Automatic Acquisition of Lexical Formality

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

Abstract

There has been relatively little work focused on determining the formality level of individual lexical items. This study applies information from large mixed-genre corpora, demonstrating that significant improvement is possible over simple word-length metrics, particularly when multiple sources of information, i.e. word length, word counts, and word association, are integrated. Our best hybrid system reaches 86% accuracy on an English near-synonym formality identification task, and near perfect accuracy when comparing words with extreme formality differences. We also test our word association method in Chinese, a language where word length is not an appropriate metric for formality.

1 Introduction

The derivation of lexical resources for use in computational applications has been focused primarily on the denotational relationships among words, e.g. the synonym and hyponym relationships encapsulated in WordNet (Fellbaum, 1998). Largely missing from popular lexical resources such as WordNet and the General Inquirer (Stone et al., 1966) is stylistic information; there are, for instance, no resources which provide comprehensive information about the formality level of words, which relates to the appropriateness of a word in a given context. Consider, for example, the problem of choice among near-synonyms: there are only minor denotational differences among synonyms such as *get*, *acquire*, *obtain*, and *snag*, but it is difficult to construct a situation where any choice would be equally suitable. The key difference between these words is their formality, with *acquire* the most formal and *snag* the most informal.

In this work, we conceive of formality as a continuous property. This approach is inspired by resources such as *Choose The Right Word* (Hayakawa, 1994), in which differences between synonyms are generally described in relative rather than absolute terms, as well as linguistic literature in which the quantification of stylistic differences among genres is framed in terms of dimensions rather than discrete properties (Biber, 1995). We begin by defining the *formality score* for a word as a real number value in the range 1 to -1, with 1 representing an extremely formal word, and -1 an extremely informal word. A formality lexicon, then, gives a FS score to every word within its coverage.

The core of our approach to the problem of classifying lexical formality is the automated creation of formality lexicons from large corpora. In this paper, we focus on the somewhat low-level task of identifying the relative formality of word pairs; we believe, however, that a better understanding of lexical formality is relevant to a number of problems in computational linguistics, including sub-fields such as text generation, error correction of (ESL) writing, machine translation, text classification, text simplification, word-sense disambiguation, and sentiment analysis. One conclusion of our research is that formality variation is omnipresent in natural corpora, but it does not follow that the identification of these differences on the lexical level is a trivial one; nevertheless,

Coling 2008: Poster Volume, pages 90–98,
Beijing, August 2010

we are able to make significant progress using the methods presented here, in particular the application of latent semantic analysis to blog corpora.

2 Related Work

As far as we are aware, there are only a few lines of research explicitly focused on the question of linguistic formality. In linguistics proper, the study of register and genre usually involves a number of dimensions or clines, sometimes explicitly identified as formality (Leckie-Tarry, 1995; Carter, 1998), or decomposed into notions such as informational versus interpersonal content (Biber, 1995). Heyligen and Dewaele (1998) provide a part-of-speech based quantification of textual contextuality (which they argue is fundamental to the notion of formality); their metric has been used, for instance, in a computational investigation of the formality of online encyclopedias (Emigh and Herring, 2005). In this kind of quantification, however, there is little, if any, focus on individual elements of the lexicon. In computational linguistics, formality has received attention in the context of text generation (Hovy, 1990); of particular note relevant to our research is the work of Inkpen and Hirst (2006), who derive boolean formality tags from *Choose the Right Word* (Hayakawa, 1994). Like us, their focus was improved word choice, though the approach was much broader, also including dimensions such as polarity. An intriguing example of formality relevant to text classification is the use of informal language (slang) to help distinguish true news from satire (Burfoot and Baldwin, 2009).

Our approach to this task is inspired and informed by automatic lexical acquisition research within the field of sentiment analysis (Turney and Littman, 2003; Esuli and Sebastiani, 2006; Taboada and Voll, 2006; Rao and Ravichandra, 2009). Turney and Littman (2003) apply latent semantic analysis (LSA) (Landauer and Dumais, 1997) and pointwise mutual information (PMI) to derive semantic orientation ratings for words using large corpora; like us, they found that LSA was a powerful technique for deriving this lexical information. The lexical database SentiWordNet (Esuli and Sebastiani, 2006) provides 0–1 rankings for positive, negative, and neutral polarity, derived automatically using relationships between words in WordNet (Fellbaum, 1998). Unfortunately, WordNet synsets tend to cut across the formal/informal distinction, and so the resource is not obviously useful for our task.

The work presented here builds directly on a pilot study (Brooke et al., 2010), the focus of which was the construction of formality score (FS) lexicons. In that work, we employed less sophisticated forms of some of the methods used here in a relatively small dataset (the Brown Corpus), providing a proof of concept, but with poor coverage, and with no attempt to combine the methods to maximize performance. However, the small dataset allowed us to do a thorough test of certain options associated with our task. In particular we found that using a similarity metric based on LSA gave good performance across our test sets, especially when the term-document matrix was binary (unweighted), the k-value used for LSA was small, and the method used to derive a formality score was cosine similarity to our seed terms. A metric using total word counts in corpora with divergent formality also showed promise, with both methods performing above our word-length baseline for words within their coverage. PMI, by comparison, proved less effective, and we do not pursue it further here.

3 Data and Resources

3.1 Word Lists

All the word lists discussed here are publicly available.[1] We begin with two, one formal and one informal, that we use both as seeds for our lexicon construction methods and as test sets for evaluation (our gold standard). We assume that all slang terms are by their very nature informal and so our 138 informal seeds were taken primarily from an online slang dictionary[2] (e.g. *wuss*, *grubby*) and also include some contractions and interjections (e.g. *cuz*, *yikes*). The 105 formal seeds 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

[1] http://www.cs.toronto.edu/~jbrooke/FormalityLists.zip

[2] http://onlineslangdictionary.com/

overt topic, and to ensure that there was some balance of sentiment across formal and informal seed sets. Part of speech, however, is not balanced across our seed sets.

Another test set we use to evaluate our methods is a collection of 399 pairs of near-synonyms from *Choose the Right Word* (CTRW), a manual for assisting writers with synonym word choice; each pair was either explicitly or implicitly compared for formality in the book. Implicit comparison included statements such as *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 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 is clearly more common among informal words, our pairs are biased toward the formal end of the spectrum; although there are some informal comparisons, e.g. *belly-ache/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 is that factors other than linguistic formality (e.g. topic, opinion) are less likely to influence performance. In general, the CTRW allows for a more objective, fine-grained evaluation of our methods, and is oriented towards our primary interest, near-synonym word choice.

To test the performance of our unsupervised method beyond English, one of the authors (a native speaker of Mandarin Chinese) created two sets of Chinese two-character words, one formal, one informal, based on but not limited to the words in the English sets. The Chinese seeds include 49 formal seeds and 43 informal seeds.

3.2 Corpora

Our corpora fall generally into three categories: formal (written) copora, informal (spoken) corpora, and mixed corpora. The Brown Corpus (Francis and Kučera, 1982), our development corpus, is used here both as a formal and mixed corpus. Although extremely small by modern corpus standards (only 1 million words), the Brown Corpus has the advantage of being compiled explicitly to represent a range of American English, though it is all of the published, written variety. The Switchboard (SW) Corpus is a collection of American telephone conversations (Godfrey et al., 1992), which contains roughly 2400 conversations with over 2.6 million word tokens; we use it as an informal counterpart to the Brown Corpus. Like the Brown Corpus, The British National Corpus (Burnard, 2000) is a manually-constructed mixed-genre corpus; it is, however, much larger (roughly 100 million words). It contains a written portion (90%), which we use as a formal corpus, and a spontaneous spoken portion (4.3%), which we use as an informal corpus. Our other mixed corpora are two blog collections available to us: the first, which we call our development blog corpus (DevBlog) contains a total of over 900,000 English blogs, with 216 million tokens.[3] The second is the 'first tier' English blogs included in the publicly available ICSWM 2009 Spinn3r Dataset (Burton et al., 2009), a total of about 1.3 billion word tokens in 7.5 million documents. For our investigations in Chinese, we use the Chinese portion of the ICSWM blogs, approximately 25.4 million character tokens in 86,000 documents.

4 Methods

4.1 Simple Formality Measures

The simplest kind of formality measure is based on word length, which is often used directly as an indicator of formality for applications such as genre classification (Karlgren and Cutting, 1994). Here, we use logarithmic scaling to derive a FS score based on word length. Given a maximum word length L[4] and a word w of length l, the formality score function, $FS(w)$, is given by:

$$FS(w) = -1 + 2\frac{\log l}{\log L}$$

[3] These blogs were gathered by the University of Toronto Blogscope project (www.blogscope.net) over a week in May 2008.

[4] We use an upper bound of 28 characters, which is the length of *antidisestablishmentarianism*, the prototypical longest word in English; this value of L provides an appropriate formality/informality threshold, between 5- and 6-letter words

For hyphenated terms, the length of each component is averaged. Though this metric works relatively well for English, we note that it is problematic in a language with significant word agglutination (e.g. German) or without an alphabet (e.g. Chinese, see below).

Another straightforward method 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 such as *-ation* and *intra-*. Here, we simply assign words that appear to have such a prefix or suffix an FS of 1, and all other words an FS of -1.

Our frequency methods derive FS from word counts in corpora. Our first, naive approach assumes a single corpus, where either formal words are common and informal words are rare, or vice versa. To smooth out the Zipfian distribution, we use the frequency rank of words as exponentials; for a corpus with R frequency 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:

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

We have previously shown that these methods are not particularly effective on their own (Brooke et al., 2010), but we note that they provide useful information for a hybrid system.

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.

4.2 Latent Semantic Analysis

Next, we turn to LSA, a technique for extracting information from a large corpus of texts by (drastically) reducing the dimensionality of a term–document matrix, i.e. a matrix where the row vectors correspond to the appearance or (weighted) frequency of words in a set of texts. In essence, LSA simplifies the variation of words across a collection of texts, exploiting document–document correlation to produce information about the k most important dimensions of variation ($k <$ total number of documents), which are generally thought to represent semantic concepts, i.e. topic. The mathematical basis for this transformation is singular value decomposition[5]; for the details of the matrix transformations, we refer the reader to the discussion of Turney and Littman (2003). The factor k, the number of columns in the compacted matrix, is an important variable in any application of LSA, one is generally determined by trial and error (Turney and Littman, 2003).

LSA is computationally intensive; in order to apply it to extremely large blog corpora, we need to filter the documents and terms before building our term–document matrix. We adopt the following strategy: to limit the number of documents in our term–document matrix, we first remove documents less than 100 tokens in length, with the rationale that these documents provide less co-occurrence information. Second, we remove documents that either do not contain any target words (i.e. one of our seeds or CTRW test words), or contain only target words which are among the most common 20 in the corpus; these documents are less likely to provide us with useful information, and the very common target terms will be well represented regardless. We further shrink the set of terms by removing all hapax legomena; a single appearance in a corpus is not enough to provide reliable co-occurrence information, and roughly half the words in our blog corpora appear only once. Finally, we remove symbols and all words which are not entirely lower

[5] We use the implementation included in Matlab; we take the rows of the decomposed U matrix weighted by the singular values in Σ for our word vectors. Using no weights or Σ^{-1} generally resulted in worse performance, particularly with the CTRW sets.

case; we are not interested, for instance, in numbers, acronyms, and proper nouns. We can estimate the effect this filtering has on performance by testing it both ways in a development corpus.

Once a k-dimensional vector for each relevant word is derived using LSA, a standard method is to use the cosine of the angle between a word vector and the vectors of seed words to identify how similar the distribution of the word is to the distribution of the seeds. To begin, each formal seed is assigned a FS value of 1, each informal seed a FS value of -1, and then a raw seed similarity score (FS') is calculated for each word w:

$$FS'(w) = \sum_{s \in S, s \neq w} W_s \times FS(s) \times \cos(\theta(w,s))$$

S is the set of all seeds. Note that seed terms are excluded from their own FS calculation, this is equivalent to *leave-one-out* cross-validation. W_s is a weight that depends on whether s is a formal or informal seed, W_i (for informal seeds) is calculated as:

$$W_i = \frac{\sum_{f \in F} FS(f)}{|\sum_{i \in I} FS(i)| + \sum_{f \in F} FS(f)}$$

and W_f (for formal seeds) is:

$$W_f = \frac{|\sum_{i \in I} FS(i)|}{|\sum_{i \in I} FS(i)| + \sum_{f \in F} FS(f)}$$

Here, I is the set of all informal seeds, and F is the set of all formal seeds. These weights have the effect of countering any imbalance in the seed set, as formal and informal seeds ultimately have the same (potential) influence on each word, regardless of their count. This weighting is necessary for the iterative extension of this method discussed in the next section.

We calculate the final FS score as follows:

$$FS(w) = \frac{FS'(w) - FS'(r)}{N_w}$$

The word r is a reference term, a common function word that has no formality.[6] This has the effect of countering any (moderate) bias that might exist in the corpus; in the Brown Corpus, for instance, function words have positive formality before this step, simply because formal words occurred more often in the corpus. N_w is a normalization factor, either

$$N_w = \max_{w_i \in I'} |FS'(w_i) - FS'(r)|$$

for all $w_i \in I'$ or

$$N_w = \max_{w_f \in F'} |FS'(w_f) - FS'(r)|$$

for all $w_f \in F'$. I' contains all words w such that $FS'(w) - FS'(r) < 0$, and F' contains all words w such that $FS'(w) - FS'(r) > 0$. This ensures that the resulting lexicon has terms exactly in the range 1 to -1, with the reference word r at the midpoint.

We also tested the LSA method in Chinese. The only major relevant difference between Chinese and English is word segmentation: Chinese does not have spaces between words. To sidestep this problem, we simply included all character bigrams found in our corpus. The drawback of this approach in the inclusion of a huge number of nonsense 'words' (1.3 million terms in just 86,000 documents), however we are at least certain to identify all instances of our seeds.

4.3 Hybrid Methods

There are a number of ways to leverage the information we derive from our basic methods. One intriguing option is to use the basic FS measures as the starting point for an iterative process using the LSA cosine similarity. Under this paradigm, all words in the starting FS lexicon are potential seed words; we choose a cutoff value for inclusion in the seed word set (e.g. words which have at least .5 or $-.5$ FS), and then carry out the cosine calculations, as above, to derive new FS values (a new FS lexicon). We can repeat this process as many times as required, with the idea that the connections between various words (as reflected in their LSA-derived vectors) will cause the system to converge towards the true FS values.

A simple hybrid method that combines the two word count models uses the ratio of word counts in two corpora to define the center of the FS spectrum, but single corpus methods to define the extremes. Formally, if m and n (word counts for the

[6] The particular choice of this word is relatively unimportant; common function words all have essentially the same LSA vectors because they appear at least once in nearly every document of any size. For English, we chose r = *and*, and for Chinese, r = *yinwei* (*because*); there does not seem to be an obvious two-character, formality-neutral equivalent to *and* in Chinese.

informal corpus *IC* and formal corpus *FC*, respectively) are both non-zero, then FS is given by:

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

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

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

where r_{IC} is the frequency 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:

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

where i is the rank of the word in IC, and R_{IC} is the total number of frequency ranks in IC). This function is undefined in the case where m and n are both zero. Intuitively, this is a kind of backoff, relying on the idea that words of extreme formality are rare even in a corpus of corresponding formality, whereas words in the *core vocabulary* (Carter, 1998), which are only moderately formal, will appear in all kinds of corpora, and thus are amenable to the ratio method.

Finally, we explore a number of ways to combine lexicons directly. The motivation for this is that the lexicons have different strengths and weaknesses, representing partially independent information. An obvious method is an averaging or other linear combination of the scores, but we also investigate vote-based methods (requiring agreement among n dictionaries). Beyond these simple options, we test support vector machines and naive Bayes classification using the WEKA software suite (Witten and Frank, 2005), applying 10-fold cross-validation using default WEKA settings for each classifier. The features here are task dependent (see Section 5); for the pairwise task, we use the difference between the FS value of the words in each lexicon, rather than their individual scores. Finally, we can use the weights from the SVM model of the CTRW (pairwise) task to interpolate an optimal formality lexicon.

5 Evaluation

We evaluate our 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 viewed as a test set (recall that our LSA method is equivalent to *leave-one-out* cross-validation).

We derive the following measures: first, the coverage (Cov.) is the percentage of words in the set that are covered under the method. The class-based accuracy (C-Acc.) of our seed sets is the percentage of covered 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 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 $FS(w_i) < FS(w_f)$. 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.

The first section of Table 1 provides the results for the basic methods in various corpora. The word length (1) and morphology-based (2) methods provide good coverage, but poor accuracy, while the word count ratio methods (3–4) are fairly accurate, but suffer from low coverage. The LSA results in Table 1 are the best for each corpus across the k values we tested. When both coverage and accuracy are considered, there is a clear benefit associated with increasing the amount of data, though the difference between the Dev-Blog and ICWSM suggests diminishing returns. The performance of the filtered Dev-Blog is actually slightly better than the unfiltered versions (though there is a drop in coverage), suggesting that filtering is a good strategy.

In our previous work (Brooke et al., 2010), we noted that CTRW set performance in the Brown dropped for $k > 3$, while performance on the seed set was mostly steady as k increased. Figure 1 shows the pairwise performance of each test set for various corpora across various k. The results here are similar; all three corpora reach a CTRW maximum at a relatively low k values (though higher than Brown Corpus); however the seed set performance in each corpus continues to improve (though marginally) as k increases, while CTRW performance drops. An explanation for this is that

Table 1: Seed coverage, class-based accuracy, pairwise accuracy, CTRW coverage, and pairwise accuracy for various FS lexicons and hybrid methods (%).

Method	Seed set			CTRW set	
	Cov.	**C-Acc.**	**P-Acc.**	**Cov.**	**P-Acc.**
Simple					
(1) Word length	100	86.4	91.8	100	63.7
(2) Latinate affix	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), Dev-Blog	100	91.4	96.8	99.0	80.5
(8) LSA (k=20), Dev-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
Hybrid					
(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

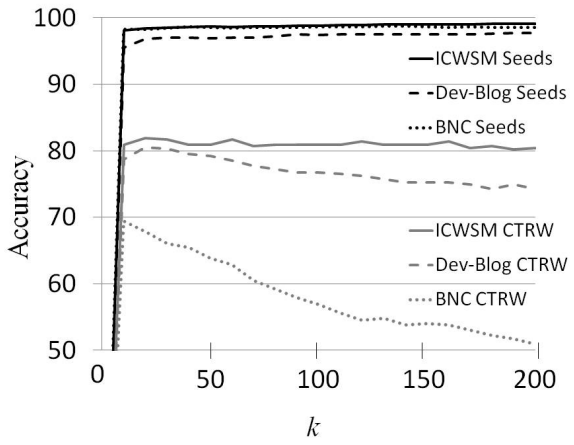

Figure 1: Seed and CTRW pairwise accuracy, LSA method for large corpora k, $10 \leq k \leq 200$.

the seed terms represent extreme examples of formality; thus there are numerous semantic dimensions to distinguish them. However, the CTRW set includes near-synonyms, many with only relatively subtle differences in formality; for these pairs, it is important to focus on the core dimensions relevant to formality, which are among the first discovered in a factor analysis of mixed-register texts (Biber, 1995).

With regards to hybrid methods, we first briefly summarize our testing with the iterative model, which included extensive experiments using basic lexicons and the LSA vectors derived from the Brown Corpus, and some targeted testing with the blog corpora (iteration on these corpora is extraordinarily time-consuming). In general, we found only that there were only small, inconsistent benefits to be gained from the iterative ap-

proach. More generally, the intuition behind the iterative method, i.e. that performance would increase with an drastic increase in the number of seeds, was found to be flawed: in other testing, we found that we could randomly remove most of the seeds without negatively affecting performance. Even at relatively high k values, it seems that a few seeds are enough to calibrate the model.

The ratio (with backoff) hybrid built from the BNC (10) provides CTRW performance that is comparable the best LSA models, though performance in the seed sets is somewhat poor; supplementing with word counts from the Brown Corpus and Switchboard Corpus provides a small improvement (11). The weighed hybrid dictionaries in (12,13) demonstrate that it is possible to effectively combine lexicons built using two different methods on the same corpus (12) or the same method on different corpora (13); the former, in particular, provides an impressive boost to CTRW accuracy, indicating that word count and word association methods are partially independent.

The remainder of Table 1 shows the best results using voting, averaging, and weighting. The voting results (14–16) indicate that it is possible to sacrifice some coverage for very high accuracy in both sets, including a near-perfect score in the seed sets and significant gains in CTRW performance. In general, the best accuracy without a significant loss of coverage came from 2 of 3 voting (15–16), using dictionaries that represented our three basic sources of information (word length, word count, and word association). The machine learning hybrids (17–18) also demonstrate a marked improvement over any single lexicon, though it is important to note that each accuracy score here reflects a different task-specific model. Hybrid FS lexicons built with the weights learned by the SVM models (19–20) provide superior performance on the task corresponding to the model used, though the simple averaging of the best dictionaries (21) also provides good performance across all evaluation metrics.

Finally, the LSA results for Chinese are modest but promising, given the relatively small scale of our experiments: we saw a pairwise accuracy of 82.2%, with 79.3% class-based accuracy ($k = 10$). We believe that the main reason for the generally lower performance in Chinese (as compared to English) is the modest size of the corpus, though our simplistic character bigram term extraction technique may also play a role. As mentioned, smaller seed sets do not seem to be an issue. Interestingly, the class-based accuracy is 10.8% lower if no reference word is used to calibrate the divide between formal and informal, suggesting a rather biased corpus (towards informality); in English, by comparison, the reference-word normalization had a slightly negative effect on the LSA results, though the effect mostly disappeared after hybridization. The obvious next step is to integrate a Chinese word segmenter, and use a larger corpus. We could also try word count methods, though finding appropriate (balanced) resouces similar to the BNC might be a challenge; (mixed) blog corpora, on the other hand, are easily collected.

6 Conclusion

In this work, we have experimented with a number of different methods and source corpora for determining the formality level of lexical items, with the implicit goal of distinguishing the formality of near-synonym pairs. Our methods show marked improvement over simple word-length metrics; when multiple sources of information, i.e. word length, word counts, and word association, are integrated, we are able to reach over 85% performance on the near-synonym task, and close to 100% accuracy when comparing words with extreme formality differences; our voting methods show that even higher precision is possible. We have also demonstrated that our LSA word association method can be applied to a language where word length is not an appropriate metric of formality, though the results here are preliminary. Other potential future work includes addressing a wider range of phenomena, for instance assigning formality scores to morphological elements, syntactic cues, and multi-word expressions, and demonstrating that a formality lexicon can be usefully applied to other NLP tasks.

Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council of Canada. Thanks to Paul Cook for his ICWSM corpus API.

References

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 Language Resources and Evaluation Conference (LREC '10), Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods*.

Burfoot, Clint and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computationsl Linguistics and the 4th International Joint Conference on Nautral Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP '09), Short Papers*, Singapore.

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

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.

Carter, Ronald. 1998. *Vocabulary: applied linguistic perspectives*. Routledge, London.

Emigh, William and Susan C. Herring. 2005. Collaborative authoring on the web: A genre analysis of online encyclopedias. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS '05)*.

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

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

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

Godfrey, J.J., 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.

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

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

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

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

Karlgren, Jussi 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.

Kessler, Brett, 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.

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. 1995. *Language Context: a functional linguistic theory of register*. Pinter.

Rao, Delip 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.

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

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

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

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