UPPSALA
UNIVERSITET

Evaluating and Improving an Automatic Sentiment Analysis System

Viktoriya Kotelevskaya

Uppsala University
Department of Linguistics and Philology
Språkteknologiprogrammet
(Language Technology Programme)
Bachelor's Thesis in Language Technology, 15 credits

October 14, 2011

Supervisors:
Mats Dahllöf, Uppsala Universitet
Bengt Dahlqvist, Uppsala Universitet
Beáta Megyesi, Uppsala Universitet
Alexandra Stålnacke, Find Agent AB

Abstract

The purpose of this thesis is to improve OpenAmplify – a system for automatic sentiment analysis by analysing OA's weaknesses and problematic areas and then modifying the resource files and lexicons by adding new linguistic items and improving the system's rules.

The performance of OA was first evaluated and the collected data was compared to opinions of human judges, allowing to identify and analyse the problematic areas and shortcomings of the system. The analysis lead to modifications of OA: idioms and missing expressions were added to the lexicons and the resource files were extended by some linguistic rules.The results of the evaluation showed that agreement between OA and human judges had increased from 41% to 55%, after modification of the system.

Acknowledgements

I am heartily thankful to Alexandra Stålnacke, Mats Dahllöf, Bengt Dahlqvist and Maria Milusheva for guidance and support. I would especially like to thank Beáta Megyesi for her encouragement and her accurate comments which were of critical importance for this paper. Also, many thanks to Dora La Flora Czifra for proof-reading this thesis. Lastly, I offer my regards to all of those who supported me in any respect during the completion of the project.

Contents

1 Introduction

As time goes by, more and more people communicate through the Internet. Millions of Internet users discuss their common interests with other people using social networks, blogs, forums, etc. People also often consult other World Wide Web users before buying a new product or eating at a certain restaurant. Many Internet users choose to express themselves in the form of ratings, reviews, discussions on forums and blogs. There is a problem though: getting that information and analysing it is certainly time-consuming and expensive. The solution might be automatic sentiment analysis tools.

The main task for automatic sentiment analysis tools is to define the polarity of a sentence or a piece of text. To define the polarity means to classify a text as positive, negative or neutral. The goal is to produce useful data and present it in a clear way so that one can get a general idea about people's opinions. For instance, such information may come in handy when a company needs to know what customers think about a certain product.

The common ways to gather public opinions used to be questionnaires and opinion polls, but as the Internet grows there is no need to do it that way - there is more than enough information on the Internet. Social networks such as Twitter and Facebook provide countless reviews, opinions, ratings, etc. And the popularity of these networks has helped to increase the interest for tools built especially for the purpose of sentiment analysis.

This study's main aims are to evaluate and improve a specific part of a system for the automatic sentiment analysis OpenAmplify - the Polarity module and understand the problems and shortcomings of the system. This is done by evaluating the performance of the existing system, identifying and analysing the problems and modifying the module's rules and resource files so that the modifications result in improved performance the whole system on the same data.

The Polarity module is an important part of the OpenAmplify system, which is responsible for assigning a polarity value to each word and then calculating the polarity score of the whole sentence or text by applying rules to it. The focus of the study presented in this paper lies on sentiment analysis on sentence-level. It means that only the Polarity module's ability to classify single sentences, rather than pieces of texts has been studied. The thesis work has been carried out in cooperation with FindAgent AB in Stockholm.

2 Sentiment Analysis

Sentiment analysis is the task of identifying positive and negative opinions, emotions, and evaluations (Wilson et al., 2005). Many of the sentiment analysis tools on the market today are commercial, but some are free, for instance Tweetfeel and Twitrratr. These tools search for Twitter posts that contain some input word and present (in a graphic way) how the majority of Twitter users feel about it. For example, if one enters the name "Michael Jackson" in the search field, Twitrratr shows that there are 136 positive, 69 negative and 1295 neutral "tweets" that contain that name. Most of the negatively marked "tweets" include words like "hate", "bad", "terrible", etc., while the positively marked "tweets" contain words like "love", "kind", "great" etc. One can observe that the "tweets" that belong to the neutral category do not include strongly positive or negative words. However, just reading through the statements makes one realize that many of them should have been categorized either as positive or negative. For example the following "tweet"

We danced to Michael Jackson at the top of the Gherkin last night. Sorta hung over, but completely worth it.

should have been classified as positive. But as there are no determinedly positive or negative words presented in the statement, Twitrratr marked it as neutral.

An essential part of "simple" tools like Tweetfeel and Twitrratr is lexicons that consist of words and phrases with either positive or negative polarity. Such tools do not seem to take into account the fact that polarity score of a strongly positive or negative word might change when that word is used in a different context. The lexicon simply consists of words like "beautiful", "nice" and "good-looking" (positive prior polarity)[1] and negative – "horrible", "disgusting" and "awful" (negative prior polarity) (Wilson et al., 2005).

Bartlett and Albright (2008) describe two main approaches to building automatic sentiment analysis tools:

- Linguistic approach
- Statistical approach

The linguistic approach relies on disambiguation using background information such as sets of rules and vocabularies. Thus, a system that is built in accordance with the linguistic approach normally contains lexicons, which consist of words and their polarity values (positive, good, negative, bad etc.).

[1] Polarity of a word taken out of context.

Another important part of such systems is sets of rules that help to produce more accurate results.

Advanced tools for automatic sentiment analysis include rules that make the process of defining polarity of statements more accurate. Those rules might be on many levels - from very simple to very advanced. Here are some elementary rules that might be included in a tool for automatic sentiment analysis:

- Adjectives that are separated by "and" have the same polarity (*He is handsome and rich.*)
- Adjectives that are separated by "but" have different polarity (*He is rich but mean.*)
- Synonyms have the same polarity (good, honorable, respectable etc.)
- Antonyms have different polarity (good = positive vs. evil = negative)
- Prefixes im- and dis- negate sentiment of a term (*possible* vs. *impossible*; *like* vs. *dislike*)

These rules are very basic, but more complex and effective rules are crucial for the performance of such tools. The system that has been investigated in this study - OpenAmplify was built using the linguistic approach: it consists of modules, sets of rules and resource files.

The statistical approach is based on methods of finding frequencies of occurrences of terms[2](Bartlett and Albright, 2008). There are several methods that can be used when building a tool for automatic sentiment analysis in accordance with the statistical approach. Boiy et al. (2007) examined and compated three classic machine learning methods: Maximum Entropy, (ME), Naive Bayes Classifier (NBC) and and Support Vector Machines (SVM). The study showed that methods that rely on SVM tend to work relatively good (85.45%), while methods based on NBC classifier (81.45%) and ME (84.80%) are somewhat poorer. Sometimes several techniques are combined in hope to achive better results (see for example Nigam et al. (2000)).

Research progress in the field of sentiment analysis is very necessary and often successful. For instance, Täckström and McDonald (2011) describe in their latest paper a semi-supervised latent variable for sentence-level sentiment analysis. The proposed method combines fine-grained (scarce, informative) and coarse-grained (abundant, but less informative) supervision in order to improve sentiment analysis on sentence-level. The results show that the proposed model performs significantly better compared to the existing ones.

Another interesting research that was conducted by Tan et al. (2011) shows how information about social relationships of Internet users can be used to improve sentiment analysis on user-level. The authors use Twitter as a source of experimental material and create models induced from the Twitter networks. The study established that including information about people's social relationships can result in significant sentiment classification improvements.

[2] An example of a tool built in accordance with this method is Clarabridge (http://www.clarabridge.com/)

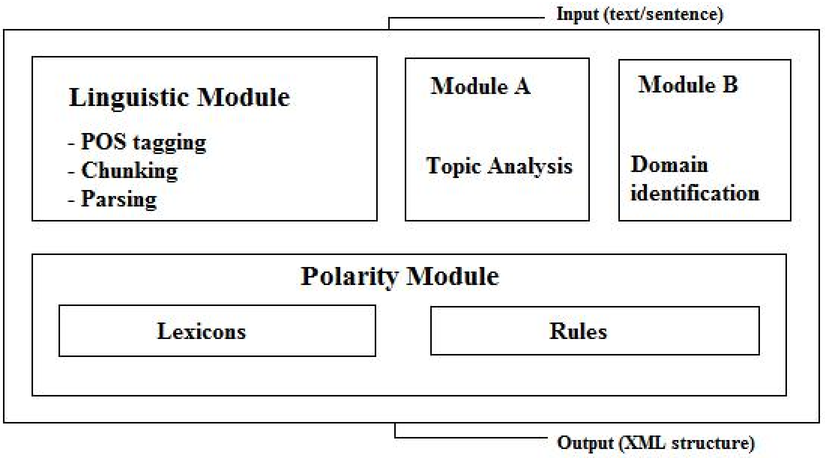
3 OpenAmplify

OpenAmplify is a web service for analysing the basic meaning of a text. It identifies the significant topics, brands and people as well as the attitudes associated with them in the context. OpenAmplify can also determine the author's style of writing (for instance the degree of formality) and his/her attitude towards the text. The service can even establish the author's properties such as age, gender, education level etc. The results of the analysis are called "signals" and can be presented either in a graphic way or written to an XML file.

OpenAmplify is a semi-free web service, which means that one can use it up to 1000 transactions per day without paying. In order to do that one has to register and get a special API (Application Program Interface) key. The free version of OpenAmplify can even be used for commercial purposes. At the moment the only available language for analysis with OpenAmplify is English.

The system consists of hundreds of resource files (lexicons, lists of words and phrases), sets of rules and modules. As Figure 3.1 demonstrates, each module is responsible for one or several tasks. The Linguistic Module includes a POS tagger, which assigns part-of-speech tag to every input word. Module A analyses input text and selects top topic of every input sentence as well as general topic of the whole input text. Module B identifies and classifies domain of analysed text. The focus of this paper, however, lies on the Polarity module, which is discussed in the next section.

Figure 3.1: The OpenAmplify system's architecture



3.1 The Polarity module

The Polarity module is responsible for automatic sentiment analysis, which makes it a central part of the OpenAmplify system. The module consists of lexicons and numerous disambiguation rules.

In order to analyse a sentence, the module first assigns a polarity value (1 = positive, 0 = neutral; default, -1 = negative) to each word of a sentence, in accordance with the module's lexicons (i.e. "bad"= -1, "good" = 1. etc.). A sentence is then divided into phrases and each phrase gets yet another polarity score, which depends on words that phrases consist of. If the structure of a sentence looks a certain way, one (or several) rules that match the structure might be applied to it. Depending on the case, that might change the polarity score of some of the phrases. The final result is presented as a number between 1 and -1.

The following step-by-step example illustrates how the sentence "I hate beautiful flowers" is analysed by the Polarity module:

(1) Each word of the statement is assigned a polarity score according to the modules's polarity score lexicons. For instance, in our example the word "flower" has the polarity score of 0.4 in module's lexicon. Words that do not affect the polarity of a statement (for instance function words) get the default score of 0. If some word is missing from the resource files, it is also given the neutral polarity score of 0.
I: personal_pronoun: 0
hate: verb: -1
beautiful: adjective: 1
flower: noun: 0.4

(2) The first set of rules is applied to a statement. Those rules divide a statement into phrases, after which each phrase is given an individual polarity score. The polarity scores of phrases are based on the words that each phrase contains. The result is phrases (instead of just words) with polarity scores. In our example, there is a phrase "beautiful flowers", which consists of two words (beautiful:adjective:1 + flower:noun:0.4). The module recognizes it as a noun phrase and its polarity score is 1.4 (1+0.4).
<NP:0>I</NP><VP:-1>hate</VP><NP:1.4>beautiful flowers</NP>

(3) If a statement has a certain structure, some rules from the second set of rules which apply to phrases might be triggered. For example, the following Rule X applies to statements, which consist of a noun phrase (subject) and a negative transitive verb phrase, which is followed by a positive noun phrase (object).
<clause>NP(sub) + VP(neg_T) + NP(pos)(object)</clause>

(4) When dealing with statements with particular structure, the system assigns the object's score to the subject and multiplies it by -0.5. Also, according to the rule the object's score is multiplied by -1.
*Rule X: **If in** <clause>: <NP_sub>*
***is followed by** <VP_neg_T>*
***followed by** <NP_obj_pos>*

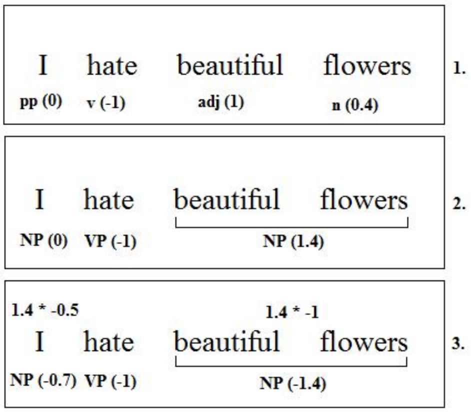


Figure 3.2: Sentence analysis by the Polarity module.

assign *<NP_sub>* ***the score of*** *<NP_obj_pos> -0.5* ***and***
multiply *<NP_obj_pos>* ***by*** *-1*

(5) Since the structure of the statement that we are considering is NP(sub) + VP(neg_T) + NP(pos_obj), the rule has to be applied to our statement:
*<NP>**I**</NP> = 1.4 * -0.5 = <NP -0.7>**I**</NP>*
*<VP -1>**hate**</VP>*
*<NP>**beautiful flowers**</NP> = 1.4 * -1 = <NP -1.4>**beautiful flowers**</NP>*

(6) The final output is the following:
Top topic: flower
Polarity: Negative (-1)

The final result is presented as a number between 1 and -1. So, the final polarity score of the statement is very different from the score that was assigned to it in the beginning, using only the polarity lexicon. In addition, the statement is now divided into two noun phrases and a verb phrase, rather than four separate words.

4 Evaluation Methods

In order to evaluate the performance of the OpenAmplify system, a corpus consisting of ca 5000 statements was composed. Each sentence contained the word "iPad". The focus of the study laid on extraction of human's opinions about "iPad". The question that human raters had to answer was:

Do you think the author's tone/attitude toward the "iPad" is positive, negative or neutral in the following sentences?

Sentence A
Sentence B
Sentence C

The sentences that the corpus consisted of were taken from four different source categories:

- International/Local press (Newspapers)
  ca 1000 sentences
- News agency/News service (Press releases)
  ca 1000 sentences
- Web magazine (Specialist press, online technology magazines etc.)
  ca 1000 sentences
- Rank X (User generated content: blogs, forums, comments, etc.)
  ca 2000 sentences

As one can see, the "Rank X" category contained twice as many sentences as the other categories. The reason for that is that this particular category consisted of UGC (User Generated Content)[1] and was bound to contain more sentiment than the other categories.

The polarity of the topic "iPad" in each sentence was identified using OA. Each statement was also rated by five Amazon Mechanical Turk judges. AMT is a service for crowdsourcing that makes collecting human opinions very easy and has been used in this study as an alternative method of collecting human opinions.

The maximum amount of time that a worker could spend on a task was five minutes. Each task required rating three sentences. For completing one task each worker got paid 0.01$. The participants were encouraged to leave

[1] Material produced by Internet users.

any thoughts or comments that they might have had about the tasks . Also, it was specified that only the AMT users from the United States were allowed to participate in this particular study. That way it was more likely that the native language of participants was English.

One of the main observations that could be made about the collected data was whether human judges agreed with each other when rating a certain statement i.e. the interrater agreement (also called interrater reliability) and the strength of association between ratings (correlation). All the statistic calculations were carried out by using the tools provided by VassarStats[2].The metrics are described in section 4.2.

4.1 Interrater Agreement

According to Stemler (2004), "across all situations involving judges, it is important to estimate the degree of interrater reliability, as this value has important implications for the validity of the study results". The author claims that there are three main categories when it comes to calculating interrater reliability, and each of the existing methods belongs to one of them. These categories are:

- Consensus estimates
- Consistency estimates
- Measurements estimates

The first category "Consensus estimates" is based on the idea that judges have a common understanding about the rating scale. Some methods that belong to that category are Cohen's kappa statistics and the figure of percent-agreement.

The "Consistency estimates" category includes for instance Pearson's correlation coefficient and Spearman's rank correlation coefficient. The common ground of that category is that judges do not have to interpret the rating scale exactly the same, as long as each and every one of them is consistent when it comes to classifying the items according to their own individual definition of the scale. That way, judges are predictable in the way they interpret the scale, hence, one can anticipate how they apply rating scale categories individually.

The "Measurements estimates" category includes methods that take into account all the available information from judges. Each individual judge is considered to be a unique source of information. Thus, a score of "2" from two different judges may be considered to be of different weight. For example, 2 points from Judge A might be looked at as being closer to 1 point, because he or she has a tendency to give lowered ratings; while 2 points from Judge B may be seen as leaning more towards 3 points, because he or she is a very strict judge.

For this study, only methods of calculating interrater agreement that belong to the first and the second category (consensus estimates and consistency estimates) were implemented. Methods of the third category (measurements

[2] www.faculty.vassar.edu

estimates) were not used as it would be rather difficult if not entirely impossible to analyse and study every rater's personal judging habits and style.The methods were chosen considering their usefulness for the current study.

4.1.1 Correlation

A correlation measures strength of association between variables. There are numerous ways to calculate correlation between two variables. For example, Pearson's product-moment is used to examine linear relationship between two variables in form of a descriptive statistic measure (Chen and Popovich, 2002).

Besides the usual Pearson's product-moment correlation for measuring relationship between two variables there are several different methods for calculating correlation depending on the data at hand. For instance, the choice of method for calculating correlation depends on the measuring scale of the data. There are different kinds of scales:

- Ratio scale (continuous, natural zero; e.g. length, weight)
- Interval scale (continuous, no natural zero; e.g. temperature, date)
- Ordinal scale (category, order; e.g. ratings, ability)
- Nominal scale (category, no order; e.g. gender, color)

4.2 Metrics

4.2.1 Percent-agreement

In order to be able to see if human judges agreed with the OpenAmplify system's polarity ratings, human answers had to be compared to each other. Since there was just one single rating provided by OA against five human ratings for each statement, the following method was chosen.
A sentence was rated by five Amazon Mechanical Turk workers.

If its not checked, no apps will be synced to your iPad.

The following answers were gathered:

AMT judges (X):
Mturk 1: Negative
Mturk 2: Neutral
Mturk 3: Negative
Mturk 4: Positive
Mturk 5: Negative

OpenAmplify (Y):
Negative
As one can see, the rating given by the majority of human raters (X) was negative. Thus, "negative" is considered to be the correct polarity for this statement.

X = Negative
Y = Negative

So, when calculating agreement between human raters and the investigated system using this method, there was a single human rating (the rating given by the majority of human judges (X)) against one machine rating (Y). Thus, human raters and OpenAmplify agreed with one another if and only if at least three (of five) workers' ratings were the same as the one given by OA.

In some cases, however, the human ratings were very random:

AMT judges (X):
Mturk 1: Negative
Mturk 2: Neutral
Mturk 3: Negative
Mturk 4: Neutral
Mturk 5: Positive

OpenAmplify (Y):
Neutral

According to the chosen method, even though two of five human ratings were the same as the rating of OA (neutral), the agreement did not take place. Thus, OA's rating was considered to be false.

4.2.2 Kappa statistics

Kappa statistics is one of the most common methods for calculating interrater agreement. The advantage of kappa statistics is that this particular method is designed to take into account the effects of chance (Ludbrook, 2002). It provides a quantitative measure of magnitude of agreement between raters. Kappa measures interrater agreement on a scale from -1 to 1. If the result of calculations is equal to 1, agreement is considered to be perfect. Result equal to -1 means that there might be potential systematic disagreement between judges. Kappa value of 0 equals agreement expected by chance (Viera, 2005).

For this study, it was necessary to calculate a so called *weighted kappa value*. Weighted kappa value accounts for the degree of disagreement between subjects. Note that weighted kappa is only appropriate for calculations with ordinal data, which makes that method appropriate for the current study (Li, 2004). There is a number of methods to calculate weighted kappa. One of the most common methods is kappa with quadratic weighting. Kappa with quadratic weighting is calculated the following way:

$$k_q = 1 - ({}_i{-}_j/k - 1)^2$$

where
i - *j* is the difference between the row category on the scale and the column category on the scale for the cell concerned; and *k* is the number of points on the scale.

(Sim and Wright, 2005)

4.2.3 Spearman's rank correlation

Since the data in this study is ordinal, a suitable method to measure the correlation is Spearman's, which is often used in cases with ordinal data. Hayslett (1981) demonstrates the way Spearman's rank correlation coefficient is calculated:

$$r_s = 1 - [6(x_i - y_i)^2]/[n(n^2 - 1)]$$

where
r_s denotes the Spearman's rank correlation coefficient;
n = number of pairs of observation;
x_i = rank of x_i; and
y_i = rank of y_i.

Spearman's rank correlation coefficient measures the dependence between two variables utilizing only the ranks. According to Hayslett (1981) ”a value of rs equal to 1 indicates perfect agreement between the ranks of x and y. A value of rs equal to -1 indicates that the ranks of y are in exactly the opposite order as the ranks of x. A value of rs near zero indicates that x and y are independent” (Hayslett, 1981).

4.2.4 Cramér's V statistic

Cramér's V is a statistic calculation measuring correlation between two categorical variables given in a contingency table. According to Agresti and Finlay (2009), it is an easy way to simply summarize association in the entire table by a single number. In order to determine Cramér's V, one has to use the following formula:

$$V = [x^2/(n(k - 1))]$$

where
x_2 stands for chi-square;
k is the number of rows or columns, whichever is less;
n is the number of rows or columns in the table.
(Agresti and Finlay, 2009)

4.2.5 Chi-square distribution

According to Murray (2006), chi-square distribution is "the probability distribution resulting from adding the squares of independent standard normal random variables. The number of squared standard normals in the sum is called the degrees of freedom of the chi-square distributed variable". The final product of calculations is a single number, that combines the difference between the expected and the observed values, commonly called x2 in probability statistics.

For this study, chi-square distribution had to be calculated in order to determine Cramér's V.

A general formula for calculating chi-square distribution for contingency table is:

$$x_2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where
^{r} is the number of rows in the contingency table;
^{c} is the number of columns in the contingency table;
O_{ij} is the observed values for each cell (row$_i$, column$_j$);
E_{ij} is the expected values for each cell (row$_i$, column$_j$).
(Hudson, 2000)

5 Evaluation of OpenAmplify

After all the necessary data was collected, some conclusions about the ability of the OpenAmplify system to define polarity of statements could be drawn. By comparing the ratings given by human raters and the system, one could establish that the frequency of agreement was 41% (agreement in 2026 of 4936 cases). Calculations also revealed that Spearman's correlation coefficient was equal to 0.1861 = weak, positive correlation, Cramér's V was 0.1624 = slight association between variables and the observed kappa was 0.1845 = slight agreement.

Table 5.1 shows the results of the individual evaluations of the source categories that the research corpus consisted of. The agreement rate for the different sources was between 40-42% and the Spearman's, Cramér's V and kappa value were very similar to each other. One can see that there were no significant differences in individual results of the four different source categories. Thus, it can be concluded that the performance of the OpenAmplify system does not depend on the type of input text.

Table 5.1: The results of original system's evaluation and individual statistics for the source categories.

	Agreement	Spearman's	Cramér's V	Kappa
Original OA	41%	0.1861	0.1624	0.1845
International/Local press	42%	0.202	0.1959	0.1977
News agency/News service	42%	0.185	0.1562	0.1818
Web magazine	40%	0.176	0.1512	0.178
Rank X	41%	0.186	0.159	0.1834

Table 5.2 shows that OA often predicted neutral opinion, rather then positive or negative. The first column of the table represents the ratings provided by OA, while column 2 contains the judgments given by the majority of human raters. Since we are considering human answers to be the golden standard (i.e. the right polarity of the statements), one can conclude that in most cases (395), OA rated positive statements as neutral.

OpenAmplify	*Human Judges*	*Number of cases*	*Percent of cases*
Neutral	Positive	395	80%
Negative	Positive	63	13%
Neutral	Negative	23	5%
Positive	Negative	11	2%
Positive	Neutral	2	0%
Negative	Neutral	1	0%
Total		**495**	

Table 5.2: Mistake tendencies.

5.1 Problems and shortcomings

In order to understand the weak areas of the Polarity module, the problematic cases needed to be detected and analysed. Thus, the cases where all five human judges gave the same rating to a certain statement (negative, neutral or positive), while OA's rating was different were selected and looked at more closely.

There were 495 cases in which the rating of human judges were the same, but dispersed with the rating given by OA. By going through those cases and analysing them one by one, we could divide them into five different groups. Each group was based on the type of error that lead to wrong analysis or the method that was chosen to solve the problem.

- **Lexicon enhancements**
  Problem: Some words, fixed expressions or idioms are missing from the resource files.
  Example: *game-changer; jailbreak; toodle away, meet [someone's] needs; follow the same curve; crash competition.*

- **Rules**
  Problem: The specific structures of some sentences lead to wrong analysis.
  Example rule: *If the top topic of a statement is a proper noun, which is followed by the verbs "rule(s)" or "suck(s)", which is in turn followed by exclamation point, the polarity of such statements should be strongly positive (1) or negative (-1) respectively.*

  iPad rules! (positive polarity)
  iPad sucks! (negative polarity)

- **Context-based polarity**
  Problem: Either human knowledge or context is needed in order to define polarity of these sentences.
  Example: *The iPad is not a computer at least not to these folks.*

- **Anaphor resolution**
  Problem: According to the current rules, only polarity score of words that belong to top topic's clause are taken into account when analysing anaphoric expressions.
  Example: *I did buy the* **iPad** *at last and* **it** *made my life so much better.*

- **Other**
  In a number of problematic cases it was rather difficult to figure out why human raters disagreed with OA.
  Example: *I just got my iPad.* (OA's rating: neutral; Human rating: positive)

Category	*Number of cases*	*Percent of cases*
Lexicon enhancement	175	35%
Context-based polarity	95	19%
Anaphor resolution	32	7%
Rules	23	5%
Other	170	34%
Total	**495**	

Table 5.3: Mistake categories.

Table 5.3 sums up these results and demonstrates how many cases each group contained. As one can see, the "Lexicon enhancement" category contained the greatest number of cases (175). This category consisted of statements that got incorrect polarity ratings because of words, fixed expressions or idioms in those statements. Those items were not a part of the Polarity module's lexicons, which lead to wrong analysis. The "Context-based polarity" category contained 95 statements that got wrong polarity rating because the system did not have the same ability to analyse sentences as human judges did (the lack of human knowledge). The "Anaphor resolution" category included 32 sentences containing anaphoric expressions, which OA failed to identify. The last category ("Rules") consisted of statements which were not correctly analysed because of their specific structures, which OA could not handle.

Some of the errors and shortcomings of the OpenAmplify system that were detected during this evaluation could be handled ("Lexicon enhancement" and "Rules"). However, the possible solutions of other types of errors were outside the scope of the current study (" Context-based polarity" and " Anaphor resolution").

6 Improvements

The analysis of the problematic cases lead to several suggestions about how the rules and lexicons of the Polarity module could be modified. The purpose of those modifications was to try to improve the shortcomings that were detected during the evaluation of the OpenAmplify system.

6.1 Lexicon Enhancements

The reason for the wrong polarity score of the statements that belonged to this category was that certain words, expressions or idioms were missing from the Polarity module's resource files. According to the system's rules, if a word is absent from the resource files, it is automatically given the polarity of 0, which means that it will not affect the polarity rating of the whole statement. On the other hand, if some idiom or fixed expression is not a part of the module's lexicon, every word of the expression contributes to the sentence's overall polarity score. That might lead to wrong results. During this study I extended the module's resource files with approximately two hundred new words, idioms and fixed expressions.

Some idioms that were added to the resource files are for example:

- skyrocket to the top (*And as the iPad continues its skyrocket to the top, they'll change their mind because nobody can ignore that large of a market.*)

- put the smack down (*Apple is putting the smack down on the iPad.*)

- put the kibosh on something (*The absence of an integrated video camera puts the kibosh on any hope of using the iPad for video chats.*)

6.2 Rules

As it has been stated in Chapter 2, effective rules are essential for the performance of tools for automatic sentiment analysis. This section presents some examples of the rules that have been added to the Polarity module's sets of rules during this study.

- *Rule 1: A proper noun, followed by the words "fan" and "lover" should be recognized as a fixed expression with positive polarity.*

  Example: I have always been a huge Mac fan.

  The study showed that a number of problematic cases included fixed expressions of the type: “Mac lover”, “Kodak fan” etc. Statements that contained such expressions usually got negative or neutral polarity rating from OpenAmplify, while all the human judges rated them as positive. One possible solution would be trying to add every single expression of this type to the lexicon. That would not, however, solve the general problem of similar cases, which are quite common in natural languages. A better way to solve this problem was to create a new type of rule (Rule 1).

- *Rule 2: If an auxiliary verb (X) is followed by a verb with negative polarity (Y), the polarity score of Y should be set to 0.*

  Example: Nokia didn't fail with Gizmo5 because of a lack of users but rather because they got cold feet and didn't promote it because they feared carriers.

  The statement above got positive rating from OpenAmplify, but it should have been classified as negative. The problem here is that the negative verb "fail" is negated by the auxiliary verb "didn't". In cases like that, the negative polarity score of the second verb is changed to positive. That often results in positive rating of the whole statement. A possible solution is to change the rule so that in such cases, auxiliary verbs like ”didn't”, ”shouldn't” etc. do not negate the score of the following verb, but rather ”reset” its polarity score.

- *Rule 3: Polarity score of sentences that start with ”never again”, followed by an auxiliary verb should be changed to negative.*

  Example: Never again will I buy anything made by the Apple company.
  Never again will I go to this terrible restaurant.

  Statements that start with ”never again” are a very common way to express one's negative opinion or emotions about something. The study showed that such sentences often got neutral rating. This rule will prevent the wrong analysis.

- *Rule 4: The imperative form "do not"/"don't" in the beginning of a statement should have a negative impact on the statement's overall polarity score.*

  Example: Don't buy Pepsi in the new can.
  Do not even consider buying an iPad.

A number of cases that were falsely classified by OpenAmplify started with "don't"/"do not". The proposed rule might cause some false alarms, but since statements of that kind are very common, it is more likely that the rule will have a positive effect on the performance of OpenAmplify.

6.3 Results

The study impelled some suggestions regarding new rules and lexicon enhancements that might be useful for improving the Polarity module. Thus, the author carried out another evaluation to establish whether the proposed modifications of OA would have the desired effect by looking at the improvements on the validation set. The modified version of OA was used to collect machine ratings, which were then compared to human ratings. The results of that evaluation showed that changes and modifications that were suggested after the original system's evaluation were effective and benefited the performance of the OpenAmplify system on the same data.

Table 6.1: Improvement in the evaluation results.

	Agreement	Spearman's	Cramér's V	Kappa
Original OA	41%	0.1861	0.1624	0.1845
Modified OA	55%	0.245	0.2063	0.2461
Improvement	14%	0.0589	0.0439	0.0616

The evaluation results of OA's modified version have improved compared to the evaluation results of the original system. The correlation values were now 0.245 (Spearman's), 0.2063 (Cramér's) and 0.2461 (kappa). Most importantly, the percent agreement value had increased and was now 55%, which means that the investigated system reached agreement with human raters in more than half of all cases.

Even though the numbers have not improved significantly, the Spearman's correlation rate has increased by 0.0589, the Cramér's V value by 0.0439 and the kappa value by 0.0616. Since the value of 1 means 100% agreement, there has been a change in the positive direction. Thus, one can establish that according to the evaluation carried out on the same data, the proposed changes in the system's resource files, lexicons and rules have improved the performance OpenAmplify system on the same corpus.

7 Discussion and Future Improvements

The evaluation of the original system showed that the performance of the current version of OpenAmplify was far from perfect. For instance, agreement between human judges (the rating of the majority of raters for every statement was considered to be the golden standard) and the system was 41% (see Table 5.1). The evaluation also revealed that there were no major differences in results of the different source categories that the research corpus consisted of: International/Local press, News agency/News service, Web magazine and Rank X. The first three categories of the research corpus consisted of formal sentences (technical magazines, news articles etc.), while the last category (Rank X) contained sentences from forum/blog entries, comments etc. (user generated content). Hence, one can establish that OpenAmplify's ability to define the polarity of statements does not depend on the formality level of texts that the system has to deal with (see Table 5.1).

One of the most interesting observations was that in the absolute majority of the problematic cases (80%), OpenAmplify classified statements as neutral, while human judges rated those statements as positive (see Table 5.2). Thus, it can be established that the system tends to miss the presence of polarity, rather than identify "false" polarity.

The results of that evaluation allowed for detection and study of the problematic cases, which lead to a number of suggestions about how the polarity module could be improved. The analysis of the problematic cases resulted in modifications of the system's resource files and rules.

The problematic cases that belonged to the group "Lexicon enhancements" could be solved by adding the missing fixed expressions, idioms or words to the resource files. However, the sentences from the group "Rules" required a much closer analysis. The target was not only to correct the existing rules or add new ones in order to get the right results for these particular statements; but rather to create a new rule (or modify the existing one), so that similar statements would get the correct polarity scores in the future. At the same time, the modifications had to be precise enough, so that they would not cause many false alarms, and broad enough in order to enfold as many cases as possible.

The "Anaphor resolution" category consisted of statements, which contained anaphoric expressions. In cases like that, OpenAmplify only took into account the polarity score of the words that belonged to the top topic's clause and ignored the rest. This is one of the most serious and intractable problem that needs to be handled. This problem occurs in anaphoric sentences in which the top topic is a part of one clause, while the positive or negative description of the topic belongs to another clause. This problem will hopefully be solved in the near future, as the system's developers are currently working on the anaphor resolution.

The cases that belonged to the "Context-based polarity" category demonstrated yet another problem. The difficulties that concerned the statements belonging to this group could not be overcome. These sentences needed to be either analysed by human beings or required additional context in order to be classified correctly.

It is important to have in mind that in this study, an alternative way of collecting human opinions was chosen - a tool for crowdsourcing Amazon Mechanical Turk. AMT is a service that makes the interaction between a requester and human raters very easy and effective. However, AMT workers are no experts, their age and education level is not defined and the main reason why they are participating in studies like this one is most likely to get the reward that comes with each completed task. This might have affected the results of the study and the interrater agreement values in particular. Nevertheless, this way of gathering human judgments has been felicitously used in a number of previous studies (see for example Callison-Burch (2009); Snow et al. (2008) and Heilman and Smith (2010)).

Another factor that might have affected the results of this study was that the same study corpus was used for the evaluation of the modified version of the system, as for the original system evaluation. That was done in order to see if the changes made to the Polarity module's lexicons and rules lead to correct analysis for the problematic cases that were detected during the original evaluation on the same corpus. Further investigations would be necessary to establish whether the improvement also yield other texts and domains.

8 Conclusion

This study's main goal was to improve a module of OpenAmplify, a system for automatic sentiment analysis and understand the week areas of the system. This was done by evaluating the performance of the existing system, identifying and analysing the problems and modifying the module's rules and resource files so that the outcome would improve. The evaluation was carried out by gathering human ratings and comparing them with polarity ratings provided by the OpenAmplify system.

The evaluation of the system, which was carried out on the original study corpus, after the modification of the module showed that the results have improved (see Table 6.1). Agreement between the investigated system and human judges had increased to 55% (compared to 41% for the evaluation of the original version of the system) and so did the correlation and interrater agreement values.

The study revealed some problematic areas and shortcomings of OA. For instance, it could be established that in some cases the lack of human knowledge and context made it impossible for the system to correctly analyse sentences. Also, it was noted that OA often had difficulties when dealing with anaphoric expressions.

The evaluation also revealed that the ability of OpenAmplify to judge the polarity of sentences did not depend on the type of input text: both UGC and more formal statements (articles, reviews etc.) got very similar results. Finally, the study showed that OpenAmplify often made mistakes in a particular direction, namely neutral, which means that the system tends to miss the presence of polarity, rather than identify false polarity.

Lastly, it should be emphasized that the purpose of systems for automatic sentiment analysis is to assist humans and not to replace them. It is unreasonable to require one hundred percent accuracy from systems that have to deal with natural languages, which can be quite complicated even for humans. The scope of such terms as positive, neutral and negative is quite indistinct. People often criticize the performance of automatic sentiment analysis tools without understanding how challenging it can be to reach even 50% accuracy. On the other hand, automatic sentiment analysis is the future of search engines, which means that more attention will be paid to the field, which will inevitably lead to better performance of systems for automatic sentiment analysis.

Bibliography

Alan Agresti and Barbara Finlay. *Statistical Methods for the Social Sciences*. Prentice Hall, 4th edition, 2009.

Jake Bartlett and Russell Albright. *Coming to a Theater Near You! Sentiment Classification Techniques Using SAS Text Miner*. Cary, NC, 2008.

Erik Boiy, Pieter Hens, Koen Deschacht, and Marie-Francine Moens. *Automatic Sentiment Analysis in On-line Text*. 2007.

Chris Callison-Burch. *Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk*. Baltimore, Maryland, USA, 2009.

Peter Y. Chen and Paula M. Popovich. *Parametric and Nonparametric Measures*. Sage Publications, 2002.

H.T. Hayslett. *Statistics Made Simple*. William Heinemann, 3rd edition, 1981.

Michael Heilman and Noah A. Smith. *Rating Computer-Generated Questions with Mechanical Turk*. 2010.

Pat Hudson. *History by Numbers: an Introduction to Quantitative Approaches*. Bloomsbury USA, 2000.

Xier Li. *Kappa - A Critical Review*. 2004.

John Ludbrook. *Statistical Techniques for Comparing Measurers and Methods of Measurement: A Critical Review*. 2002.

Michael P. Murray. *Econometrics: A Modern Introduction*. Pearson Education, 2006.

Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. *Text Classification from Labeled and Unlabeled Documents Using EM*. 2000.

Julius Sim and Chris C. Wright. *The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements*. 2005.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. *Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks*. 2008.

Steven E. Stemler. *A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability*. 2004. URL `http://PAREonline.net/getvn.asp?v=9&n=4`.

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. *User-Level Sentiment Analysis Incorporating Social Networks*. 2011.

Oscar Täckström and Ryan McDonald. *Semi-supervised Latent Variable Models for Sentence-level Sentiment Analysis*. 2011.

Garrett Viera. *Understanding Interobserver Agreement: The Kappa Statistic*. 2005.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis*. Vancouver, British Columbia, Canada, 2005.