Rule-Based Sentiment Analysis in Narrow Domain

Detecting Sentiment in Daily Horoscopes Using *Sentiscope*

Željko Agić and Danijela Merkler

University of Zagreb
Faculty of Humanities and Social Sciences

SAAIP 2012, Mumbai, India, 2012-12-15

Overview

- motivation
- system design and implementation
  1. collecting horoscope texts from the web on a daily basis
  2. rule-based module for polarity phrase detection designed in NooJ linguistic development environment
  3. web-based wrapper application for counting polarity phrases and assigning overall sentiment scores
  4. simple visualization module
- evaluation
- rule-based component demo and visualization demo

Document collection

- developed a simple focused crawler
- collected horoscopes from largest websites (in Croatian)
  - selected by Google search index
  - eight different newspaper portals and specialized portals
- collected from 2012-02-11 to 2012-05-10
- 7,716 articles, 484,179 tokens

Inter-annotator agreement

- development set of 333 articles manually annotated by two human annotators for overall sentiment and polarity phrases
- lineary weighted kappa: 0.593 $\rightarrow$ moderate agreement
- excluding neutral sentiment, kappa: 0.989 $\rightarrow$ very good agreement

	+	−	x	Σ
+	94	0	26	120
−	1	82	31	114
x	18	4	77	99
Σ	113	86	134	333

Overall article sentiment and polarity phrases

- positive phrases imply positive overall sentiment and vice versa
- also applies when both types of phrases are present
- even distribution of phrases for neutral sentiment articles
- justifies theoretical baseline that overall sentiment is assigned from the polarity group with the highest count

	<p>	<n>	both	<p> in both	<n> in both
+	410	27	23	85	27
−	19	321	15	19	53
x	142	145	67	117	115

Phrase detection

- designed in two stages — from scratch and by observing the development set
- grouped in two NooJ local grammars
  - positive and negative sentiment detection
- focus on three POS
  - adjectives, nouns and verbs
  - adverbs are homographic with adjectives in singular nominative case in neuter gender
- 170 negative and 139 positive words and phrases
- aggregate of positive and negative words which occur with a negation, which results in expressing the opposite sentiment
  - 33 negated positive and 17 negated negative words and phrases
- a total of 203 words and phrases for negative and 156 words and phrases for positive sentiment detection

Demo

Polarity phrase detection in NooJ

Evaluation

- conducted on a manually annotated held-out test set
  - initial run also on portion of development set
  - approximately 11,500 tokens in 168 articles each
- polarity phrase detection accuracy of the rule-based component

sample	**precision**	**recall**	**F_1-score**
initial	0.371	0.283	0.321
development	0.435	0.469	0.451
test	0.413	0.393	0.402

Evaluation

- system accuracy on overall sentiment detection and confusion matrix for overall sentiment assignment
- system performance is high in discriminating between positive and negative overall sentiment
- accuracy steeply decreases upon inclusion of neutral sentiment
- positive words and phrases are more accurately detected

	$+^*$	$-^*$	x^*	**precision**	**recall**	**F_1-score**
$+$	40	3	17	0.677	0.666	0.671
$-$	2	25	17	0.555	0.568	0.561
x	17	17	30	0.468	0.468	0.468

Demo

Prototype web interface for data visualization

Conclusions and future work

- detecting sentiment in narrow domain such as daily horoscope texts is not easy to achieve
  - complex phrases and syntax
  - specific style, even for each individual author
- obtained results as baseline for further work
  - overall F_1-score: 0.566
  - F_1-score for phrase detection: 0.402
  - moderate inter-annotator agreement
- obtained data can be used for different types of linguistic analysis
- re-implementation of the link between polarity phrases and overall sentiment
  - elimination of neutral sentiment category
- model adjustment and application for sentiment annotation and visualization in other domains
  - precision and recall shown to be much higher (0.9, 0.6) using the same framework for financial texts

Thank you for your attention! ☺