Comparable Corpora for Cross-Linguistic Sentiment Analysis

Maite Taboada

Department of Linguistics
Simon Fraser University
Vancouver, Canada
mtaboada@sfu.ca

AACL 2009

American Association for Corpus Linguistics

Edmonton, Alberta
8-11 October, 2009

Sentiment analysis

- Evaluation, subjectivity, opinion, sentiment, stance
- Sentiment
  - Opinion expressed in discourse
  - About anything
    - a book, a movie, a new car, a politician, a group of people
  - Anywhere
    - a personal conversation, an e-mail message, an editorial, a newspaper article, a conference presentation, a business report
  - Multiple applications
    - Market intelligence, stock price analysis, political trends, public opinion tracking, …

Extracting sentiment from text

Ok let me get this out of the way: I did NOT like this movie. Borderline hated it. Now if you are only after positive reviews, then you might skip this, but if you are interested in why I thought it sucked, read on.

To begin with, I only mildly like Will Ferrell. I don't think he’s funny at all on SNL, but I was willing to give him a chance. I had seen previews for this a few weeks before it came out and it looked interesting, so I figured I would give it a shot.

I'm very thankful that a local theater gives military discounts, I only paid $5.50 to see it, and after watching the movie, I feel like I COMPLETELY WASTED my money.

It's a 100% kids movie. There is ZERO bad language in it. The special effects and props are EXTREMELY low budget and so fake its not even funny. You can clearly tell what parts were filmed on a sound stage, as the backgrounds are blatently fake, no effort was put into them.

www.epinions.com

My goal

- Classification of texts based on subjective content (=sentiment)
  - Positive
  - Negative
- Input a text, and produce a numeric value that expresses its subjective content → the text’s sentiment
  - -3: quite negative
  - -5: very negative
  - 2: somewhat positive
  - …

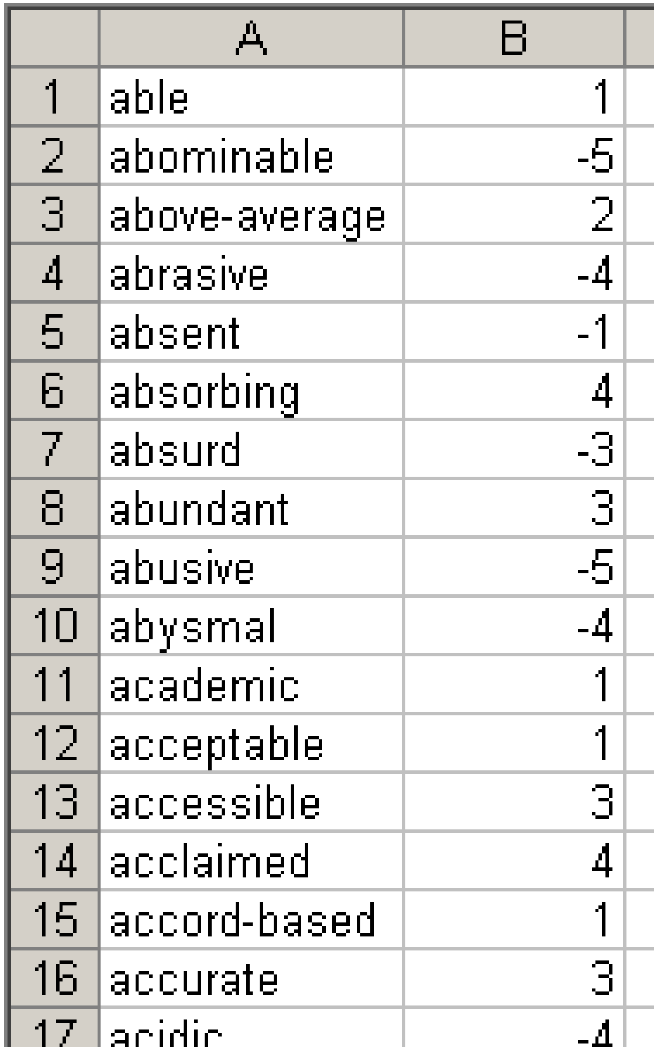
- Other types of information in the future
  - Type of opinion expressed (Appraisal Theory)
  - Holder and target of the opinion
  - Summary/key words that contain the opinion
  - Structure of the argument

Outline

- Sentiment analysis
- Corpora in sentiment analysis
  - Semantic, or lexicon-based systems
  - Machine Learning, or corpus-based systems
- Collecting corpora for development and testing
  - The English SO-CAL
- Working on a new language: Options
- Comparable vs. parallel corpora
- What is lost in translation
- Conclusions

Two approaches in sentiment analysis

- Semantic, or lexicon-based
  - Dictionaries of opinion-bearing words
  - Annotated with polarity and strength
  - Extract those words from texts, average their values across the text
- Machine learning, or "corpus-based"
  - Collect a corpus with some labelling
    - Positive and negative texts
  - Build a classifier that learns to distinguish one type of text from another
  - Different types of features used (n-grams, POS, text features)



	A	B
1	able	1
2	abominable	-5
3	above-average	2
4	abrasive	-4
5	absent	-1
6	absorbing	4
7	absurd	-3
8	abundant	3
9	abusive	-5
10	abysmal	-4
11	academic	1
12	acceptable	1
13	accessible	3
14	acclaimed	4
15	accord-based	1
16	accurate	3
17	acidic	-4

SO-CAL: The Semantic Orientation Calculator

- Dictionary-based
  - 2,257 adjectives, 1,142 nouns, 903 verbs, and 745 adverbs
  - 177 intensifying expressions
- Valence shifters
  - Intensifiers and negation (shift rather than switch)
- Irrealis
  - Modals, verbs of opinion (*expect),* punctuation

Brooke 2009 (SFU thesis), Taboada, Brooke and Stede 2009 (SIGDial), Brooke, Tofiloski and Taboada 2009 (RANLP)

Why we follow a semantic approach

- Linguistically motivated, transparent features
- Good performance on unseen data
- Good cross-domain performance

development

Corpus (online reviews)	% correct
400 Epinions 1	80.25%
400 Epinions 2	79.75%
2000 Movie	77.70%
2400 Camera	80.30%

Moving to a new language: Spanish

- Options
  - Create resources from scratch
  - Translate existing resources (e.g., dictionaries)
  - Translate target texts into English
- Recent work in sentiment analysis suggests that translating texts into English is the way to go
  - Regardless of approach (semantic or machine learning) (Bautin et al. 2008, Wan 2008)
  - We may lose some information, since translation is automatic
  - But sentiment translates well
    - Or does it?

Comparable vs. parallel corpora

- Old debate, about how to do corpus linguistics in general (Johansson 2007)
- Parallel corpora
  - Exactly the same meanings across languages
- Comparable corpora
  - More representative of the actual linguistic expression in the language
  - But the range and type of linguistic phenomena may be different in each language
- For sentiment analysis
  - Parallel corpora do not provide the full range of sentiment expressed in each language

The options

1. Create resources from scratch	• 28 hours
• Collect a development corpus	• 5-6 hours
• Build dictionaries from that corpus and other sources	• 12 hours
• Install Spanish part-of-speech tagger	• 4 hours
• Lemmatizer (not necessary in English)	• 6 hours
2. Translate existing English dictionaries	• 4-6 hours
• Use an on-line dictionary (Spanishdict.com)	
• Take the first sense	• 4 hours
• Optional manual editing	• 2 hours
3. Translate target texts into English	• 4 hours
• Using Google translate	

Results

- Only “unseen” data: corpora that we had not used in development

Corpus language	System	Method	% correct
English	English	SO-CAL	79.75%
Spanish	Spanish	Hand-built resources	72.00%
Spanish	Spanish	Spanishdict.com	67.25%
Spanish	English	Translated into English	66.50%

What is lost in translation

- Translation of any kind seems to come at a cost
- Translating dictionaries
  - Translated and hand-built dictionaries contain different words (20-40% of the words are different)
  - Translated dictionaries tend to be much more formal
- Translating the corpus
  - Of course, Google is not a very good translator
  - Would a good translation help overcome these problems?

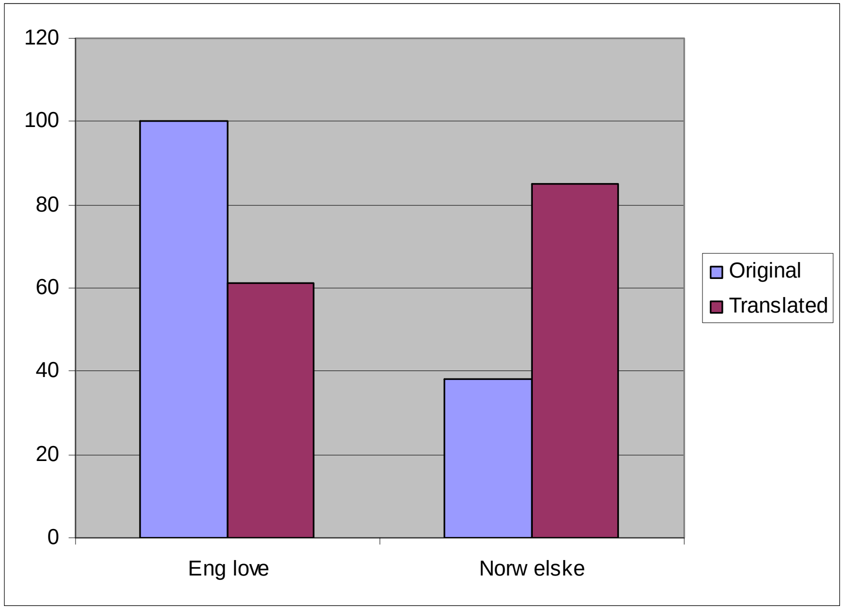
Using parallel corpora

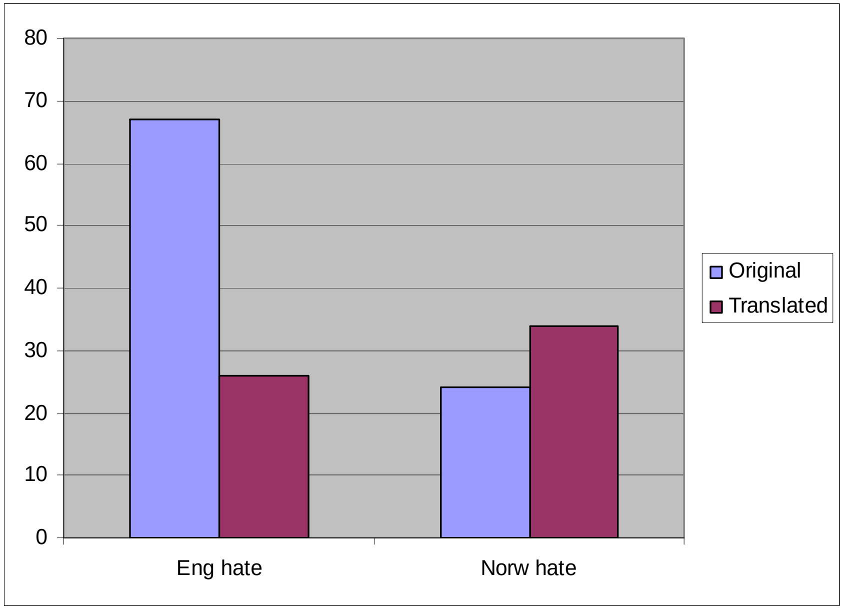
- Translated columns by Paul Krugman
  - 10 columns in *The New York Times*
  - 10 columns in *El País* (translated into Spanish)
- Europarl
  - 200 texts from the English-Spanish parallel version
    - 100 English
    - 100 Spanish
- In both cases, there were differences in the score for each text
  - Not huge (under 1 standard deviation)
  - But often meaningful enough (i.e, the polarity of the text was positive in one language, and negative in the other)

Why don’t translations work? (1)

1. Differences in the languages themselves

- Contrastive studies have shown differences in the more subtle aspects of the language, such as modal particles, discourse markers or negation (Degand 2003, Fischer 2006, Johansson 2007, Stefanowitsch 2004, …)
- Johansson (2007) on *love* and *hate* in English and Norwegian





SFU SIMON FRASER UNIVERSITY
THINKING OF THE WORLD

Why don’t translations work? (2)

2. Problems with translations

- Word sense disambiguation not working well in one of the languages
  - Eng: *free fall* (“free” interpreted as positive)
  - Spa: *caída en picado*

  - Eng: *late 2008* (“late” interpreted as negative)
  - Spa: *finales de 2008*

Why don’t translations work? (3)

3. Differences in the genre

- In Spanish, the titles are often enough to get the whole meaning of the review, but are difficult to translate (or don’t play a big role in English)
  - About a children’s movie:
    - *Hasta los niños se duermen en el cine* (‘Even the children fall asleep in the movie theatre’)
  - About a cordless phone:
    - *Mejor que fabriquen peines* (‘They should make combs instead’)

Conclusions

- For sentiment analysis
  - A semantic approach is the more robust, domain-independent approach
- For sentiment analysis in a new language
  - Creating a new corpus and new dictionaries works better than
    - Translating from target language into English
    - Translating existing English dictionaries into the target language
  - Because languages differ in which categories and words they use to express sentiment

References

Bautin, Mikhail, Lohit Vijayarenu and Steven Skiena. (2008). International sentiment analysis for news and blogs*. Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*. San Jose, CA.

Brooke, Julian. (2009). *A Semantic Approach to Automatic Text Sentiment Analysis.* Unpublished M.A. thesis, Simon Fraser University.

Brooke, Julian, Milan Tofiloski and Maite Taboada. (2009). Cross-linguistic sentiment analysis: From English to Spanish*. Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing*. Borovets, Bulgaria.

Degand, Liesbeth and Henk Pander Maat. (2003). A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. In A. Verhagen and J. van de Weijer (Eds.), *Usage based approaches to Dutch* (pp. 175-199). Utretcht: LOT.

Fischer, Kerstin (Ed.). (2006). *Approaches to Discourse Particles*. Amsterdam: Elsevier.

Johansson, Stig. (2007). *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies*. Amsterdam and Philadelphia: John Benjamins.

Stefanowitsch, Anatol. (2004). HAPPINESS in English and German: A metaphorical-pattern analysis. In M. Achard and S. Kemmer (Eds.), *Language, Culture, and Mind* (pp. 137-149). Stanford: CSLI.

Taboada, Maite, Caroline Anthony, Julian Brooke, Jack Grieve and Kimberly Voll. (2008). SO-CAL: Semantic Orientation CALculator. Vancouver: Simon Fraser University.

Taboada, Maite, Julian Brooke and Manfred Stede. (2009). Genre-based paragraph classification for sentiment analysis*. Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue* (pp. 62-70). London, UK.

Wan, Xiaojun. (2008). Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis*. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 553-561). Honolulu.

Comparable Corpora for Cross-Linguistic Sentiment Analysis

Maite Taboada

Simon Fraser University

mtaboada@sfu.ca

http://www.sfu.ca/~mtaboada/

http://www.sfu.ca/~mtaboada/research/nserc-project.html