Multilingual Sentiment Analysis

Comparing techniques in sentiment analysis on different languages

LING 575
Claire Jaja, Andrea Kahn

Problem Definition

Sentiment analysis techniques are typically developed on English.

Current approaches to other languages often involve automatic translation or use of “language agnostic” techniques like machine learning.

This raises two research questions:

1. Are machine learning techniques really language agnostic?

2. How do the results obtained when using resources translated/pivoted from English compare to those with resources developed in the test language?

Datasets

IMDb movie reviews (Pang and Lee, 2004)

- English
- 1000 positive, 1000 negative reviews, pre-processed

CorpusCine movie reviews (Cruz Mata, 2011)

- Spanish
- 3878 reviews with 1 - 5 star ratings
- processed by us, discarding 3 star reviews, then choosing 1000 positive and 1000 negative

quotations from newspaper articles, annotated for polarity

- English (Balahur-Dobrescu and Ralf, 2009)
    - 1590 total, where annotators agree: 863 obj, 193 pos, 234 neg
- German (Balahur-Dobrescu, 2011)
    - 2387 total, where annotators agree: 591 obj, 514 pos, 379 neg

Methodology

classifiers: MaxEnt, Naive Bayes

features: unigram (with and without frequency cut-off), bigram, trigram, unigrams from General Inquirer sentiment lexicon

use 10-fold cross validation

Results: MaxEnt

features	IMDb			CorpusCine		
	average	min	max	average	min	max
unigram	86.00%	81.00%	91.50%	83.40%	81.50%	86.00%
unigram > 4	68.20%	62.00%	71.50%	67.45%	60.50%	73.00%
bigram	84.65%	80.50%	88.00%	83.10%	78.00%	87.00%
trigram	50.05%	49.50%	51.00%	81.00%	76.00%	87.00%
unigram + bigram	85.35%	82.50%	89.00%	82.70%	79.50%	86.50%
sentiment lexicon	78.70%	74.00%	83.00%	?	?	?

Results: Naive Bayes

features	IMDb			CorpusCine		
	average	min	max	average	min	max
unigram	81.65%	75.00%	87.00%	82.70%	79.00%	86.50%
unigram > 4	69.20%	62.50%	74.50%	64.75%	59.50%	69.50%
bigram	81.15%	73.50%	85.50%	81.80%	78.50%	85.00%
trigram	80.95%	73.00%	86.00%	81.55%	78.50%	85.00%
unigram + bigram	81.45%	74.50%	85.50%	81.85%	78.00%	85.00%
sentiment lexicon	78.50%	75.00%	82.50%	?	?	?

Results: Discussion

- using a unigram frequency cut off of 4 drastically drops results
- MaxEnt is better than Naive Bayes on IMDb using unigram and/or bigram features
- MaxEnt is weirdly bad using trigram features on IMDb
- CorpusCine results are worse than IMDb results using MaxEnt and unigram and/or bigram features
- IMDb and CorpusCine results are comparable using Naive Bayes - NB is more language agnostic? (when it comes to two similar languages like English and Spanish…)

Future Work

- translate sentiment lexicon into Spanish, use for CorpusCine
- find Spanish sentiment lexicon, use for CorpusCine
- translate CorpusCine test set(s) into English, use IMDb trained classifiers
- address negation in the text
- lemmatize text
- try subjectivity classification for English and German newspaper quotes

Thanks for listening!