Julian Brooke Tong Wang Graeme Hirst

jbrooke, tong, gh@cs.toronto.edu Department of Computer Science, University of Toronto

Automatic Acquisition of Lexical Formality

1. Introduction

- **Goal**: Quantify the *formality* of lexical items, assigning a formality score (FS) in the range −1 to 1 to each word (Brooke et al. 2010)
- **Theoretical basis**: Formality as a cline (Leckie-Tarry 1991; Biber 1995)
- **Approach**: Primarily corpus-based, inspired by similar research in lexical sentiment (Turney and Littman 2003)

Motivations

- Near-synonym word choice (*get* vs. *acquire* vs. *snag*)
- Languages where word length is not a usable metric (e.g. Chinese)

2. Data and Resources

Word Lists

- Seed sets
  - 138 informal, slang (*wuss*) and interjections (*yikes*)
  - 105 formal, discourse cues (*hence*) and adverbs (*adroitly*)
- Near-synonym pairs from *Choose the Right Word* (Hayakawa 1994)
  - 399 pairs of near-synonyms, e.g. *determine/ascertain*
- Chinese seeds: 49 formal, 43 informal (based on English list)

Corpora

- Brown Corpus (development corpus, both mixed and formal)
- Switchboard Corpus (spoken, informal)
- British National Corpus (mixed), 90% written (formal), 4.3% spontaneous spoken (informal)
- UofT Blog Corpus, 216 million tokens, from 900,000 blogs (mixed)
- ICWSM Blog Corpus (Burton et al. 2009)
  - English: 1.3 billion word tokens, from 7.5 million blogs (mixed)
  - Chinese: 25.4 million character tokens, from 86,000 blogs (mixed)

3. Methods

Simple

- Word length
- Latinate affixes, e.g. *-ation*
- Word count in corpora
  - *Formality is rare* and *informality is rare* assumptions
  - Ratio between counts in formality-divergent corpora

Latent Semantic Analysis (Landauer and Dumais 1997)

- Collapse word–document matrix to *k* dimensions
- Calculate cosine similarity to seed words
- Filtering necessary for large corpora
- Normalizing for corpus and seed set bias
- Chinese tokenization using character bigrams

Hybrid

- Combine word-count methods (back-off to *Rare* assumptions)
- Voting (decide only if *n* lexicons agree)
- Classification with ML algorithms (SVM, Naïve Bayes)
- (Weighted) average across lexicons
- Iterative methods using LSA cosine similarity

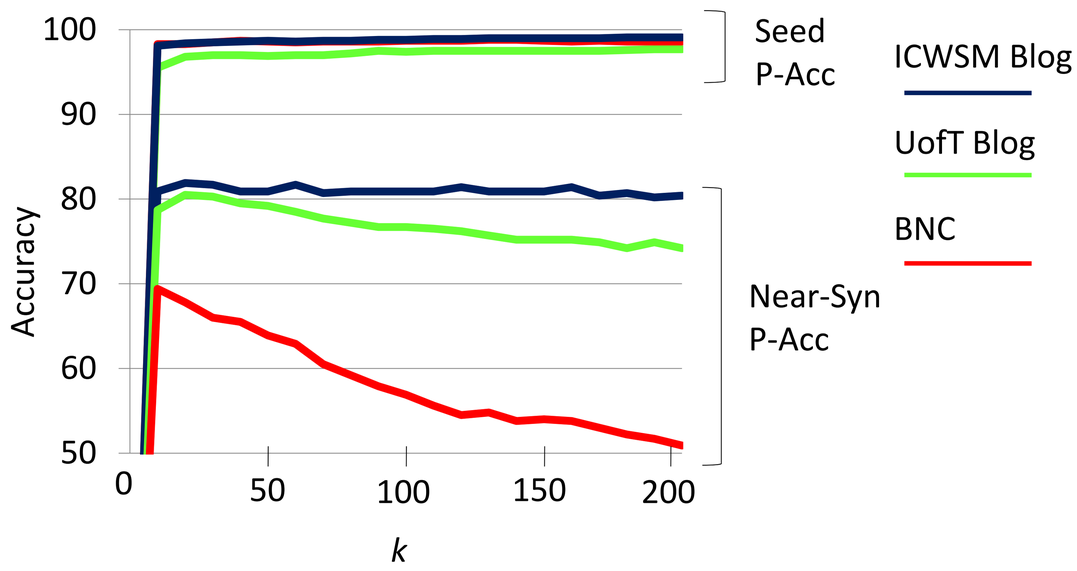
4. Evaluation

- Seed sets (*leave-one-out* cross-validation)
  - Coverage (% of words included in lexicon)
  - Class-based accuracy (FS > 0 implies formal, FS < 0 implies informal)
  - Pairwise accuracy (all possible formal/informal pairings)
- Near-synonym pairs
  - Same, but no class-based accuracy (only relative judgements)

5. Results

Method	Seeds			Near-Syns	
	Cov.	C-Acc.	P-Acc.	Cov.	P-Acc.
(1) Word length	100	86.4	91.8	100	63.7
(2) Latinate affixes	100	74.5	46.3	100	32.6
(3) Word count ratio, Brown and Switchboard	38.0	81.5	85.7	36.0	78.2
(4) Word count ratio, BNC Written vs. Spoken	60.9	89.2	97.3	38.8	74.3
(5) LSA (k=3), Brown	51.0	87.1	94.2	59.6	73.9
(6) LSA (k=10), BNC	94.7	83.0	98.3	96.5	69.4
(7) LSA (k=20), UofT Blog	100	91.4	96.8	99.0	80.5
(8) LSA (k=20), UofT Blog, filtered	99.0	92.1	97.0	97.7	80.5
(9) LSA (k=20), ICWSM, filtered	100	93.0	98.4	99.7	81.9

- Our corpus methods offer a marked improvement over word length
- LSA with large blog corpora is by far the best individual method
- Filters reduce blog word–doc matrices to 1/16 size, no loss of accuracy
- Lexicon induced from English ICWSM corpus includes 750,000 entries



- Low *k* values preferred for near-synonyms: formality-relevant dimensions are fundamental aspects of text variation (Biber 1995)
- Seed pairs are quite semantically distinct; thus increasing *k* helps
- Consistent across corpora, though the slope of change varies

Hybrid Method	Seed			Near-Syns	
	Cov.	C-Acc.	P-Acc.	Cov.	P-Acc.
(10) BNC ratio with backoff (4)	97.1	78.8	75.7	97.0	78.8
(11) Combined ratio with backoff (3 + 4)	97.1	79.2	79.9	97.5	79.9
(12) BNC weighted average (10,6), ratio 2:1	97.1	83.5	90.0	97.0	83.2
(13) Blog weighted average (9,7), ratio 4:1	100	93.8	98.5	99.7	83.4
(14) Voting, 3 agree (1, 6, 7, 9, 11)	92.6	99.1	99.9	87.0	91.6
(15) Voting, 2 agree (1, 11, 13)	86.8	99.1	100	81.5	96.9
(16) Voting, 2 agree (1, 12, 13)	87.7	98.6	100	82.7	97.3
(17) SVM classifier (1, 2, 6, 7, 9, 11)	100	97.9	99.9	100	84.2
(18) Naive Bayes classifier (1, 2, 6, 7, 9, 11)	100	97.5	99.8	100	83.9
(19) SVM (Seed, class) weighted (1, 2, 6, 7, 9, 11)	100	98.4	99.8	100	80.5
(20) SVM (CTRW) weighted (1, 6, 7, 9, 11)	100	93.0	99.0	100	86.0
(21) Average (1, 6, 7, 9, 11)	100	95.9	99.5	100	84.5

- Different corpora, different methods are complementary
- Voting allows for extremely high accuracy at the cost of coverage
- ML methods effective, but averaging provides best all-around lexicon
- Chinese results: C-Acc 79.3%, P-Acc 82.2%
  - Normalization is important for imbalanced Chinese corpus
- Iterative methods ineffective, LSA does not improve with more seeds

References and Acknowledgements

Biber, Douglas. 1995. *Dimensions of Register Variation: A cross-linguistic comparison*. Cambridge University Press.

Brooke, Julian, Tong Wang, and Graeme Hirst. 2010. Inducing lexicons of formality from corpora. *In Proceedings of the LREC 2010 Workshop on Methods for the Automatic Acquisition of Language Resources and their Evaluation Methods*, pages 17-22. Malta.

Burton, Kevin, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media* (ICWSM 2009), San Jose, CA.

Hayakawa, S.I., editor. 1994. *Choose the Right Word*. HarperCollins Publishers, second edition. Revised by Eugene Ehrlich.

Landauer, Thomas K. and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Leckie-Tarry, Helen. 1991. *Language Context: a functional linguistic theory of register.* Pinter.

Turney, Peter and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. A*CM Transactions on Information Systems*, 21:315–346.

This work was supported the Natural Sciences and Engineering Research Council of Canada. Thanks to Paul Cook for the suggestion of the ICWSM data, and the use of his corpus API.