it is important to be able to distinguish the summarizing assessment. Indeed, such reviews mostly contain retelling of a storyline or an irrelevant discussion. The same time the statements that truly characterize the review are contained in a few sentences in the beginning or the end of the text.

Secondly, it is desirable to be able to identify the object being reviewed. The point is the same review can discuss several books simultaneously, e. g.: "Сегодня

я читал X и мне не понравилось. Гораздо хуже замечательной книги Y, которую я читал вчера". If the object is not specified the system should be able to identify it itself.

Thirdly, in classification into three classes the author's opinion should be distinguished from outer sources opinions ("говорят книга хорошая, но мне не очень понравилась"). In this case, the author's opinion obviously has a greater weight. In classification into three classes this factor is not so critical and different sources can be assigned with similar weights.

Conclusion

We tested several methods of classification into 2 and 3 classes and improved the linguistic approach, based on application of evaluative vocabulary, by application of automated filters generation. Additionally, main errors made by the classifier were analyzed and categorized. Finally, the direction for future work has been set.

References

1. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* (2011), Sentiment Analysis Track at ROMIP 2011, available at: www.dialog-21.ru/digests/dialog2012/materials/pdf/83.pdf.
2. *Chetviorkin I., Loukachevitch N.* (2012), Extraction of Russian Sentiment Lexicon for Product Meta-Domain, Proceedings of COLING 2012.
3. *Chetviorkin I., Loukachevitch N.* (2012), Sentiment analysis track at ROMIP'12.
4. *Joachims T.* (1998), Making large-scale support vector machine learning practical. Advances in Kernel Methods: Support Vector Machines, MIT Press: Cambridge, MA.
5. *Pleshko V. V., Polyakov P. Yu., Ermakov A. E.* (2009), RCO at RIRES 2009 [RCO na ROMIP 2009]. Trudy ROMIP 2009 [Proc. ROMIP 2009]. Petrozavodsk, Saint Petersburg, pp. 122–134.
6. *Polyakov P. Yu., Kalinina M. V., Pleshko V. V.* (2012), Research of applicability of thematic classification to the problem of book review classification. Dialog '12. Naro-Fominsk.
7. *Sentiment 140* (2012), Available at: www.sentiment140.com.