The Algorithm of Deep Sentiment Analysis of Ukrainian Reviews

M. Romanyshyn

This paper describes the most common approaches to sentiment analysis and defines an optimal approach for deep sentiment analysis of restaurant reviews in Ukrainian.

Introduction

Sentiment analysis is the task of natural language processing aimed at identifying positive and negative opinions, emotions, attitudes and evaluations. Nowadays sentiment analysis is widely used in such areas as sociology (e.g. collecting data from social networks about people's likes and dislikes), political science (e.g. collecting data about political views), marketing (e.g. creating product ratings), medicine and psychology (e.g. detecting signs of depression in users' messages), etc.

This paper dwells upon an attempt of implementing deep sentiment analysis of reviews in Ukrainian, as there is no such tool available yet. To find the best solution, the most common approaches to sentiment analysis have been studied. These include rule-based sentiment analysis, statistical analysis based on sentiment dictionaries, and approaches based on machine learning algorithms.

Recent Research in the Area of Sentiment Analysis

A large number of projects have appeared in the field of sentiment analysis during the last ten years: sentiment analysis of hotel reviews, bank reviews, restaurant reviews, comments on movies, products, messages in blogs and social networks, etc.

A very interesting research has been conducted by W. Kasper and M. Vela from DFKI GmbH [3]. The developers managed to combine statistical and rule-based approaches to implement sentiment analysis for hotel reviews in German. Another important research in this area was conducted by K. Moilanen and S. Pulman from University of Oxford [4], where compositional semantics was used. Sentiment analysis of Russian messages, implemented at ZAO "Ai-Teko", Moscow, used a specially created sentiment dictionary for analysis of reviews in Russian [1]. Another research for Russian used a rule-based approach for analyzing emotions in text messages [2]. These and some other studies will be mentioned further in this article.

Approaches to Sentiment Analysis

The main approaches to sentiment analysis are:

- statistical approach based on sentiment dictionaries;
- rule-based approach;
- supervised machine learning;
- unsupervised machine learning.

1. Statistical Approach Based on Sentiment Dictionaries

The first approach uses so called sentiment dictionaries. Sentiment dictionary is a list of words with their sentiment values. The sentiment value can be a number from some range (e.g. 1-10, where 1 is a negative word, and 10 is a positive word) or a certain category (e.g. positive or negative). Very often only nouns, verbs, adjectives and adverbs are listed in a dictionary, as, for example, in study [1]. All the words in a sentiment dictionary usually refer to a specific domain, as it is much more difficult to implement sentiment analysis for general domain. The most commonly used sentiment dictionaries are SentiWordNet (http://sentiwordnet.isti.cnr.it/) and General Inquirer Lexicon [8]. In this approach every word in a review is assigned a sentiment

value, stated in a dictionary, and after that the sentiment of the whole review is computed. Although this approach is rather easy to implement, it does not give high accuracy and does not give space for deeper analysis.

2. Rule-Based Approach

The majority of commercial systems use a rule-based approach. This kind of sentiment analyzers uses a sentiment dictionary and a collection of rules, based on which the system decides on the sentiment of the review. One of the plainest examples of the usage of such an approach can be found in [2]. Here the researchers took into account words-invertors and opposite conjunctions of Russian. A more complex implementation of this approach was conducted in [4], which involved part of speech tagging and parsing. Starting from the predicate, the words in the sentence were added one by one and the sentiment is computed, depending on the main word in the text fragment. The sentiment analysis of German hotel reviews [3] uses the rule-based approach together with statistical analysis, which gives the ability to both get the general sentiment of the message and conduct a deep analysis of every clause. This approach is very effective but only if the system possesses a sufficient number of manually written rules, which defines how time-consuming implementing of this approach may be. This kind of systems also needs a sentiment dictionary and may use part of speech tagging or parsing, which depends on how deep and accurate we expect the analysis to be.

3. Supervised Machine Learning

Using machine learning algorithms became popular during the last few years. Sentiment analysis on the basis of supervised machine learning algorithms involves training a classifier on a sentiment-annotated corpus and using the trained model for defining sentiments of new reviews. Machine learning classifiers can be used plainly or hierarchically, meaning that it is possible, for example, to train a binary classifier first to define neutral and subjective reviews, and then another classifier to differentiate between positive and negative reviews, like in [5]. Linear regression, as one of machine learning algorithms can be used, when we are trying to find a numerical sentiment value. Supervised machined learning algorithms can define sentiments of reviews rather accurately but only if there is a sentiment-annotated corpus with a sufficient amount of data for learning (supposedly more than 500 thousand words). This approach is very often used together with a rule-based approach in order to get both accurate general sentiment of a review and deep analysis of every clause.

4. Unsupervised Machine Learning

Unsupervised machine learning has not proven itself to be very effective for sentiment analysis yet. To implement this approach a corpus is also needed, but it doesn't need to be annotated. The task of the system is to find text fragments, which seem to be subjective. Then the general direction of the sentiment of the review can be defined.

Defining an approach for deep sentiment analysis of Ukrainian

The aim of this study is to implement deep sentiment analysis of reviews in Ukrainian. We are setting our eye specifically on deep sentiment analysis, as it sets the aim of the research not just on defining the sentiment of the review, but also on the analysis of the information conveyed in each sentence. Having conducted a detailed research of existing approaches to sentiment analysis, we found that the most suitable approach for deep sentiment analysis will be a rule-based approach. This approach makes it possible to define subjective text fragments that reflect the opinion of the author.

The implementation of deep sentiment analysis for Ukrainian language can be divided into the following steps:

- defining and implementing text preprocessing tools;
- creating a sentiment dictionary;
- constructing rules for sentiment compounding;

• creating a result representation tool.

To be more precise, the algorithm of deep sentiment analysis for Ukrainian reviews can be presented with the scheme in Figure 1.

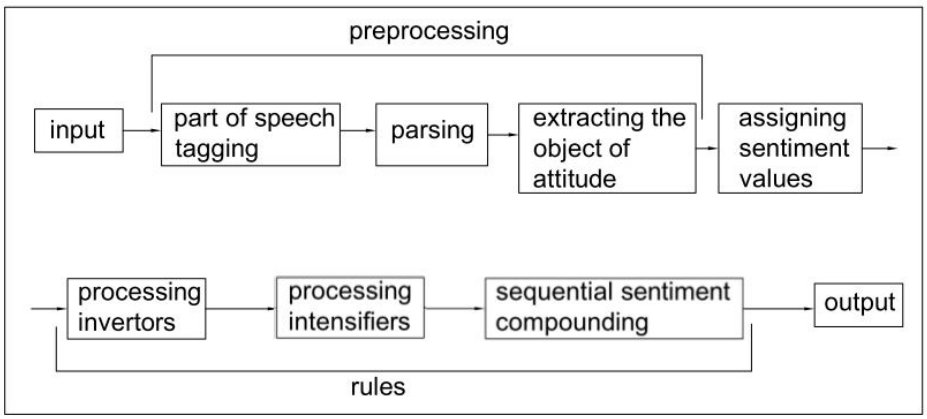

Figure 1. The main steps of sentiment analysis.

The text preprocessing step involves shallow part of speech tagging and parsing. Part of speech tagging for Ukrainian language was implemented in the project UGTag [6]. Although this analyzer lacks morphological disambiguation, we still can use it, as we need only shallow tagging and the information about the part of speech will be enough. Unfortunately there is no available parser for Ukrainian language, and the implementation of such a tool goes beyond this study. Thus, it was decided to create part-of-speech patterns for distinguishing a subject, a predicate and an object of each sentence.

Stemming and lemmatization will not be needed as the morphological analyzer already provides information about the initiative form of the word. This initiative form will be useful, when the desired word form is not in the dictionary. Then the word will be assigned the sentiment value of its initiative form.

At the stage of preprocessing the object of subjectivity will also be defined. Since we have chosen restaurant reviews as the domain for our study, the objects of subjectivity will be the names of the restaurants. The main methods to find these objects are looking for named entities (capital letters, foreign words, unusual combination of nouns, etc.), punctuation (such as quotes), and surrounding words (using examples collected from the corpus). Consider Figures 2 and 3 for examples of input and preprocessed data.

kvVItKA
14.09.2011
а я сьогодні була в Amigos, стейки і салати дуже смачні, мохіто не файне, але обслуговування хороше.

Figure 2. The input data example.

а/con я/pron сьогодні/adv **була/v** в/prep Amigos/unk
стейки/n і/con салати/n **дуже/adv смачні/adj**
мохіто/n **не/par файне/adj**
але/con обслуговування/n **хороше/adj**

Figure 3. The preprocessed data example.

Figure 3 shows an already tagged message with a defined object of subjectivity and shallow parsing done, meaning that the sentence is divided into clauses and a predicate in each clause is defined (shown in bold).

It is worth mentioning that the sentiment of each clause will be defined separately, as one complex sentence may contain text fragments with opposite sentiments.

Getting to the next step, we must say that there is no available sentiment dictionary for Ukrainian language, which defines the problem of creating one. Sentiment dictionaries are usually created with the use of ontologies or sentiment-annotated corpora. In order to generate a sentiment dictionary, we created a sentiment-annotated corpus of restaurant reviews in Ukrainian (600 annotated reviews) using the Gate 7.0 environment [7]. Restaurant reviews in Ukrainian, which became the basis of the corpus, were taken from a popular forum http://posydenky.lvivport.com/ and a website on all kinds of reviews http://v.lviv.ua/. These websites were chosen because of the big number of reviews that meet the chosen topic, and because the majority of the reviews on these websites were written in Ukrainian. An annotation scheme for the corpus was developed with the help of CREOLE package. The developed annotation scheme has the following structural units: nickname, date, review, citing, sentence, clause, target, word, and a url-address. Each unit together with its attributes was described in a separate xlm-file.

On the basis of the sentiment-annotated corpus we managed to get the main part of sentiment dictionary. It is worth mentioning that the dictionary contains only nouns, verbs, adjectives and adverbs. Each word is assigned a sentiment (positive or negative) and emotion, if such is present (we used basic human emotions by P. Ekman: anger, disgust, fear, joy, sadness, surprise). In future we are going to semiautomatically extend this dictionary with the help of the dictionaries of synonyms and antonyms. Words that play the role of invertors (”no”, ”not”, etc.) and amplifying words (”very”, ”extremely”, ”really”, etc.) are processed, too. The dictionary is going to be extended with the help of the dictionaries of synonyms and antonyms.

The sentiment dictionary is further used to assign sentiment values to every word of the review. Consider Figure 4 for an example of sentiment-tagged clauses. The subjective vocabulary has been highlighted.

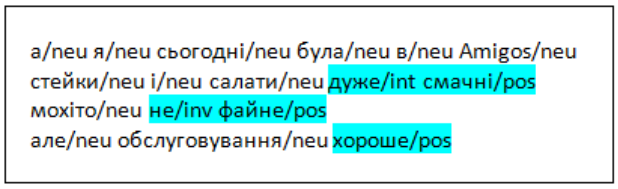

Figure 4. The sentiment tagged review.

The next step is writing the rules. The first part of the rules refers to the words-invertors. When such a word is found, the sentiment of the next word or set of words (up to five words) is changed to the opposite. The second part refers to processing the amplifying words: if such a word is found, the sentiment of a positive clause is changed to very positive and the sentiment of a negative clause is changed to very negative. In the end, the sentiment of the clause is defined. This is done with the help of sequential composing of sentiments, used in the work [4] for English. The words are added one by one, starting from the predicate, and the sentiment is defined depending on the main word in the text fragment. Consider Figure 5 for an example of sentiment compounding rules.

In this way it can be seen that applying a rule-based approach provides more information about the author's attitude toward a certain object, than all kinds of statistical approaches do, as in case of applying a statistical approach, we would just get a general positive sentiment, and that is all.

The final step is creating the result representation tool. The results are going to be presented

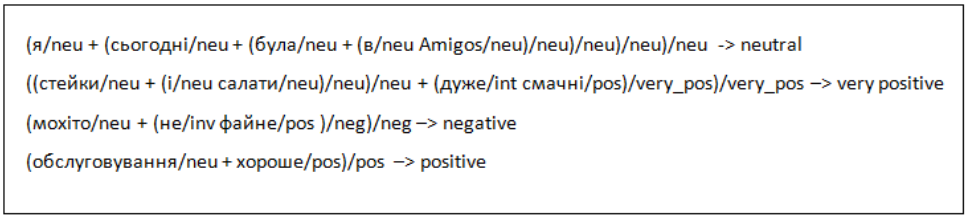

Figure 5. An example of sentiment compounding for the given example.

in a table with positive, negative, very positive and very negative clauses from every review. If a clause possesses any specific emotion, it is stated, too.

The accuracy of the system is going to be measured on the test set with the help of precision and recall formulas for each category (positive, negative, very positive, very negative and neutral clauses):

$$Precision = tp/(tp + fp), \tag{1}$$

$$Recall = tp/(tp + fn), \tag{2}$$

where, for example, for positive clauses, tp (true positives) will represent the number of positive clauses that have been actually assigned a positive category by the system;

fp (false positives) - the number of clauses of other categories that have been assigned a positive category by the system;

fn (false negatives) - the number of positive clauses that have been assigned some different category by the system;

In this way the first two steps of sentiment analysis algorithm have been successfully implemented. The third and the fourth steps, though, still require a lot of work. We expect the system to have more than 90 percent precision.

Conclusion

Having conducted a detailed analysis of recent research in the field of sentiment analysis, we have managed to define the optimal algorithm for implementing sentiment analysis for reviews in Ukrainian. A sentiment-annotated corpus and a sentiment dictionary for the domain of restaurant reviews have been created. The rules of sentiment compounding are being developed.

References

[1] Pazelskaya A. Method of defining emotions in Russian texts / A. H. Pazelskaya, A. N. Soloviov. – Computational linguistics and intellectual technologies: vol. 10 (17). – Moscow: RHHU, 2011. – pp. 510-522.

[2] Kan D. Rule-based approach to sentiment analysis at ROMIP 2011 / Dmitry Kan. – Available from: http://www.slideshare.net/dmitrykan/rule-based-approach-to-sentiment-analysis-at-romip-2011

[3] Kasper W. Sentiment Analysis for Hotel Reviews / Walter Kasper, Mihaela Vela. – Proceedings of the Computational Linguistics-Applications Conference. – Jachranka, Poland: Polskie Towarzystwo Informatyczne, Katowice, 10/2011. – pp. 45-52.

[4] Moilanen K. Multi-entity Sentiment Scoring / Karo Moilanen, Stephen Pulman. – Proceedings of Recent Advances in Natural Language Processing (RANLP 2009). – Borovets, Bulgaria, September 14-16 2009. – pp. 258-263.

[5] Python NLTK Demos for Natural Language Text Processing. – Available from: http://text-processing.com/demo/sentiment

[6] UGTag – a morphological tagger for Ukrainian language. – Available from: http://www.domeczek.pl/ polukr/parcor/
[7] Using GATE Developer. - Available from: http://gate.ac.uk/sale/tao/splitch3.html#chap:developer
[8] Agrin N. Developing a Flexible Sentiment Analysis Technique for Multiple Domains / Nate Agrin. - 2006. - Available from: http://courses.ischool.berkeley.edu/i256/f06/projects/agrin.pdf

Authors

Mariana Mykhailivna Romanyshyn — the 2nd year post-graduate student, department of Computer-Aided Design, Lviv Polytechnic National University, Lviv, Ukraine; E-mail: *mariana.scorp@gmail.com*