Sentiment analysis on conversational texts

Birgitta Ojamaa **Päivi Kristiina Jokinen** **Kadri Muischnek**

University of Tartu
{b14606, kristiina.jokinen, kadri.muischnek}@ut.ee

Abstract

This paper describes ongoing work related to the analysis of spoken utterance transcripts and estimating the speaker's attitude towards the whole dialogue on the basis of their opinions expressed by utterances. Using the standard technology used in sentiment analysis, we report promising results which can be linked to the conversational participants' self-evaluation of their experience of the interaction.

1 Introduction

One of the popular research topics in current NLP is the sentiment analysis or opinion mining on texts corpora. Sentiment analysis has its roots in natural language processing and linguistics, and it appeared as the field of study in the early 2000s (Pang & Lee (2004, 2005) on document polarity analysis), and has become more popular due to widespread Internet usage and the texts freely available online on social media (Liu 2012: 1-2).

Sentiment analysis or opinion mining deals with using automatic analysis to find sentiments, emotions, opinions and attitudes from a written text towards a subject. This subject may be a product, an organisation, a person, a service or their attributes (Liu 2012: 1).

Based on the words associated with negative, neutral or positive sentiments, the documents are classified into positive, negative and neutral categories, and ratings for various aspects of a given topic (restaurant, movies) can be predicted.

Challenges with the short informal texts concern their unconventional characteristics as text: they contain shortenings, abbreviations, spelling mistakes, etc.

One of the interesting questions that we aim to study in this short paper is how well the standard techniques of sentiment analysis can be applied to conversational data. Since the utterances in conversations are short and produced alternatingly by the participants, transcribed dialogue texts resemble tweets or short SMS messages rather than long documents. However, face-to-face conversations are unique in that they are highly situational, and utterance meaning is constructed jointly by the participants in the dialogue context. The goal of this project was to use sentiment analysis on the conversational texts and compare the linguistic results with participant's own description of the interaction.

It must be noticed that the sentiment analysis we describe in this paper is not to be mixed with the participants' emotion analysis. Our goal is to study, if the sentiment analysis tools can be applied to conversational data and extract sentiments (positive and negative attitudes) that may be mapped onto the speakers' experience of the dialogue as a whole.

The short student paper is structured as follows. We introduce our data in Section 2, and

discuss its cleaning and method in Section 3. We present results in Section 4, and discuss them with future prospects in Section 5.

2 Data

The texts used in this project are from the MINT (Multimodal INTeraction) project that deals with Artificial Intelligence and multi-modal agents (Jokinen and Tenjes 2012). One particular field where intelligent agents need a lot of development is the study of emotion and sentiment, not only in gestures, but language as well – for example in speech synthesis where it soon becomes important for an agent to learn different tones for communicating more effectively (Vainik 2014: 335).

The dialogues are first-encounter dialogues where the speakers are unfamiliar with each other and they are expected to make acquaintance with their partner. They are expected to describe their likings to the partner but not to start emotional arguments on due to social politeness rules.

Each utterance is a continuous vocalization by a speaker rather a grammatically "correct" sentence. We used the transcriptions of utterances in 23 dialogue files, altogether 2902 utterances. Although it might have been useful to divide the transcribed text according to the speakers, we did not do this due to the small amount of data.

Cleaning of the data included removing of the XML notation that was used in the transcriptions (made using Praat (Boersma 2001)), as well as the English translations of the text. In total, there were 2902 sentences or utterances. Since the Praat output texts are grouped by the speaker, the utterances had to be rearranged to display individual utterances by time.

The corpus is accompanied by self-evaluation of the participant's experience of the interactions. This is based on a questionnaire which the participants filled after each interaction, describing how well certain positive and negative adjectives (e.g. pleasant, stressful, interesting) describe their experience.

The method to clean the files was to:

1) clean some parts of code and all of the translated text manually;
2) remove code around the timestamps to sort the text using UNIX shell-script;
3) sort the text by timestamps using shell-script;
4) remove the remaining code using shell-script, leaving only text;
5) text segmentation using a Perl script[1]
6) morphological analysis/disambiguation using Filosoft's t3mesta in shell-script[2].

The final result of the cleaning was text separated into sentences by the markers <s> </s> and individual wordforms on each line. What was left is displayed in Figure 1. The utterances are annotated using tags <s> and </s>. A row begins with the word-form as it was used in text, followed by its lemma and inflectional endings, separated from the lemma by +. Then come the part of speech tags and morphological categories between double slashes //. English glosses for every word-form are added in the end of the row, translation of the whole utterance in the end of every utterance.

```
<s>
nüüd nüüd+0 //_Y_ ?, //  now
me mina+0 //_P_ pl n, // we
peame pida+me //_V_ me, //
must-1.pl
rääkima rääki+ma //_V_ma, //
speak-inf
</s>
'We must speak now'
```

1 Provided by Kaili Müürisep

2 www.filosoft.ee

```
<s>
*naer* naer+0 //_S_ sg n, //
laughter
</s>
'laughter'

<s>
jaa jaa+0 //_D_ // yes
</s>
'yes'
```

Figure 1. Text after cleaning and morphological analysis.

3 Finding lemmas

After cleaning the file, it was necessary to find the most frequently used lemmas to compile a suitable lexicon. Finding the most frequently used lemmas was done using another shell-script. Since the regular morphological ending of infinitive form of Estonian verb used in dictionariy entries (-ma) wasn't particularly useful for context, the verb stems were used. There didn't seem to be much variation amongst the texts, most of them shared the most frequent words, which are displayed in Figure 2.

```
1412 olema 'to be'
1033 mina 'I'
748 see 'this'
734 et 'that'
634 ja 'and'
524 siis 'then'
497 ei 'no'
470 nagu 'as, like'
362 jah 'yes'
360 naer 'laughter'
326 sina 'you'
```

Figure 2. The most frequent lemmas from all the texts together with their Estonian translations.

While comparing the vocabulary of the material with that of the general (written) Estonian, one could say that the differences can be described as general differences between written and spoken language – personal pronouns *mina* 'I' and *sina* 'you' are more frequent as well as various spoken language particles, e.g. *nagu* 'like', *noh*, *okei* 'okay'.

4 Compiling sentiment lexicons

The lexicon was separated into two categories: positive and negative words. Both of these categories were compiled by using the most prototypical lemmas (*good*, *bad*, *interesting, hard* etc.) and some frequent lemmas from the texts (such as conversational cues: *mhm, yes*, *okay,* etc). Altogether the dictionary was quite small: 46 words, most of those positive. Although some (such as Vainik (2014: 346)) have argued that splitting words by valence isn't enough for most applications, the decision was made to use just two lexicons, since there hasn't been much detailed research into emotional categories and corresponding words. Of course, statistical methods are very popular too. Figure 3 gives some positive and negative words from the lexicon.

```
Positive sentiment words

jajaa 'yes-yes'
julge 'brave'
legendaarne 'legendary'
lihtne 'simple'
meeldiv 'pleasant'


Negative sentiment words

häbi 'shame'
hull 'crazy'
igav 'boring'
imelik 'strange'
keeruline 'complicated'
```

Figure 3. Some words from the lexicons together with their English translations.

As mentioned, Estonian is a morphologically rich language, so there is a need to use the lemmas rather than operate on all the different word forms. To make the task easier, the word-forms used in text were all made (using shell-script) into lemma variants of the same word. Compared to Figure 1, Figure 4 might be hard to understand for an actual language speaker, but it

keeps the lexicons concise and shouldn't change the meaning much, when already analysing only words, not phrases. At this stage, all the texts were joined together into one file.

```
oot kas siis keegi  teine
(oot kas siis kedagi teist)
wait if then anybody other
              -part  -part

siin ei ole    juures
(siin ei olegi  juures)
here no be-neg presence

näge või
(näha või)
see-inf  or

'wait, isn't anybody else
seen nearby'

vist    küll   jah
perhaps surely yes
'perhaps yes, sure'

väike naer
'some laughter'

ei aga mina ei tead   keegi
ei aga ma ei tea,     keegi
no but I  no know-neg
anybody

nagu ei
nagu ei
like  no
'no, but I don't know,
anybody like not...'
```

Figure 4. The text is lemmatized. Original text is in the parenthesis, followed by the English glosses. English translation is given in the end of every utterance.

5 Sentiment analysis

The sentiment analysis program was written in Python, using some of the code developed by the first autor for her bachelor thesis (Ojamaa 2014). The utterances were divided into four groups according by their sentiment: positive, negative, neutral and ambiguous.

Lexicons were read into lists and if matches were found, they were compared with the rest of the sentence to look for negation or other recognised sentiment words. Negation (formed by regular expression to account for three words that negate in Estonian: *ei, pole, mitte*) was allowed to influence words up to four words in the right- or left-hand context of the negation word. In the output of the program (the results file) the program displayed the sentiment evaluation for the sentences as well as the number of the sentence and the sentiment words found with their sentiment in context (since if negated, a positive word should have a negative polarity).

Out of 2902, the program annotated 576 sentences or 20% as positive, 38 or 1.3% as negative, and 3 as ambiguous. The rest or almost 79% of the utterances were either neutral or contained no sentiment.

6 Evaluation

200 analyzed sentences (utterances) were evaluated manually to see possible problems with the rules and lexicons. Of those, only 30 had got the wrong polarity tag, i.e. the overall correctness was 85%.

Of those 30, 20 erroneous decisions occurred because of the meta-comment "laughter" that was included in the positive lexicon, but in dialogues often signified awkwardness instead. The rest of the errors could feasibly be solved using regular expressions to find conversationally positive utterances (such as *mm* (a form of *mhm*) or *jaah* ('yees')). There was also a slight problem with negation, where the four word context might have been too large or punctuation should also have been taken into account. A few problems were caused by the small size of the lexicon assome sentiment words weren't recognized.

As mentioned, we can also compare the sentiment analysis results with the speaker's self-

evaluation of the interaction. Comparing the results with data published by Jokinen and Tenjes (2012), where participants were mostly positive in their descriptions of their participating in the video collection, it seems that sentimental analysis is consistent with the results that point to the conclusion that the participants felt happy discussing in front of cameras and they could self-reflect on it later.

7 Discussion and Future work

This paper started to explore the use of standard sentiment analysis tools and methods in analyzing transcribed conversational data. The results show that the methods can be applied with fairly good classification results, and even though most sentences are neutral, the speakers are mostly positive when showing sentiment.

Still there are many issues to be taken into account and to be studied further. The text represents a spoken natural language, and there can be errors in speaking and in transcription. Estonian morphological analysis should also be more accurate, so as to improve the use of statistical methods and additional conversational cues when analyzing the texts.

Also, as mentioned, dialogues are joint efforts so we need to distinguish the two speakers, and take into account the fact that the speakers influence each other and their utterances are dependent on the previous utterances. The transcribed text should thus take into account the interaction context, and differentiate between the different speakers.

When dealing with spoken dialogues, speech signal is an clear source of predicting. We can also use speech signal analysis as the corpora contain videos from which the signal properties can be extracted.

References

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10): 341-345.

Kristiina Jokinen and Silvi Tenjes. 2012. Investigating Engagement – Intercultural and Technological Aspects of the Collection, Analysis, and Use of Estonian Multiparty Conversational Video Data. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mar (Ed.). *Proceedings of the Eight International Conference on Language Recourses and Evaluation* (LREC'12):2764 – 2769.

Diane Litman and Kate Forbes. 2003. Recognizing emotions from student speech in tutoring dialogues. *Workshop on Automatic Speech Recognition and Understanding*, ASRU '03. IEEE.

Bing Liu 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool. Synthesis Lectures on Human Language Technologies.

Birgitta Ojamaa. 2014. Tartu Ülikooli üliõpilaste tagasiside hoiakute analüüs. Bachelor's thesis.

Patrizia Paggio, Jens Allwood, Elisabeth Ahlsén, Kristiina Jokinen and Costanza Navarretta. 2010. The NOMCO Multimodal Nordic Resource – Goals and Characteristics. *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC'10): 2968-2973

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity. *Procs of the Association for Computational Linguistics* (ACL): 271-278.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Procs of the Association for Computational Linguistics* (ACL):115–124.

Ene Vainik. 2014. Kelle emotsioonile anda teksti ette lugedes hääl? Vaatlus mitmest perspektiivist. *Eesti Rakenduslingvistika Ühingu aastaraamat.* Vol 10 (2014): 335-351.