Cross-Linguistic Sentiment Analysis: From English to Spanish

Julian Brooke
Department of Linguistics

Milan Tofiloski
School of Computing Science

Maite Taboada
Department of Linguistics

SFU SIMON FRASER UNIVERSITY
THINKING OF THE WORLD

Introduction

Sentiment Analysis

- Subjectivity, Polarity, and Strength
- Sentence or Text
- Lexical (Semantic) or Machine Learning (Corpus-based)
- Semantic Orientation (SO)

SA in a New Language

- Build new resources?
- Adapt existing ones?
- Machine Learning?
- Machine Translation?

The English SO Calculator

Dictionaries

- 2,257 adjectives, 1,142 nouns, 903 verbs, and 745 adverbs,
- 177 intensifiers, includes multi-word expressions
- Hand-ranked, committee-reviewed SO of between 5 and -5
- Taken from multi-domain Epinions Corpus, Polarity Dataset, General Inquirer

Contextual Valence Shifters

- Negation: shift instead of switch (e.g. *not terrible* -5 → -1 instead of -5 → 5)
- Intensifiers as Percentage Modifiers (e.g. *very* = +25%, *extraordinary* = +50%
- Irrealis blockers (e.g. modal *would* in *it would be good*)
- Extra weight on negative expressions to counteract positive bias

Evaluation of Features

- Movie = 2000 movie reviews
- Camera = 3000 camera reviews
- All features useful
- Consistent across domains

Percent Correct by Corpus

Features	Epinions	Movie	Camera	Total
All	80.3	76.4	80.3	78.7*
No Neg	75.8*	74.6	76.1*	75.4*
No Int	79.0*	74.7	77.5*	76.5*
No Irreal	78.8*	74.8	79.6	77.6*
No Neg W	71.8*	75.6	71.5*	73.2*

* = $p < 0.05$ significance

Adapting SO-CAL to Spanish

Dictionaries

- Translated English SO Dictionaries
  - Google
  - Spanishdict.com
  - Spanishdict.com hand-fixed (2 hours work)
- Hand-ranked words from Ciao Corpus (12 hours work)
- Combined translated and hand-ranked
  - 2,049 adjectives, 1,324 nouns, 739 verbs, and 548 adverbs
  - Comparable in size to English dictionaries

Spanish Differences

- Complex inflections
- Flexible word order (e.g. adjective position)
- Verb mood irrealis (e.g. subjunctive)

Otherwise, same software used for both English and Spanish SO calculation

Alternative Approaches

- Machine Translation
  - Each Corpus Translated Using Google Translator
  - Spanish Corpus → English, use English SO Calculator
  - English Corpus → Spanish, use Spanish SO Calculator
- Machine Learning
  - Support Vector Machine Classifier
  - Unigram Feature Model (Pang et al. 2002)
  - 10-fold cross-validation on each corpus

Related Work

- Wan et al. 2008
  - Use machine translation to improve Chinese sentiment classification
- Original Chinese lexical resources were of little benefit
- Mihalcea et al. 2007
  - Translate lexical resources for subjectivity detection
  - Limited benefits, instead projected subjectivity annotations
- Bautin et al. 2008
  - Tracking sentiment in different languages over time
  - Used an English system and various machine translators
- Yao et al. 2006, using a bilingual dictionary to build a Chinese sentiment dictionary
- Banea et al. 2008, subjectivity detection in Spanish

Corpora

- Two English Corpora
  - Epinions, words used to populate English SO dictionaries, www.epinions.com
  - Epinions 2, unseen corpus
- Two Spanish Corpora
  - Ciao Corpus, words used to populate Ciao dictionary, www.ciao.es
  - Dooyoo Corpus, unseen corpus, www.dooyoo.es
- 400 texts each
- Balanced for polarity
- 50 texts in each of 8 consumer product areas (movies, books, music, phones, hotels, computers, cars, and cookware/appliances)

Evaluation

- Polarity (recommendation) provided by original users
- Accuracy: percentage of texts whose polarity is correctly identified by the classifier
- Test each classifier/dictionary and corpus combination (use translations when needed)
- Compare all original corpora with all translated corpora to evaluate the effect of translation

Results

Classification Method		Corpus				
		English		Spanish		Overall
SO Calculator	Dictionary	Epinions	Epinions 2	Ciao	Dooyoo	
English	English SO-CAL	**80.25**	**79.75**	72.50	**73.50**	**76.50**
Spanish	Google-translated	66.00	68.50	66.75	66.50	66.50
Spanish	Spanishdict.com	68.75	68.00	67.25	67.25	67.94
Spanish	Fixed Spanishdict.com	69.25	69.75	68.25	68.00	68.81
Spanish	Ciao (manual)	66.00	67.50	**74.50**	72.00	70.00
Spanish	Ciao + Fixed Combined, Ciao Preferred	68.75	**72.50**	74.25	72.25	**71.93**
Spanish	Ciao + Fixed Combined, Fixed Preferred	69.50	68.75	73.50	70.75	70.87
Support Vector Machine, English versions		76.50	71.50	72.00	64.75	71.25
Support Vector Machine, Spanish versions		71.50	68.75	72.25	69.75	70.56

Method	Texts	Accuracy
SO Calculation	Original	76.62
	Translated	71.81
SVM	Original	72.56
	Translated	69.25

Discussion

- Translated dictionaries do poorly, except with translated texts ($p < 0.05$)
- SVM does poorly, mostly due to small training set
- Translating texts and Spanish SO calculator are comparable
- Spanish SO-CAL is still rough, English SO-CAL has been thoroughly tested
- However, translated texts overall do significantly ($p < 0.01$) worse than originals using SO-CAL, similar for SVM
- Only small preference for high coverage dictionary/corpus combinations, not significant
- Overall, building new dictionary resources on top of existing software seems preferable to machine translation, especially in the long term

References

C. Banea, R. Mihalcea, J. Wiebe and S. Hassan. Multilingual subjectivity analysis using machine translation. Proc. of EMNLP. Honolulu, 2008.

M. Bautin, L. Vijayarenu and S. Skiena. International sentiment analysis for news and blogs. Proc. of 3rd AAAI International Conference on Weblogs and Social Media. San Jose, CA, 2008.

A. Kennedy and D. Inkpen. Sentiment classification of movie and product reviews using contextual valence shifters. Computational Intelligence 22(2): 110-125, 2006.

R. Mihalcea, C. Banea and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. Proc. of ACL. Prague, Czech Republic, 2007.

B. Pang, L. Lee and S. Vaithyanathan. Thumbs up? Sentiment classification using Machine Learning techniques. Proc. of EMNLP, 2002.

L. Polanyi and A. Zaenen. Contextual valence shifters. In Computing Attitude and Affect in Text: Theory and Applications, J.G. Shanahan, Y. Qu, and J. Wiebe, Eds. Springer: Dordrecht, pp. 1-10, 2006.

X. Wan. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. Proc of EMNLP. Honolulu, 2008.

J. Yao, G. Wu, J. Liu and Y. Zheng. Using bilingual lexicon to judge sentiment orientation of Chinese words. Proc. of 6th International Conference on Computer and Information Technology (CIT'06). Seoul, Korea, 2006.

Acknowledgements

This work was supported by an NSERC Discovery Grant (261104-2008) to Maite Taboada.

Contact: **Maite Taboada** ▸ **mtaboada@sfu.ca** ▸ **http://www.sfu.ca/~mtaboada**