Simple Opinion Mining Tool for Estonian Language

Fanny-Dhelia Pajuste & Maarja Lepamets
Institute of Computer Science, Faculty of Mathematics and Computer Science, University of Tartu
curriculum: Computer Science 2013/14
Project repository: https://github.com/fannydhelia/opinion-mining

Introduction

Opinion mining is generally used for extracting people's opinions on movies, books, dishes, etc based on texts of natural languages. Unlike English language Estonian uses several suffixes for nouns, adjectives and verbs which make the task more difficult. Therefore, the lemmas of the words must be used instead of the actual words for detecting the same word in different cases. We implemented a tool for opinion mining for Estonian language that uses a morphological analysator [1] and considers negation and other sentiment shifters.

Dictionaries

The tool uses **three different dictionaries.** The first two are for all nouns, verbs and adjectives that can be associated with **positive** or **negative** emotions, respectively. The third dictionary contains words that can alter the meaning of an adjective when situated close to it. Those words are called **sentiment shifters.** All words in the dictionaries are converted to their canonical forms using morphological analyser of an Estonian language called **t3mesta** [1]. This helps to find the lemma of a word and exclude all the suffixes of different cases. For example, the **lemma** for word "KENALE" ("to pretty") is "KENA" ("pretty"). It must be taken into consideration that all such dictionaries are often very subjective and depending on what one considers positive or negative. The dictionaries used are created by the authors of this project and include only a small amount of opinion related words. Additional words can be added to the dictionaries using the tool.

The first two dictionaries are divided into two. The first part contains the words that show mild positive/negative opinion, e.g "HEA" ("good"), "KENA" ("pretty"), "HALB" ("bad"), etc, the second part consists of words that express strong emotions, e.g "IMELINE" ("wonderful"), "ASENDAMATU" ("irreplacable"), "KOHUTAV" ("awful"), etc. Sentiment shifters are divided into three categories. There are **negation** words or phrases that switch the meaning of a word followed by them to the opposite, e.g "EI OLE" ("is not"), "MITTE" ("not"), etc. Other sentiment shifters can **strengthen** or **weaken** the emotion. The examples of the strengthening words are "VÄGA" ("very") and "ERITI" ("especially"), the examples for weakening word are "ÜPRIS" ("quite") and "VEIDI" ("slightly").

Algorithm

The tool takes a **user-provided text** as an input. At first it divides the text into subsentences by splitting it on full stops, commas, colons, semi-colons, exclamation and question marks. The tool then **lemmatizes** the text and searches it for all words that are given in any of the three dictionaries and marks their locations. For every found positive or negative word the programm starts looking at the positions close to it within one subsentence to detect sentiment shifters. The **neighbouring locations** are considered first and then moved further from there. The shifters are searched as long as we reach the locations that are already closer to another positive or negative word.

References

1. Kaalep, H.-J., Vaino, T. (OÜ Filosoft, 1998). t3mesta – Estonian language morphological analysator.

Sentiment Scores

Every positive and negative word has a **sentiment score.** For mildly opinionated words the score is +/-2, for strong words the score is +/-4. The sentiment shifters alter the score by 1 in either direction and the negation switches the score with a value having an opposite sign. In current implementation the score is switched to -/+2 in case of negation.

To illustrate the sentiment scores let us consider the **examples below.**

Example 1: **strongly negative opinion**
TÄNANE ILM ON KOHUTAV. (Today's weather is awful.)
KOHUTAV (awful) -4
Final opinion: -4

Example 2: **strongly positive opinion**
TA ON VÄGA TARK JA KENA. (He is very smart and handsome.)
TARK (smart) & VÄGA (very) +3
KENA (handsome) +2
Final opinion: +5

Example 3: **multiple sentiment shifters**
FILM EI OLNUD VÄGA HUVITAV JA NÄITLEJAD OLID KOLEDAD, AGA VÄHEMALT TOIT OLI HEA. (Movie was not very interesting and actors were ugly but at least food tasted good.)
HUVITAV (interesting) & VÄGA (very) & EI OLNUD (was not) -2
KOLEDAD (ugly) -2
HEA (good) +2
Final opinion: -2

Discussion

The aim of the project was to implement a programme which could detect the opinion of a writer based on the words they use in their text. We succeeded in writing a simple **dictionary-based tool** which also uses negation and other sentiment shifters. Still, to advance the tool several improvements can be made:

- detect opinions on different aspects of a described thing
- consider that sentences with "AGA" ("but") and "SIISKI" ("nevertheless") have different opinions on either side of the word
- consider grammatical correctness and slang

Study IT in .ee sponsored by Skype™
ARVUTITEADUSE INSTITUUT
TARTU ÜLIKOOL 1632 UNIVERSITAS TARTUENSIS