Julian Brooke Tong Wang Graeme Hirst
jbrooke, tong, gh@cs.toronto.edu Department of Computer Science, University of Toronto

Inducing Lexicons of Formality from Corpora

1. Introduction

- **Goal**: Quantify the *formality* of individual lexical items, assigning a formality score (FS) in the range −1 to 1 to each word
- **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
  - 399 pairs of near-synonyms, e.g. *determine*/*ascertain*
  - Compared for formality in *Choose the Right Word* (Hayakawa 1994)

Corpora

- Brown Corpus (development corpus, both mixed and formal)
- Switchboard Corpus (SB) (spoken, informal)
- British National Corpus (BNC), 90% written (formal), 4.3% spontaneous spoken (informal)
- UofT Blog Corpus (5 days of blogs, see www.blogscope.net)
  - 216 million tokens, from 900,000 blogs (mixed)
- ICWSM Blog Corpus (Burton et al. 2009)
  - 1.3 billion tokens, from 7.5 million blogs (mixed)

3. Methods

Simple

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

Co-occurrence

- Pointwise Mutual Information (Church and Hanks 1990)
- Latent Semantic Analysis (Landauer and Dumais 1997)
  - Collapse word–document matrix to *k* dimensions
  - Calculate cosine similarity to seed words
  - Other options: weights (*td·idf*), lemmatization, linear regression (LR)
  - Filtering necessary for large corpora

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

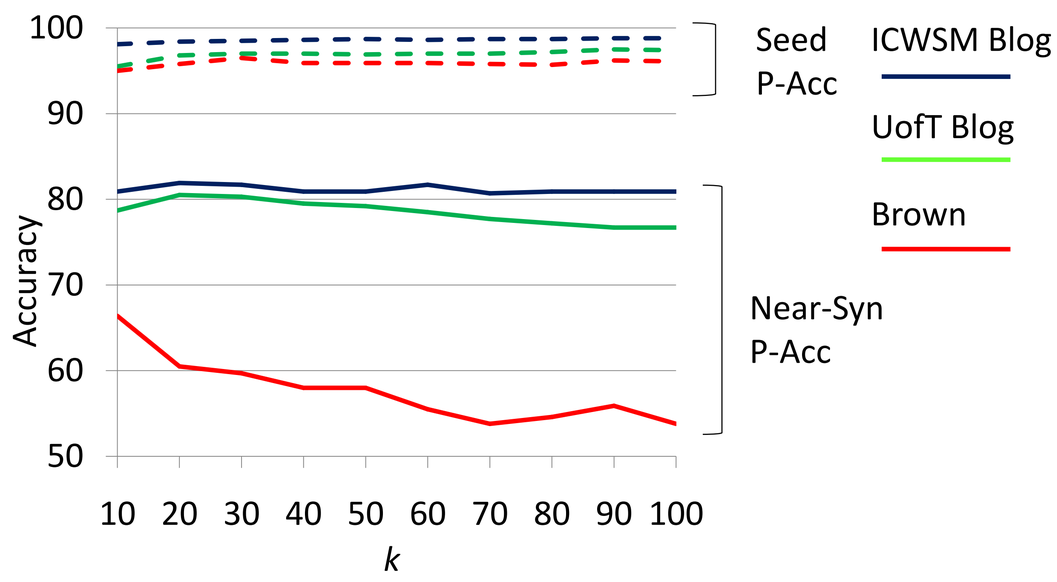
4. Evaluation

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

5. Results

Lexicon Construction Method	Seeds			Near-Syns	
	Cov.	C-Acc.	P-Acc.	Cov.	P-Acc.
Word length	100	86.4	91.8	100	63.7
Latinate affixes	100	74.5	46.3	100	32.6
Word Counts, Brown, informal is rare	51.0	63.7	68.3	59.6	18.5
Word Counts, Brown, formal is rare	51.0	36.3	19.5	59.6	55.0
Ratio, Brown vs. SB	39.7	81.5	85.7	35.6	78.2
Ratio, BNC Written vs. Spoken	60.9	89.2	97.3	38.8	74.3
Ratio, Brown + BNC-W vs. SB + BNC-S	68.7	86.2	96.7	53.4	77.5
PMI (Brown)	51.0	80.6	84.4	59.6	73.2
LSA (Brown, *k*=3, binary, cosine, not lemmatized)	51.0	87.1	94.2	59.6	73.9
LSA (as default, but lemmatized)	50.2	86.9	94.0	54.4	71.9
LSA (as default, but *td·idf*)	51.0	48.4	48.6	59.6	52.9
LSA (as default, but LR, *leave-one-out*)	51.0	75.8	86.8	59.6	58.4
LSA (as default, but filtered)	43.6	87.7	95.5	43.9	74.9
LSA, UofT Blogs (*k*=20, default)	100	91.4	96.8	99.0	80.5
LSA, UofT Blogs (*k*=20, filtered)	99.0	92.1	97.0	97.7	80.5
LSA, ICWSM Blogs (*k*=20, filtered)	100	93.0	98.4	99.7	81.9

- Our corpus methods offer an marked improvement over word length
- LSA with large blog corpora is by far the best individual method
- Lemmatization, weighting, regression generally not effective
- Filters reduce blog word–doc matrices to 1/16 size, no loss of accuracy
- ICWSM lexicon 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.
Brown/BNC-W vs. SB/BNC-D hybrid (BB-SB)	97.1	79.2	79.9	97.5	89.9
3 Agree (WL, ICWSM *k*=20, BB-SB)	67.5	99.4	100	51.6	96.1
2 Agree (WL, ICWSM *k*=20, BB-SB)	86.4	99.0	100	80.5	96.9
SVM (WL, ICWSM *k*=20, BB-SB, Affix)	100	95.9	99.7	100	84.5
Naïve Bayes (WL, ICWSM *k*=20, BB-SB, Affix)	100	97.1	99.5	100	83.7
Average (WL, ICWSM *k*=20, BB-SB)	100	88.5	98.2	100	84.5
Weighted (SVM) (WL, ICWSM *k*=20, BB-SB)	100	93.4	98.7	100	85.7

- Hybrid methods offer performance beyond that of basic methods
- Voting allows for extremely high accuracy at the cost of coverage
- SVM weighted average provides best all-around lexicon; ICWSM (LSA) lexicon twice as useful as word length and word count lexicons

References and Acknowledgements

Biber, Douglas. 1995. *Dimensions of Register Variation: A cross-linguistic comparison*. Cambridge University Press.
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.
Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguististics*, 16(1):22–29.
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.