USING BASIC SYNTACTIC RELATIONS FOR SENTIMENT ANALYSIS

Mavljutov R. R. (m-ceros@yandex.ru),
Ostapuk N. A. (nataxane@yandex.ru)

Yandex, Moscow, Russia

The paper describes a rule-based approach to sentiment analysis. The developed algorithm aims at classifying texts into two classes: positive or negative. We distinguish two types of sentiments: abstract sentiments, which are relevant to the whole text, and sentiments referring to some particular object in the text. As opposed to many other rule-based systems, we do not regard the text as a bag of words. We strongly believe that such classical method of text processing as syntactic analysis can considerably enhance sentiment analysis performance. Accordingly, we first parse the text and then take into account only the phrases that are syntactically connected to relevant objects. We use the dictionary to determine whether such a phrase is positive or negative and assign it a weight according to the importance of the object it is connected with. Than we calculate all these weights and some other factors and decide whether the whole text is positive or negative. The algorithm showed competitive results at ROMIP track 2012.

Keywords: sentiment analysis, opinion mining, syntactic relations, context-free grammar, thesaurus

1. Introduction

Automatic sentiment analysis is a comparatively new field in computational linguistics. With developing of Web and particularly blogosphere every Internet user got the opportunity to leave a review, expressing his or her opinion about some product or service. Such information is useful both for other users and for market departments of service providers. The problem is this information is large, so it cannot be processed manualy. As an illustration, the website TripAdviser.com publishes about 40 reviews every minute, and booking.com has almost 18 million reviews overall. The methods of natural language processing may be helpful to takle the issue with big amount of data. On the basis of these methods systems of sentiment analysis are being developed. The goals of the SA systems vary from text tone assessment to extraction and assessment of specific parameters, which are discussed in the text.

Automatic sentiment analysis task encounters a lot of problem, such as implicit expression of emotional component in the text, too informal language of reviews and until recently lack of annotated corpus for Russian to measure the quality. To settle the last problem, ROMIP offers sentiment ananlysis track, which aims at classifying blog posts about books, films and cameras according to the sentiment they express into 2, 3 or 5 groups.

The current version of our system classifies reviews into two groups. The algorithm is based on rules, which take into account syntactic relations in the text. The main goal of our participation in ROMIP 2012 was to measure the quality of work of our system and to compare it with others in order to understand if we are on the right way and at what else we have to work.

2. Related works

All existing approaches to sentiment analysis can be divided into two large categories: rule-based and machine learning based.

Sentiment analysis based on machine learning in general is similar to classical task of text classification, where sentiment words act as features. The commonly used method here is support vector machines trained on large annotated corpora [3], [5], [8].

Rule-based methods make use of sentiment lexicon of the text. Such methods vary from simple lists of positive and negative words to more sophisticated methods, using sentiment patterns and syntactic relations between words in the text. Approaches which involve syntactic relations are mostly developed for English language [11], [14]. For the Russian language the task of constructing syntactic tree is much more complicated, taking into account rich morphology and free word order.

In [7] the syntactic approach to sentiment analysis for Russian was implemented. This system aimed at determining news texts tone. It extracts the object of evaluation as well as syntactic groups with opinion words and according to some set of rule combines them.

3. Method description

In our work we implemented the following algorithm: first, we gathered object thesaurus, including terms to which opinion phrase could refer. Then we detected phrases syntactically connected to objects from the thesaurus, as well as negations relevant to these phrases — such syntactic groups became potential entries of our sentiment dictionary. It's worth noticing, that we considered the whole syntactic group including the object as a sentiment; not just opinion phrase: this issue will be considered in details in Section 3.3. After that we compiled a sentiment dictionary using mined syntactic groups and some additional resources and finally we searched for sentiments in the text and weighed them to determine text tone.

3.1. Objects thesaurus

For each class of objects (films, books, digital cameras) we have gathered a thesaurus, that has three categories of terms:

1. Common nouns that denote objects of the class. For digital cameras such terms are *“камера”, “фототехника”, “аппарат”, “фотоаппарат”,* etc.

2. Proper names of objects of the class. The names of the camera models and movies and books titles.
3. Common nouns that denote parameters, properties, and parts of objects of the class. As an illustration, for digital cameras the parameters are *"формфактор", "качество фото", "разрешение", "матрица", "объетив", "вспышка"* etc. For films and books they are *"автор", "режиссер", "игра актеров", "атмосфера", "дубляж"* etc.

Each element of the thesaurus had its unique id, class id and type id.

The distribution of terms quantity by object class and type was the following:

	books	**films**	**digital cameras**
common nouns	69	74	475
proper names	2,713	208	1,412
parameters	161	252	512

We have filtered ambiguous proper names (e.g., *"камень"*), to be sure that we wouldn›t mix up class objects with other entities in texts. For the digital cameras we have also made a vocabulary of contracted proper names that consists of company names and parts of the full names of the models. This vocabulary is helpful since the camera names are usually complex, so writers (especially in blogs and comments) prefer to use simplified versions. For example, instead of *"BenQ DC C1450"* they may write *"BenQ DC"*, *"BenQ", "benq", "benq dc", "c1450"*, and so on.

Gathering data for thesaurus

At the beginning of ROMIP competition we were given a vocabulary of proper names for each class as a source data. We used this vocabulary to mine common nouns and parameters. To perform this task we executed the following algorithm:

1. Gather text snippets where proper names from the thesaurus are mentioned. Each text got a class id according to the class of the proper name that was found in it. We used a part of Russian Web as a source, and we restricted the search area with texts enclosed by the paragraph tag <p>.
2. Extract all noun phrases (which do not coincide with the matched proper name), and sequences of noun phrases connected by genitive case. Let's call them potential thesaurus terms.
3. Calculate Pointwise Mutual Information between a potential term and text class, where it was found:

$$PMI\,(potential\ term,\ text\ class) = \log_2\left(\frac{p(term,\ text\ class)}{p(term) \times p(text\ class)}\right)$$

where p is probability.

The idea was that common nouns and parameters that denoted objects of a certain class would have the value of PMI for this class much bigger, than for the other two classes. So, we could choose the closest class for each potential term and calculate its affinity to the class:

$$affinity\ (term\_i,\ class\_j) = MIN\ (PMI(term\_i,\ class\_j) - PMI(term\_i,\ class\_k))$$

where $k\ != j$ is probability.

Now for each class we have a set of potential terms, and for each potential term we have the value of its affinity to the class.

4. Sort potential terms for each class by the value of their affinity and filter manually the part of them with highest values.

In our case on the first stage we have gathered 2 billion of text snippets in which proper names from the thesaurus were mentioned. On the second stage we got 60 thousand of potential thesaurus terms. We cut off a part of them with low value of affinity, and only 17 thousand were left. After correction of misprints 8 thousands were left. Then we have filtered those that left manually, and only 1.5 thousand terms became a part of the thesaurus.

3.2. Syntactic relations used for opinion extraction

Unlike to other approaches that use syntax, we didn't make full text parsing. According to our experience, there is a set of the specific syntactic relations are generally used to express subjectivity.

Previously we conducted a research which aimed at determining how subjective evaluation of an object could be expressed in the text. The training set consisting of 10 thousand hotel reviews was annotated manually. According to this markup the following distribution was received:

1. 80% of subjective evaluations are grammatical modifiers expressed by adjectives, e.g. *"громкая музыка", "плохое обслуживание"*.
2. 7% — predicates expressed in different ways: *"бармен кричал", "обслуживание было плохим", "обслуживание оставляло желать лучшего", "отель чудовищен", "отношение к клиентам просто ужас"*.
3. 4% — adverbials expressed by adverbs and prepositional phrases connected to predicate, grammatical modifier or directly to an object: *"кран работал плохо", "плохо работающий кран", "связь на троечку"*.
4. 9% — other ways. This ways include expression of subjectivity with interjections (*"брр", "фууу",* etc.), objects comparison (*"А лучше Б", "А понравился меньше, чем Б"*), reference to self (*"мне стало плохо", "я замучался его смотреть"*), and expressions, where object and opinion phrase are not connected syntactically (*"Вчера посмотрел этот фильм. До сих пор противно"*).

We didn't make a detailed study for classes proposed by the ROMIP task and texts related to blogs; however, we made an assumption that the trend would remain the same. In current research we concentrated on the first three ways of subjectivity expression. We also considered independently cases where opinions were expressed by reference to self.

We have used the Tomita-parser[12] for extracting syntactic relations between object and other parts of a sentence. The Tomita-parser is an instrument for extracting structured data (facts) from texts in natural language by means of context-free grammars. To extract a fact, we should write a set of rules, describing the structure of this fact in the text. For example, to extract an adjective agreed with a noun, we should write the next rule:

$$S \rightarrow Adj\text{<}gnc\text{-}agr[1]\text{>}\ Noun\text{<}gnc\text{-}agr[1],\ rt\text{>};$$

For our task we have written set of rules for each of three syntactic structures. In sum we got about 50 rules. The main difficulty was to describe predicates and adverbials, expressed by collocation (*оставляло желать лучшего, на троечку* etc). We searched for such collocations in the text and tried to generalize them and to describe their structure. Of cause, we could not find all of them, and that's why the grammar did not cover all desired syntactic structures — empirically, we managed to detect about 80–90% of them.

Text chunks, which were find by the grammar, were converted into facts. In Tomita, fact is a structured entity, which consists of fields. To convert text chunk into fact means to point out, with which part of the chunk we should fill every fact field. In our case facts consisted of four fields:

1. *an object from the thesaurus*
2. *type of syntactic relation between the object and the other part of the sentence*
3. *related part of the sentence*
4. *negation*

For example, the initial phrase is *"Неделю назад я купил водонепроницаемую камеру от Nikon."* In this sentence the object is *"камера"*. From all syntactic connections of the object, only one may potentially express subjectivity (the grammatical modifier), so one fact will be extracted:

1. *object: "камера"*
2. *relation: grammatical modifier*
3. *related part: "водонепроницаемый"*
4. *negation: false*

Negation extraction

Determining negations is an important part of sentiments extraction. We define negation as a part of text structure that inverts the sign of a sentiment.

In Russian negation is expressed in different ways for different parts of speech. So for each type of syntactic relations in facts we wrote a different set of rules for extraction of negations.

Examples:

(1) *'нет'* | *'без'* | *'отсутствие'* | *'лишенный'* | *'лишивший'* | *'мало'* | *'никакой'* | *'ни'* + *noun in genitive case*

(2) *‘не’ | ‘мало’ + verb in a finite form*

(3) *‘нельзя› | ‹невозможно› + verb in an infinite form*

(4) *‹не› | ‹мало› | ‹ничего› + adjective*

(5) *‘не’ + adverb, preposition phrase*

The presence of “не” (particle of negation) doesn’t necessarily express negation. For example, the expression *“не только мерзкий”* doesn’t change the sign of *“мерзкий”*. Therefore, we have also described the class of expressions, where negation words didn’t express negation.

3.3. Sentiment dictionary

As opposed to usual practice, we don’t consider opinion words apart from their context. An entry in our sentiment dictionary is a fact, not a separate word. This approach is justified by the fact that a sentiment sign depends not only on an opinion word, but also on the object, and type of the syntactic relation that characterize their connection.

Compare two facts with the same opinion word, but different objects:

1) object: “официант”	1) object: “скорость обработки сигнала”
2) relation: grammatical modifier	2) relation: grammatical modifier
3) related part: “бешеный”	3) related part: “бешеный”
4) negation: false	4) presence of negation: false

In the first case the fact describes a negative sentiment; in the second — a positive sentiment; however, the opinion word “бешеный” stays the same.

Also, some sentiments don’t base on opinions words. For example, let’s consider phrase “Брюс уже не тот”. The fields of the fact are:

1. *object: “Брюс”*
2. *relation: predicate*
3. *related part: “мом”*
4. *negation: true*

This fact denotes a sentiment; but, the word “тот” cannot be classified as an opinion word.

The task of compiling the sentiment dictionary was to collect facts, that express a subjective evaluation.

In addition to facts with all fields filled, we also considered their modifications, where values of some fields were empty. It could be a fact with empty “object” or “related part of sentence” field.

A fact with empty “object” field denotes context-free sentiment (the sign of which doesn’t depend on object). For example, the phrase *“что-то было ужасным”* represents a negative attitude regardless of the object.

A fact with empty "related part of sentence" field denote object, which convey a subjective evaluation by itself. For example, the parameters of digital cameras, like *"блики экрана", "поломка", "царапина", "битый пиксель"*, convey a negative attitude.

Compiling the sentiment dictionary

We used several sources to compile our dictionary:

1. Object-independent sentiments, which we gathered at the previous stage of our research.
2. Filtered manually and translated to our format vocabulary of sentiments given for the competition. Again, we used only object-independent sentiments.
3. The training set. The algorithm was very similar to that we used for thesaurus mining. In this case, the classes were negative and positive reviews. For each fact we have calculated its PMI with each of two classes. Then for each class we made a list of facts with the highest values of affinity to it. These facts formed the sentiment dictionary.

The size of the final vocabulary was 43 thousands of facts. Among them 5.5 thousands of facts were with empty field "object" (object-independent sentiments).

3.4. Two class classification of blog texts

After the Tomita-parser extracted facts from a text, we searched for these facts in the sentiment dictionary. Those sentiments which were found became features for the review classification.

The class of the texts was defined by the sign of the weighed sum:

$$predicted\ class = SUM\big(object\_i\_weight \times relations\_in\_sentiment\_i\_weight \times sentiment\_i\_class\big) - TRESHOLD,\ sum\ of\ all\ found\ sentiments$$

We have made the following assumptions:

1. the weight of the object expressed by a proper name or by a common noun is 1. The weight of the object parameter is 0.5
2. if the text has more than two mentions of different proper names, we consider this text as not a review, and refuse to classify it.

Thereby, the weighed sum has 4 variables to define: 3 weights for different types of relations in sentiments (modifier, predicate and adverbial) and the TRESHOLD parameter.

We used the training set to find optimal values for the parameters. As an algorithm for learning we chose SVM with cross-validation. The best results on the training set were precision 0.94, recall 0.89 for the positive class.

4. Results and further work

Here are official results from ROMIP 2012 for 2-class sentiment classification track. Our results are highlighted with blue color:

System_ID	Precision_P	Recall_P	F_Measure_P	Precision_N	Recall_N	F_Measure_N	Accuracy
Object — book							
xxx-17	0.914530	0.955357	0.934498	0.583333	0.411765	0.482759	0.883721
xxx-8	0.868217	1.000000	0.929461	0.000000	0.000000	0.000000	0.868217
xxx-27	0.873016	0.982143	0.924370	0.333333	0.058824	0.100000	0.860465
xxx-10	0.898305	0.946429	0.921739	0.454545	0.294118	0.357143	0.860465
xxx-41	0.872000	0.973214	0.919831	0.250000	0.058824	0.095238	0.852713
xxx-39	0.866142	0.982143	0.920502	0.000000	0.000000	0.000000	0.852713
xxx-3	0.910714	0.910714	0.910713	0.411765	0.411765	0.411765	0.844961
xxx-25	0.901786	0.901786	0.901786	0.352941	0.352941	0.352941	0.829457
Object — film							
xxx-23	0.857534	0.948485	0.900719	0.604651	0.333333	0.429752	0.830882
xxx-12	0.836788	0.978788	0.902235	0.681818	0.192308	0.300000	0.828431
xxx-18	0.823980	0.978788	0.894737	0.562500	0.115385	0.191489	0.813725
xxx-15	0.854749	0.927273	0.889535	0.520000	0.333333	0.406250	0.813725
xxx-14	0.817043	0.987879	0.894376	0.555556	0.064103	0.114943	0.811275
xxx-17	0.808824	1.000000	0.894309	0.000000	0.000000	0.000000	0.808824
xxx-13	0.860000	0.912121	0.885294	0.500000	0.371795	0.426471	0.808824
xxx-19	0.895899	0.860606	0.877898	0.494505	0.576923	0.532544	0.806373
Object — camera							
xxx-5	0.965937	1.000000	0.982673	0.000000	0.000000	0.000000	0.965937
xxx-13	0.975062	0.984887	0.979950	0.400000	0.285714	0.333333	0.961071
xxx-15	0.970297	0.987406	0.978777	0.285714	0.142857	0.190476	0.958637
xxx-14	0.965602	0.989924	0.977612	0.000000	0.000000	0.000000	0.956204
xxx-20	0.972431	0.977330	0.974874	0.250000	0.214286	0.230769	0.951338
xxx-2	0.977099	0.967254	0.972152	0.277778	0.357143	0.312500	0.946472
xxx-10	0.977041	0.964736	0.970849	0.263158	0.357143	0.303030	0.944039
xxx-17	0.972010	0.962217	0.967089	0.166667	0.214286	0.187500	0.936740

Precision, recall and F-measure were counted separately for positive and negative texts. Accuracy is proportion of correctly classified objects in all objects processed by the algorithm it is calculated according the following formula:

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$$

where tp is correct results, fp — unexpected results, fn — missing results and tn — correct absence of results. [2]

Our classifier has the second result (among 26 participants) in film classification and the third — in book classification (among 40 participants). A little bit worse we performed at camera classification — we are the sixth of 25. It can be explained by the fact that reviews about books and films are very much alike both in sentiment lexicon and parameters which are evaluated. Camera reviews have more specific lexicon and it was more complicated to extract sentiment facts from them. In such cases training process should be more domain-specific with less "object-independent" sentiments.

From complete result table one can see that regardless to object class precision and recall of classification of negative reviews is considerably lower than positive ones. The explanation is that negative reviews form only 10% of the flow. This correlation is true both for training set and for the Web in general. Prevalence of one class impacts on machine learning. Moreover, it complicates the process of gathering sentiment dictionary for negative class.

Despite pretty bad performance in negative reviews classification, total accuracy is still high enough. It means that test set also contained less negative reviews.

On the basis of existing system we are going to implement 3 or 5 groups classificator. Moreover, at the previous stage of our research we tried to evaluate not the whole text, but separate parameters of it, such as service, beach, rooms for hotel reviews or service, interior, food for restaurant reviews. We believe, that for such objects as hotels and restaurants, as well as cameras, cars and so on, such parametric evaluation is much useful, and that's why we are going to continue our investigation in this area.

References

1. *Chetviorkin I. I.* (2012), Testing the sentiment classification approach in various domains — ROMIP 2011, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 747–755.
2. *Chetviorkin I. I., Braslavski P. I., Loukachevitch N. V.* (2012) Sentiment analysis track at ROMIP 2011 Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 739–746.
3. *Kotelnikov E. V., Klekovkina M. V.* (2012) Sentiment analysis of texts based on machine learning methods [avtomaticheskij analiz tonal'nosti tekstov na osnove metodov machinnogo obuchenija], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 756–763.
4. *Nakagawa T., Inui K., and Kurohashi S.* (2010), Dependency tree-based sentiment classification using crfs with hidden variables, In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, Morristown, NJ, USA, pp. 786–794

5. *Pak A., Paroubek P.* (2012) Language independent approach to sentiment analysis (LIMSI participation in ROMIP '11), Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 764–771.
6. *Pang B. & Lee L.* (2008), Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, v.2 n.1–2, pp.1–135.
7. *Pazel'skaja A. G., Solov'jev A. N.* (2011) A method of sentiment analysis in Russian texts [metod opredelenija emocij v russkih tekstah], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2011"], Bekasovo, pp. 510–522
8. *Polyakov P. Yu., Kalinina M. V., Pleshko V. V.* (2012), Research on applicability of thematic classification methods to the problem of book review classification [issledovanie primenimosti metodov tematicheskoj klassifikacii v zadache klassifikacii otzyvov o knigah], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 772–779.
9. *Poroshin V.* (2012), Proof of concept statistical sentiment classification at ROMIP 2011, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 780–788.
10. *Prabowo R. and Thelwall M.* (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2) pp. 143–157.
11. *Shilpa Arora, Elijah Mayfield, Carolyn Penstein-Rose and Eric Nyberg* (2010), Sentiment classification using automatically extracted subgraph features. NAACL workshop on Computational approaches to analysis and generation of emotion in text
12. *Tomita-parser:* http://api.yandex.ru/tomita/
13. *Vasilyev V. G., Khudyakova M. B., Davydov S.* (2012), Sentiment classification by fragment rules [klassifikacija otzyvov pol'zovatelej s ispol'zovaniem fragmentnyh pravil], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 789–796.
14. *Yi J., Nasukawa T., Niblack W. & Bunescu R.* (2003), Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques, In Proceedings of the 3rd IEEE international conference on data mining (ICDM 2003), Florida, USA, November 19–22, pp. 427–434.
15. *Zirn Cacilia, Niepert Mathias, Stuckenschmidt Heiner, Strube Michael.* (2011), Fine-Grained Sentiment Analysis with Structural Features. In Proceedings of 5th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, Chiang Mai, Thailand