Inducing Lexicons of Formality from Corpora

Julian Brooke, Tong Wang, Graeme Hirst

Department of Computer Science, University of Toronto
Toronto, ON, Canada M5S 3G4
jbrooke, tong, gh@cs.toronto.edu

Abstract

The spectrum of formality, in particular lexical formality, has been relatively unexplored compared to related work in sentiment lexicon induction (Turney and Littman, 2003). In this paper, we test in some detail several corpus-based methods for deriving real-valued formality lexicons, and evaluate our lexicons using relative formality judgments between word pairs. The results of our evaluation suggest that the problem is tractable but not trivial, and that we will need both larger corpora and more sophisticated methods to capture the full range of linguistic formality.

1. Introduction

The derivation of lexical resources for use in computational applications has been primarily focused on the semantic or denotational relationships among words, for instance the synonym and hyponym relationships encapsulated in a database like WordNet (Fellbaum, 1998). Largely missing from popular resources like WordNet and the General Inquirer (Stone et al., 1966) is information about the formality of a word, which relates directly to the appropriateness of a word in a given context. The concept of formality has of course received a certain amount of interest in computational linguistics, for instance in studies of text generation (Hovy, 1990; Inkpen and Hirst, 2006). The lexical work on formality, however, generally assumes a static, discrete conception of formality. Theoretical and empirical work on genre and register (Leckie-Tarry, 1995; Biber, 1995; Heylighen and Dewaele, 2002) belies this idea; instead, linguistic formality and the dichotomies that underlie formality, e.g. spoken/written, interpersonal/abstract, contextual/context-independent, are generally conceived as dimensions, clines, or spectrums upon which particular genres may vary. Quantification of this spectrum, however, is rarely pursued beyond calculation of the easily countable surface features such as part of speech, providing broad metrics for text classification, but very little that can be applied to more subtle tasks such as word choice. In *Choose the Right Word* (Hayakawa, 1994), a manual intended to help writers select the best English word from among a group of near-synonyms, there is a clear assumption that the notion of a formality spectrum also applies at the lexical level; there, small differences between the formality of words are enumerated using relative, continuous language (i.e. *A is more formal than B* rather than *A is formal*).

In this work, we investigate methods for deriving a continuous spectrum of formality in the form of a formality lexicon. Our work is inspired and informed by the recent interest in sentiment lexicon acquisition that has formed a major part of work in Sentiment Analysis (Turney and Littman, 2003; Esuli and Sebastiani, 2006; Taboada and Voll, 2006; Rao and Ravichandra, 2009). We believe that construction of formality lexicons is a related but distinct problem, and so we will adapt methods used in the Sentiment research but also apply techniques which are distinct to the variations of formality. We predict that formality is a somewhat easier problem, due to the stronger co-occurrence relationships among formal words. One of the goals of this preliminary work is to show, however, that quantifying formality is far from trivial, particularly if the relationships among words are to be applied to tasks that require attention to linguistic detail, for instance word choice or phrase-level formality classification. More generally, we believe that a deeper understanding of formality may lead to applications that allow for capturing the variation of language in ways that avoid the pitfalls of domain-specificity, e.g. the need to train models for any possible location on the spectrum of register. Finally, one of our key goals is to develop methods for deriving lexical formality that are language-independent; the small-scale, corpus-based methods we investigate here are suitable for almost any language for which a varied corpus of a reasonable size is available.

2. Data and Resources

2.1. Word Lists

As the starting point for this work, we collected two lists of words, one formal and one informal, that we use both as seeds for our dictionary construction methods and as test sets for evaluation (our 'gold standard'). We assume that all slang terms are by their vary nature informal and so our 138 informal seeds[1] were pulled primarily from an online slang dictionary[2] (e.g. *wuss*, *grubby*) and also includes some contractions and interjections (e.g. *cuz*, *yikes*). The 105 formal seeds[3] were selected from a list of discourse markers (e.g. *moreover*, *hence*) and adverbs from a sentiment lexicon (e.g. *preposterously*, *inscrutably*); these sources were chosen to avoid words with overt topical content, and to ensure that there was some balance of emotional bias across formal and informal seed sets. The imbalance in the seed set counts (more informal than formal) is offset here by the fact that our formal seeds are much better represented in

This work is supported by the Natural Sciences and Engineering Research Council of Canada.

[1] See http://www.cs.toronto.edu/~jbrooke/informal_seeds.txt

[2] http://onlineslangdictionary.com/

[3] See http://www.cs.toronto.edu/~jbrooke/formal_seeds.txt

our primary corpus.

To allow for a more objective, fine-grained evaluation, we manually extracted a set of 399 pairs of near-synonyms[4] from *Choose the Right Word* (CTRW); all these pairs were either explicitly or implicitly compared for formality in the book. Implicit comparison included blanket statements like *this is the most formal of these words*; in those cases, and more generally, we avoided words appearing in more than one comparison (there are no duplicate words in our CTRW pair set), as well as multiword expressions and words whose formality is strongly ambiguous (i.e. word-sense dependent). An example of this last phenomenon is the word *cool*, which is used colloquially in the sense of *good* but more formally as in the sense of *cold*. Partly as a result of this polysemy, which we observe is more common among informal words, our pairs are clearly biased toward the formal end of the spectrum; although there are some informal comparisons, e.g. *bellyache/whine*, *wisecrack/joke*, more typical pairs include *determine/ascertain* and *hefty/ponderous*. Despite this imbalance, one obvious advantage of using near-synonyms in our evaluation metric is that factors other than linguistic formality (e.g. topic, opinion) are less likely to influence performance.

2.2. Corpora

Our primary corpus for the word co-occurrence methods presented here (section 3.3) is the Brown corpus (Francis and Kučera, 1982). Although extremely small by modern corpus standards, it has the advantage of being compiled explicitly to represent a range of American English genres (and, by extension, formalities). It includes four genres (reportage, formal documents, fiction, and miscellaneous) divided into 15 sub-genres; for our split-corpus method, we consider reportage and formal documents as formal. Its small size (approximately 1 million words in 499 documents) means that our results using it are likely to represent a lower bound rather than anything approaching optimal performance; nonetheless, we have found that it serves as a useful development set for selecting appropriate methods and testing various options. We note here that it contains at least one use of 53 (38%) of our informal seeds and 71 (67%) of our formal seeds. For our word count comparison methods (section 3.2) it is also useful to have a spoken corpus, representing the more informal end of the formality spectrum: for this, we use word counts for another publicly available corpus, the Switchboard (SW) corpus of American telephone conversations (Godfrey et al., 1992), which contains roughly 2400 conversations with over 2.6 million word tokens.

3. Methods

Each method described below derives a formality score (FS) in the range 1 to -1 for any word within its coverage, similar to the quantification of SentiWordNet (Esuli and Sebastiani, 2006). Since some methods do not have full coverage, in our evaluation we will also sometimes consider hybrid methods that back-off to a higher coverage (baseline) model; we do not, however, test more-complex hybrid systems (e.g. weighted sums) here.

3.1. Baselines

The most obvious baseline is based on word length, which is often used directly as an indicator of formality for applications like genre classification (Karlgren and Cutting, 1994). Given a shortest word of length n and a longest word of length m in some vocabulary V (the Brown corpus), we derive FS scores for any word based on this set by dividing up the formality scale into equal partitions; for a word w of length l, the formality score function, $FS(w)$, is given by:

$$FS(w) = -1 + 2\frac{l}{m-n}$$

A special exception is made for hyphenated terms, which can be extremely long in the case when an entire phrase is hyphenated, biasing the maximum word length: for those terms, we use the average length of constituent words rather than the total length. Though this metric works fairly well for English, we note that it might be problematic in a language with word agglutination (e.g. German) or without an alphabet (e.g. Chinese).

Another straightforward baseline is the assumption that Latinate prefixes and suffixes are indicators of formality in English (Kessler et al., 1997), i.e. informal words will not have Latinate affixes like *-ation* and *intra-*. Here, we simply assign words that have appear to have such a prefix or suffix an FS of 1, and all other words an FS of -1.

3.2. Frequency Methods

These methods derive FS based on word counts in corpora. Our first approach assumes a single corpus, where formal words are common and informal words are rare, or vice versa. To smooth out the Zipfian distribution, we use the rank of words as exponentials; for a corpus with R ranks, the FS for a word of rank r under the *formal is rare* assumption is given by:

$$FS(w) = -1 + 2\frac{e^{(r-1)}}{e^{(R-1)}}$$

Under the *informal is rare* assumption:

$$FS(w) = 1 - 2\frac{e^{(r-1)}}{e^{(R-1)}}$$

A more sophisticated method is to use two corpora that are known to vary with respect to formality and use the relative appearance of words in each corpus as the metric. If word appears n times in a (relatively) formal corpus and m times in an informal corpus (and one of m, n is not zero), we derive:

$$FS(w) = -1 + 2\frac{n}{m \times N + n}$$

Here, N is the ratio of the size (in tokens) of the informal corpus (IC) to the formal corpus (FC). We need the constant N so that an imbalance in the size of the corpora does not result in an equivalently skewed distribution of FS.

A hybrid method combines these two models by using the ratio of word counts in two corpora to define the center of the FS spectrum, but single corpus methods to define the edges. Formally, if m and n (word counts for the IC and FC, respectively) are both non-zero, then FS is given by:

$$FS(w) = -0.5 + \frac{n}{m \times N + n}$$

[4] See http://www.cs.toronto.edu/~jbrooke/CTRWpairs.txt

However, if n is zero, FS is given by:

$$FS(w) = -1 + 0.5\frac{e^{(r_{IC}-1)}}{e^{(R_{IC}-1)}}$$

where r_{IC} is the rank of the word in IC, and R_{IC} is the total number of ranks in IC. If m is zero, FS is given by:

$$FS(w) = 1 - 0.5\frac{e^{(r_{FC}-1)}}{e^{(R_{FC}-1)}}$$

where i is the rank of the word in IC, and R_{IC} is the total number of ranks in IC). This function is undefined in the case where m and n are both zero. Here we also consider the effect of lemmatization, treating various inflected forms as a single type.

3.3. Co-occurrence Methods

We test the co-occurrence methods used by Turney and Littman (2003) to derive Semantic Orientation (positive or negative word bias), with small modifications specific to our situation. The general idea is to derive an FS value for any given word by calculating the degree of association between it and the words in our seed sets. One such metric of association is Pointwise Mutual Information (PMI) (Church and Hanks, 1990); we derive probabilities using a word versus document matrix, with the FS of each word calculated as follows:

$$FS(w) = \frac{1}{N}(\sum_{f \in F}\frac{P(w\&f)}{P(w)P(f)} - \sum_{i \in I}\frac{P(w\&i)}{P(w)P(i)})$$

Here, F is the list of formal seeds, I is the list of informal seeds, and N is a normalization factor, either $argmax|FS'(w_F)|$ (for all w $FS'(w) > 0$) or $argmax|FS'(w_I)|$ (for all w, $FS'(w) < 0$), where $FS'(w)$ is the calculation before normalization; this last insures that the FS will be the range 1 to -1. $P(w\&f)$ is the probability (the count) of the word appearing with a particular formal seed in the same document.

The other method used by Turney and Littman, Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997), is a technique for extracting information from a large corpus of texts by (drastically) reducing the dimensionality of a word–passage matrix, i.e. a matrix where the row vectors correspond to the appearance or (weighted) frequency of words in a set of passages (the columns). The mathematical basis for this transformation is singular value decomposition[5]; for the details of the matrix transformations as relevant to this task, we refer the reader to the discussion in Turney and Littman (2003). The number of columns in the compacted matrix is given by the factor k, an important variable in any application of LSA, and one that is best determined by trial and error. Another factor is the size of a passage, which could be as large as a full document or as small as a sentence; here, we consider documents and paragraphs as possible passages[6]. A third variable that we investigated is the weighting of values in the original matrix; Turney and Littman, for instance, used *tf·idf* (term frequency times inverse document frequency), however it was not clear that this was appropriate for our task, and so we tested various possible options (binary, *tf*, *idf*, and *td·idf*). Again, we consider the effect of lemmatization.

Once a k-dimensional vector for each word appearing in a corpus is derived using LSA, a standard method is to use the cosine of the angle between a word and sets of seed words to identify how similar the distribution of the word is to the seeds. In our case, FS calculated as:

$$FS(w) = \frac{1}{N}(\sum_{f \in F}\cos(\theta(w,f)) - \sum_{i \in I}\cos(\theta(w,i)))$$

Again, F and I are the formal and informal seed sets, and N is a normalization factor, calculated in the same way as with PMI, above.

Another method that is available to us, due to the relatively large size of our seed sets, is derivation of FS by means of regression, using machine learning algorithms. We speculate that this might be preferable to the cosine method since the irrelevant dimensions might be discarded from the model, whereas in the cosine calculations these dimensions would show up as noise. To investigate the effectiveness of this approach, we tested various regression algorithms included in the WEKA software suite (Witten and Frank, 2005); below, we present results for two, linear regression and Gaussian processes, which preformed well according to the r^2 value with 10-fold cross-validation; for both we used the default settings for WEKA (version 3.6.2), which for Gaussian processes entails a classifier with an RBF kernel. Training was carried out using the k-dimensional vectors of our formal and informal seeds; for the purposes of training the former were assigned a value of 1, the latter -1. Since the model applied to new data could potentially fall outside that range, appropriate normalization of the output (dividing by the most extreme FS values) is also necessary in this case.

4. Evaluation

We evaluate our lexicon dictionary methods using the gold standard judgments from the seed sets and CTRW word pairs. To differentiate the two, we continue to use the term *seed* for the former; in this context, however, these 'seed sets' are being used as a test set. For computation of the co-occurance-based FS of a word that is part of our seed set, we apply *leave-one-out* cross-validation, removing that word from list of seeds for the purposes of calculating cosine difference from the seeds or when training a model to predict its FS value. The coverage (Cov.) is the percentage of words in the set which appear in the induced dictionary. The class-based accuracy (C-Acc.) is the percentage of words which are correctly classified as formal (FS $>$ 0) or informal (FS $<$ 0). The pair-based accuracy (P-Acc.) is the result of exhaustively pairing words in the

[5] We use the Divisi Python implementation of SVD, http://divisi.media.mit.edu; our vectors are taken from 'weighted U' matrix after SVD is applied and all but the top k singular values are removed.

[6] Preliminary testing with sentences suggested that the resulting matrices were far too sparse to be useful, we omit those results here.

two seed sets and testing their relative formality; that is, for all $w_i \in I$ and $w_f \in F$, the percentage of w_i/w_f pairs where $FS(w_i) < FS(w_f)$. The average FS difference (FS-Dif.) is just $FS(w_i) - FS(w_f)$ for each of the w_i/w_f pairs created as above; we wish to maximize this number on the basis that our seeds represent relatively extreme examples of the formality spectrum. For the CTRW pairs there are only two metrics, the coverage and the pair-based accuracy; since the CTRW pairs represent relative formality of varying degrees, it is not possible to calculate a class-based accuracy and there is no guarantee that the average distance should be maximized.

5. Results

The results of evaluation for all the various methods are shown in Table 1; the numbers in parentheses below indicate the corresponding line of the table. In the first section of the table, the baseline provided by the word length (1) is quite high, particularly for seed set pairwise accuracy, indicating that nearly all the informal seed words are shorter than the formal seed words. Word length is not as effective with the fine-grained differences, however, and the class-based accuracy is low, as many formal seeds are incorrectly labeled as informal using our linear method.[7] It is clear from the class-based accuracy score that Latinate suffixes and prefixes (2) are indicative of formality; they do not, however, provide information that allows for relative, more fine-grained distinctions. The advantage of these methods, of course, is their coverage.

The first two results in the second part of Table 1 (3–4) show that neither assumption (i.e. that formal words are rare or that informal words are rare) is particularly successful, though they fail in different ways that are indicative of the formality make-up of the corpus and the test sets. Since the Brown corpus is a corpus of published written texts, and therefore more formal, the *informal is rare* hypothesis (3) is a better one for the extreme seed sets; however, in the CTRW test sets, which is more indicative of the formal end of the spectrum, this assumption fails spectacularly, with the model performing much worse than chance. The opposite is true for the *formal is rare* model (4), since it makes opposite predictions. Neither is directly useful for the task as a whole.

Much better is the word ratio model using the Brown corpus as the formal dictionary and the Switchboard corpus as the informal dictionary (5); although the coverage is quite low, the score for pairwise accuracy in the CTRW set is the highest in Table 1, and the scores for the seed test are also quite good. The hybrid model, with the ratio model converting the middle of the spectrum and the *rare* models applied at either end (6), provides us with the best class-based accuracy in the table, and comparable performance among CTRW pairs with a 20% increase in coverage. A hybrid model that splits the Brown corpus into two halves (7), i.e. the relatively formal genres of reportage and formal documents and the relatively informal genres of fiction and miscellaneous, does not, however, perform nearly as well, suggesting that very distinct corpora are required for this method to be useful. The effects of lemmatization (8) are harder to interpret; the drop for seed words is marked, but there is a modest increase for CTRW, and a small boost in coverage. In general, this suggests that the inflectional differences among words might be somewhat indicative of formality, and should not necessarily be disregarded. Finally, the use of a word length backoff (9) provides superior performance with respect to the seed sets, but is slightly worse than the word length baseline in the CTRW set.

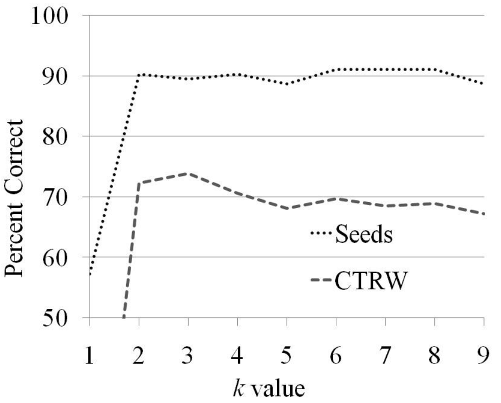

Figure 1: Seed class-based accuracy and CTRW pairwise accuracy, LSA cosine method for various k, $1 \leq k < 10$

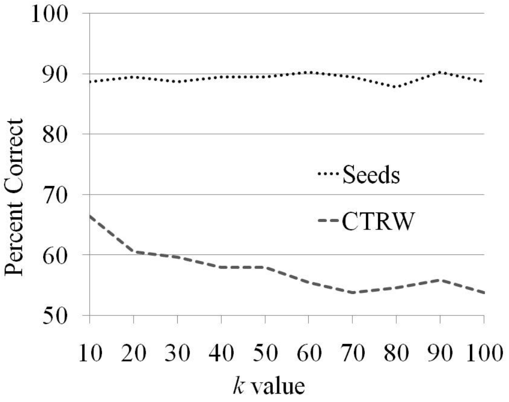

Figure 2: Seed class-based accuracy and CTRW pairwise accuracy, LSA cosine method for various k, $10 \leq k \leq 100$

The co-occurrence results are presented in the third part of Table 1. The PMI results (10) are quite promising, given the simple nature of the calculation, though the LSA results (11–14) are better, particularly when the optimal value of k is used (14). To find that value, we tested all values between 1 and 10, and at intervals of 10 thereafter; graphs showing the class-based accuracy for seeds and pairwise accuracy for the CTRW set are presented in Figures 1 and 2.

For the CTRW set, the performance peaks at $k = 3$; beyond $k = 3$, the overall trend is down, though there are small jumps at particular values of k, and we often see corresponding fluctuations in seed set performance, even though the overall picture there is much flatter. We posit

[7] Switching to logarithmic FS function for word length would likely improve the class-based accuracy, though fine-tuning this function would take us beyond a simple baseline, and have no effect on the pairwise accuracy.

	Seed set				CTRW set	
Dictionary construction method	**Cov.**	**C-Acc.**	**P-Acc.**	**FS-Dif.**	**Cov.**	**P-Acc.**
Baseline methods						
(1) Word length	100	74.9	91.8	0.49	100	63.7
(2) Latinate affixes	100	74.5	46.3	0.86	100	32.6
Word count methods						
(3) Word Counts, Brown, informal is rare,	51	63.7	68.3	0.45	59	18.5
(4) Word Counts, Brown, formal is rare	51	36.3	19.5	−0.45	59	55.0
(5) Ratio, Brown and Switchboard	38	81.5	85.7	0.75	36	78.2
(6) Hybrid, Brown and Switchboard	58	90.8	89.4	0.81	56	74.3
(7) Hybrid, split Brown	51	51.6	70.0	0.49	60	38.2
(8) Hybrid, Brown and Switchboard, lemma	63	78.6	79.1	0.51	62	77.0
(9) Hybrid, Brown and Switchboard, WL backoff	100	87.7	92.3	0.72	100	61.2
Co-occurence methods						
(10) PMI, Brown	51	80.6	84.4	0.33	60	73.2
(11) LSA (*k*=100), cosine (Brown, binary, document)	51	88.7	96.1	0.74	60	53.8
(12) LSA (*k*=10), cosine	51	88.7	95.0	1.00	60	66.4
(13) LSA (*k*=3), cosine	51	89.5	94.5	1.07	60	73.9
(14) LSA (*k*=3), cosine, WL backoff	100	88.9	95.1	0.94	100	62.2
(15) LSA (*k*=3), cosine, lemma	51	88.5	94.4	1.02	60	70.5
(16) LSA (*k*=100), cosine, paragraph	51	83.1	96.6	0.65	60	53.8
(17) LSA (*k*=10), cosine, paragraph	51	83.1	95.0	0.86	60	61.8
(18) LSA (*k*=3), cosine, paragraph	51	83.1	91.7	0.86	60	73.5
(19) LSA (*k*=3), *tf*, cosine	51	66.1	74.9	49.2	60	49.2
(20) LSA (*k*=3), *idf*, cosine	51	55.6	57.7	0.02	60	52.5
(21) LSA (*k*=3), *td·idf*, cosine	51	54.8	39.7	−0.07	60	52.5
(22) LSA (*k*=100), Gaussian	51	71.8	83.8	0.42	60	38.2
(23) LSA (*k*=10), Gaussian	51	81.5	92.3	0.45	60	56.3
(24) LSA (*k*=3), Gaussian	51	87.1	92,7	0.39	60	56.7
(25) LSA (*k*=100), linear	51	58.9	57.6	0.04	60	53.4
(26) LSA (*k*=10), linear	51	79.0	88.9	0.12	60	58.4
(27) LSA (*k*=3), linear	51	75.8	86.8	0.14	60	61.8

Table 1: Seed coverage (%), class-based accuracy (%), pairwise accuracy (%), average FS difference, CTRW coverage (%) and pairwise accuracy (%) for various FS dictionaries

that the more fine-grained CTRW set is much more sensitive to the noise that comes with the increase in dimensionality; clearly, the second dimension (the one that is 'added' at $k = 2$) is the strongest indicator of formality, and though other dimensions (e.g. $k = 3,6$) also provide information that boost performance. More generally, however, the addition of dimensions is a losing proposition, as the best dimensions for detecting formality are among the first discovered by using the LSA method, and beyond that the noise outweighs the relevant information. The results with a word length backoff suggest that overall the LSA method is slightly better than the hybrid word-count method, though the differences are not significant.

Looking at the options for LSA, lemmatization (15) has a small but consistently negative effect. More notable is the drop in performance when paragraphs rather than documents are taken as the unit in our word–passage matrix (16-18), suggesting that a *one formality per document* assumption is a relatively good one; the pairwise accuracy in the seed sets, though, is consistently high. With respect to weights, our original intuition was that a binary feature for appearance in a document was the best way to approach the construction of a word-document matrix; intuitively, there does not seem to be useful information that can be gleaned from the number of appearances of a formal or informal word in a document, nor should a word be weighted solely based on its rarity in a corpus. Indeed, our results (19–21) confirm this; applying *td·idf* or either of its component results in a major drop in performance across the board.

Finally, we look at the results using machine learning regression methods rather than cosine distance to derive FS (22–27). Neither of the algorithms perform well on the CTRW set, with the Gaussian Processes method (22–24) particularly poor, despite its relative sophistication; one explanation is that it tries to maximize the extreme cases, failing on the more-subtle word distinctions. The performance differences related to increases in k (22, 25) are consistent with cosine but more marked, revealing themselves in all three accuracy measures, though with a great deal more variation across the methods. Regression might prove to be more effective with more-numerous and more-nuanced training examples (for instance, including seed words that represent the middle of the spectrum).

One gratifying result is that, despite particular inconsistencies, the four performance metrics used here show clear correlation; for instance, even when the seed accuracies are flat, increases in k are associated with both a drop in CRTW accuracy and a drop in the average FS difference between seed words. Thus, we can be confident that the performance differences among our models are robust, reflecting variation across the full spectrum of formality.

6. Conclusions and Future Work

Though preliminary, the work we have presented in this paper suggests that quantifying formality is a tractable but not trivial problem. Surprisingly, despite significant variation in the underlying features from which they are derived, several of the models investigated here reached an impressive accuracy in distinguishing extreme differences in formality, using information derived from a small yet diverse corpus. Less encouraging, however, is the performance of these same methods in identifying more-subtle variations among near-synonyms; at present, our guess based on the word count and co-occurrence is no better than one based simply on word length.

The next step in this project will involve an expansion of our data. There are a number of larger publicly available corpora that could be applied to our problem, for instance the British National Corpus (Burnard, 2000); informal testing suggests that word count information from the BNC will easily boost our word-count performance well beyond the baselines provided by word length. Blogs are a natural, inexhaustible source of information on register variation, though there are potential pitfalls and challenges related to using large amounts of web data, in particular the fact that LSA, our most promising method, does not scale up well (Turney and Littman, 2003).

With respect to refining our methods, one way forward is to see how the information represented by these various methods can be integrated to improve performance, i.e. with some kind of meta-classifier. There is certainly room for the methods to inform each other, since agreement for our best word count classifier and best co-occurrence classifier in the CTRW test set is a mere 66.3%, almost 10% below the accuracy in both cases; agreement on the seed sets is much higher, of course, but still below 90% for both metrics. One difficulty here is the lack of reliable training data, and one option we are exploring is the use of semi-automated methods to derive larger, more objective seed sets. A related idea is to use, for instance, word count or PMI FS as a starting point, and then use the LSA co-occurence information to iteratively refine those scores until convergence. In short, the methods described here just represent a basic toolbox for the continued exploration of the formality spectrum, moving beyond English-specific approaches to those that can be applied in any language.

7. References

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

Lou Burnard. 2000. User reference guide for British National Corpus. Technical report, Oxford University.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguististics*, 16(1):22–29.

Andrea Esuli and Fabrizio Sebastiani. 2006. Senti-WordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Internation Conference on Language Resources and Evaluation(LREC)*, Genova, Italy.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Nelson Francis and Henry Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin, Boston.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:517–520.

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

Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3):293–340.

Eduard H. Hovy. 1990. Pragmatics and natural language generation. *Artificial Intelligence*, 43:153–197.

Diana Inkpen and Graeme Hirst. 2006. Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics*, 32(2):223–262.

Jussi Karlgren and Douglas Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1071–1075.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38.

Thomas K. Landauer 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.

Helen Leckie-Tarry. 1995. *Language Context: a functional linguistic theory of register.* Pinter.

Delip Rao and Deepak Ravichandra. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Lingusitics*, Athens, Greece.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilivie. 1966. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press.

Maite Taboada and Kimberly Voll. 2006. Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th Internation Conference on Language Resources and Evaluation (LREC)*, Genova, Italy.

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

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, San Francisco.