Hybrid Models for Lexical Acquisition of Correlated Styles

Julian Brooke
Department of Computer Science
University of Toronto
jbrooke@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
gh@cs.toronto.edu

Abstract

Automated lexicon acquisition from corpora represents one way that large datasets can be leveraged to provide resources for a variety of NLP tasks. Our work applies techniques popularized in sentiment lexicon acquisition and topic modeling to the broader task of creating a stylistic lexicon. A novel aspect of our approach is a focus on multiple related styles, first extracting initial independent estimates of style based on co-occurrence with seeds in a large corpus, and then refining those estimates based on the relationship between styles. We compare various promising implementation options, including vector space, Bayesian, and graph-based representations, and conclude that a hybrid approach is indeed warranted.

1 Introduction

Though lexical resources are useful for many NLP tasks, manual lexicon creation is often onerous, particularly for aspects of language for where full coverage requires hundred of thousands of annotations. This work deals with one such aspect which we refer to as stylistic variation. This should not be understood in a purely aesthetic sense, but as reflecting various high-level aspects of the text, including genre and social identity. Some tasks relevant to style so defined include genre classification (Kessler et al., 1997), author profiling (Rosenthal and McKeown, 2011), social relationship classification (Peterson et al., 2011), sentiment analysis (Wilson et al., 2005), readability classification (Collins-Thompson and Callan, 2005), and text generation (Hovy, 1990). Following the classic work of Biber (1988), computational modeling of style has often focused on textual statistics and the frequency of function words and syntactic categories. There are, of course, manually-constructed lists which capture some aspects of style, for instance resources related to psycholinguistics (Coltheart, 1980), but these are necessarily limited in scope. Our interest is in providing broad lexical coverage, potentially in any language. Here, we will show that style is particularly amenable to corpus-based automated lexical acquisition.

Our approach to this problem is grounded in methods popularized for polarity lexicon creation (Turney and Littman, 2003), but we take a more holistic view than is typical, simultaneously tackling the acquisition of several styles in a single model. Not only is this theoretically warranted, due to the correlation effects resulting from the oral/literate spectrum of register, but we also show it can offer practical gains: our hybrid models first derive initial estimates of each style from a large social media corpus, and then refine these estimates based partially on the results from other styles. We demonstrate that various popular methods are applicable to this problem, and indeed a single method might not provide the best results for all styles. For evaluation, we use a consensus annotation, the results of which also raise interesting questions about annotation for more continuous kinds of variation.

2 Related Work

In English manuals of style and other prescriptivist texts (Strunk and White, 1979; Kane, 1983), writers are urged to pay attention to various aspects of lexical style, including elements such as familiarity, readability, formality, fanciness, colloquialness, specificity, concreteness, and objectivity; these stylistic categories reflect common aesthetic judgments about language, but are also inextricably linked to the conventions of register and genre. See Biber and Conrad (2009) for a discussion of the relationship between register, genre,

International Joint Conference on Natural Language Processing, pages 82–90,
Nagoya, Japan, 14-18 October 2013.

and style as traditionally defined in descriptive linguistics. Some researchers have posited a few fixed styles (Joos, 1961) or a small, discrete set of situational constraints which determine style and register (Halliday and Hasan, 1976); by contrast, the applied approach of Biber (1988) and theoretical framework of Leckie-Tarry (1995) offer a more continuous interpretation of register variation. In Biber's approach, functional dimensions such as *Involved vs. Informational*, *Argumentative vs. Non-argumentative*, and *Abstract vs. Non-abstract* are derived in an unsupervised manner from a mixed-genre corpus, with the labels assigned depending on where features (a small set of known indicators of register) and genres fall on each spectrum. The theory of Leckie-Tarry posits a single main cline of register with one pole (the oral pole) reflecting a reliance on the context of the linguistic situation, and the other (the literate pole) reflecting a reliance on cultural knowledge. The more specific elements of register are represented as subclines which are strongly influenced by this main cline, creating probabilistic relationships between related dimensions.

Computational linguistics research most similar to ours has focused on classifying the lexicon in terms of individual aspects relevant to style (e.g. formality, specificity, readability, and concreteness) (Brooke et al., 2010; Pan and Hsieh, 2010; Kidwell et al., 2009; Turney et al., 2011). Of particular methodological relevance is work on the induction of polarity lexicons based on co-occurrence in large corpora (Turney and Littman, 2003; Velikovich et al., 2010), or connections in WordNet (Rao and Ravichandra, 2009; Baccianella et al., 2010); semi-supervised vector space and graph methods are common, and several of the methods we apply here are taken directly from or inspired by work in this area.

3 Word annotation

In this study, we consider six styles—colloquial, literary, concrete, abstract, subjective, and objective—which are clearly represented in the lexicon, which are mentioned often in the relevant English linguistics literature, and which have strong positive and negative correlations with other styles in the group. Many (but not all) of these correlations are related to the oral/literate distinction. Our definition of each style (adapted from our annotation guidelines) is given below.

Colloquial Words which are used primarily in very informal contexts, for instance slang words and internet abbreviations.

Literary Words which you would expect to see primarily in literature; these words often feel old-fashioned or flowery.

Concrete Words which refer to events, objects, or properties of objects in the physical world that you would be able to see, hear, smell, or touch.

Abstract Words which refer to something that requires major psychological or cultural knowledge to grasp; complex ideas which can't purely be defined in physical terms.

Subjective Words which are strongly emotional or reflect a personal opinion.

Objective Words which are emotionally distant, explicitly avoiding any personal opinion, instead projecting a sense of disinterested authority.

Our method and evaluation relies on having a set of seed words for each style. The words used in this study were originally collected from various sources by the authors; we included words that we considered clear members of a particular stylistic category—though they might also belong to other categories—with little or no ambiguity with respect to that style. Colloquial seeds consist of English slang terms and acronyms, e.g. *cuz*, *gig*, *asshole*, *lol*. The literary seeds were primarily drawn from web sites which explain difficult language in texts such as the Bible and *Lord of the Rings*; examples include *behold*, *resplendent*, *amiss*, and *thine*. The concrete seeds all denote physical objects and actions, e.g. *shove* and *lamppost*, while the abstract seeds all involve nontrivial concepts *patriotism* and *nonchalant*. For our subjective seeds, we used an edited list of strongly positive and negative terms from a manually-constructed sentiment lexicon (Taboada et al., 2011), e.g. *gorgeous* and *depraved*, and for our objective set we selected words from sets of near-synonyms where one was clearly an emotionally-distant, formal alternative, e.g. *residence* (for *home*) or *occupied* (for *busy*). We filtered initial lists to 150 of each type (900 in total), removing words which did not appear in the corpus or which occurred in multiple lists.

Relying on a single annotator, however, is problematic, and a more serious issue with our original

Table 1: Fleiss's kappa for 5-way annotation, by style.

Style	Kappa
Literary	0.61
Abstract	0.37
Objective	0.55
Colloquial	0.85
Concrete	0.67
Subjective	0.63
Average	0.61

seed sets is that many of the seeds belong on multiple lists, reflecting the fact that stylistic correlations occur at the lexical level. This interferes with evaluation, since we need to to be fairly certain not only which seeds are in a category, but which are not. Therefore, we carried out a full annotation study with 5 annotators, asking each annotator to tag all 900 words for each of the 6 styles according to guidelines we prepared. One of the authors was included as an annotator (this annotation was carried out prior to all the others), but the other four were unfamiliar with the project; all were native English speakers with at least an undergraduate degree, and all reported reading a variety of text genres for work and/or pleasure. We provided written guidelines explaining each style in detail, and asked annotators to make judgments based on what they felt to be the most common sense. Communication among annotators was restricted during the process, but we allowed access to other resources (e.g. the internet) and answered general questions about the guidelines that came up during the process. A few annotators had obviously skewed numbers for certain styles relative to other annotators due to misinterpretation of the guidelines, and we provided non-specific feedback for revision in these cases. The Fleiss's kappa (Fleiss, 1971) values for our 5-way annotation study are presented in Table 1.[1]

The kappa values in Table 1 indicate agreement well above chance, but several of the dimensions (and the average) are below the 0.67 standard for reliable annotation (Artstein and Poesio, 2008), and only one (colloquial) reaches the higher 0.8 standard. This suggests that there is a sizable subjective aspect to these judgments and we should be somewhat skeptical of the judgment of any particular annotator. However, we had forced our annotators to make a boolean choice for each style, which may be somewhat inappropriate for somewhat non-discrete phenomenon like style. Taboada et al. (2011), when validating their fine-grained manual polarity lexicon (which included annotation of both polarity and strength), demonstrated that Mechanical Turk worker disagreement on a boolean task seemed to correspond fairly well to ranges on a scale: there was agreement at the extremes of polarity, but increasing disagreement towards the middle.

With this in mind, we used our initial annotations to create a new annotation task for two of our external annotators: the goal was to investigate whether annotators can identify relative differences in degree suggested by either agreement or disagreement with their choices by other annotators. First, we extracted minority opinions, defined here as word/style combinations where the annotator agreed with exactly one other annotator and disagreed with the three others, and consensus opinions, defined as those where all the annotators agreed. We randomly paired each minority opinion word/style with a consensus opinion; for both opinions, the annotator in question had made the same judgment (both yes, or both no), but some of the other annotators had made different choices. We then asked our annotators (who were unaware of the exact nature of the experiment) to pick, among two words they had tagged the same in the first round, the word which had 'more' of the relevant stylistic quality.

In the negative case (where the annotator had originally marked both as not having the style), the results are stark: in 97% of the cases, the annotator picked the minority opinion (i.e. the word which some other annotators had marked yes), suggesting that the annotator could identify the stylistic tendencies of the (mixed-agreement) word, but had nonetheless excluded it, probably because there were much clearer examples of this style and other styles which could be more clearly applied to the word. In the positive case, the annotators preferred the word with group consensus 82.7% of the time, which is indeed the pattern we would predict if the minority opinion is less extreme; the positive case is more subtle than the negative case, where many of the words used for comparison very clearly do not belong to the relevant style. These results are consistent with the

[1] The annotations and our guidelines are available at http://cs.toronto.edu/~jbrooke/style_annotations.zip .

Table 2: Number of seeds, by style.

Style	Positive	Negative
Literary	132	660
Abstract	107	599
Objective	245	495
Colloquial	163	684
Concrete	190	572
Subjective	258	487

idea that disagreement is a rough indicator of degree, and that not all disagreement should be dismissed as noise or some other failure of annotation. Of course, this also indicates that relative or continous (e.g. Likert scale) judgments might be preferable to boolean ones, but in this case boolean annotation is far more practical, and indeed desirable for both model creation and evaluation.

For our final seed set, our positive annotations include all word/style combinations where a majority of annotators marked yes, whereas our negative annotations include only terms where there was complete consensus; words where only 1 or 2 annotators marked yes were removed from consideration as seeds (for that particular style). The summary of the counts for main seed set are presented in Table 2.

4 Methods

Our method for stylistic lexicon acquisition breaks down into three steps. The first is to apply one of several methods which leverages co-occurrence in a large corpus to derive, for each word, a raw score for each style. We then take that raw score and normalize it; the resulting number can be used directly to compare words relevant to a style. Finally, we consider the vector formed by these normalized style scores, and apply other methods which further refine this vector, implicitly taking into account the correlations among styles. The elements of the refined vector correspond to the degree of each style, so if we apply this method for all words in our vocabulary we create a full-coverage lexicon.

4.1 Corpus analysis

For all the methods in this section, we use the same corpus, the ICWSM Spinn3r 2009 dataset (Burton et al., 2009), which has been used successfully in earlier work (Brooke et al., 2010). Social media corpora are particularly appropriate for research on style, since they contain a variety of registers. Here, we include all 2.46 million texts in the Tier 1 portion which contained at least 100 word types. Hapax legomena were excluded, since they could not possibly offer any co-occurrence information, but otherwise we did not filter or lemmatize words: our full vocabulary is 1.95 million words.

Our simplest method uses pointwise mutual information (PMI) (Church and Hanks, 1990), a popular metric for measuring the association between words. Since standard PMI has a lower bound of $-\infty$ when the joint probability is 0 (a common occurrence since many of our words are relatively rare), we actually use a normalized version, NPMI, which has an upper bound of 1 and a lower bound of -1.

$$NPMI(x,y) = \left(\log \frac{p(x,y)}{p(x)p(y)}\right)\left(\frac{1}{\log p(x,y)}\right)$$

Following earlier work (Brooke et al., 2010), here and elsewhere we do not use the term frequency within a document (which is less relevant to style). Instead the probabilities are calculated using the number of documents where the word or words appear divided by the total number of documents. The raw score r_{ij} for style i of word w_j is simply the sum of its NPMI with the associated set of seeds S_i:

$$r_{ij} = \sum_{s \in S_i} NPMI(w_j, s)$$

Our second method, LSA, was applied to formality by Brooke et al. (2010) and concreteness by Turney et al. (2011). We begin by converting our corpus into a binary word-document matrix, and carry out latent semantic analysis (Landauer and Dumais, 1997), which includes a singular value decomposition of the matrix and dimensionality reduction to k dimensions. Assuming $\mathbf{v}_w$ denotes the resulting k-dimensional vector for word w, we calculate r_{ij} as:

$$r_{ij} = \sum_{s \in S_i} \cos(\theta(\mathbf{v}_{w_j}, \mathbf{v}_s))$$

Our third method, using latent Dirichlet allocation (Blei et al., 2003), is more novel for lexical acquisition, and we address the specifics of this method in more detail in other work (Brooke and Hirst, 2013). Briefly, LDA is a Bayesian topic model which assumes that texts are generated via a

distribution of topics for each text (θ), and a distribution of words for each topic (β); given a corpus, appropriate values for θ and β are derived using inference, in this case variational Bayes inference using the original implementation provided by Blei et al. (2003). Our method works by seeding each of six topics in an LDA model (corresponding to our six styles) by dividing the entire initial probability mass among the seeds and running two iterations of the model, which distributes some of the probability mass to co-occurring words. In our previous work, we found further iterations had no benefit and even slightly degraded the model. For the LDA method, r_{ij} corresponds directly to β_{ij} of the resulting model which is just the probability of topic (style) i generating w_j.

4.2 Normalization

The raw numbers derived from corpus analysis methods discussed above cannot be used directly as indicators of style: the frequencies of both the seeds and the words being predicted have significant effect on the relative and absolute magnitudes of each style for all our methods, and performance using just these numbers is near chance. However, in two steps we can normalize these numbers to a form where the magnitude does directly reflect degree of a style. Again, r_{ij} refers to the raw score for style i and word j from some corpus analysis method. First, we take steps to ensure that r_{ij} is nonnegative. For LDA this is unnecessary (since r_{ij} is based on a probability distribution), but for NPMI and LSA it is needed, since both involve summing over items which vary between -1 and 1. We can ensure that these are positive by adding a constant equal to the number of seeds. Next, we convert the result to a style 'distribution' for each word:

$$r'_{ij} = \frac{r_{ij} + |S_i|}{\sum_{k=1}^{6} r_{kj} + |S_k|}$$

The result is still not useful, since frequency (and count) of seeds clearly still has an effect. To focus on the differences between words, we subtract the means for each style and divide by the standard deviation

$$b_{ij} = \frac{r'_{ij} - \overline{r'_i}}{\sigma_{r'_i}}$$

to reach b_{ij}, the base for the 'style space' methods in the next subsection.

4.3 Style Vector Optimization

Given a vector that represents the styles for a given word, we wish to refine the vector to improve performance on relative judgments for individual styles. Here, we test two options: the first transforms the stylistic vectors into k-Nearest Neighbor (kNN) graphs, where we can apply label propagation. The second option treats the vector as a set of features for supervised linear regression, one for each style, using a specialized loss function. Both methods rely on having a style vector representation of not only our target words, but also our seed (training) words. For LSA and NPMI, we used leave-one-out crossvalidation to create these vectors; for LDA, however, it was impractical to do a full run of the model for each word, and so we used 10-fold crossvalidation instead.

A vector-space representation offers a number of obvious similarity functions for building a kNN graph: we test two here, inverse Euclidean distance (L2) and cosine similarity (cos). A more difficult problem is the choice of k (for kNN k): here, we estimate a good k from the training set. Since the training set and dimensionality of the data is (now) fairly small, we simply test on all possible intervals of 5, and choose the best (often near 50, though we saw values as low as 10 and as high as 90) using our pairwise evaluation (see Section 5.1). Since our label propagation method works independently for each style, we can choose a different k for each.

For label propagation, we use the simple one-step propagation function from Kang et al. (2006). Here, K is our similarity function (which returns zero if seed s is not one of the k nearest neighbors), and z_{ij} is the resulting confidence score, which we use as our new estimate for the style:

$$z_{ij} = \sum_{w_s \in S_i} K(w_j, w_s)$$

Obviously, the main work here is done by the similarity function, which implicitly includes information from other stylistic dimensions by preferring words which are close not just on the relevant dimension, but in the stylistic space as a whole. There are of course more sophisticated, multi-step approaches to label propagation, e.g. the one used by Rao and Ravichandran (2009), but a single-step approach has clear advantages in light of our large vocabulary and dense graph; we leave exploration of whether unlabeled words can help further to fu-

ture work. We did test the one-step correlated label propagation method proposed by Kang et al. but found it was ineffective, probably because it increases the effects of correlation, which is actually counter to our needs.

The information provided by label propagation is distinct enough that it can be successfully combined with the original (base) vector. As with k for kNN, we estimated a good weighting for this combination using the training data, testing at 0.01 intervals. Since we noted some interdependence, we combined this step with the selection of (kNN) k. Again, this ratio can be different for each style.

Our second vector optimization technique is an adaption of supervised linear regression. Linear regression usually involves minimizing squared distance of the output of the model from the training set, assuming there are known values of expected output. In this case, however, we don't have reliable values for specific degrees of a style. We proceed by replacing the least-squared loss function with a loss function based on our evaluation metric (see Section 5.1):

$$L(\theta) = \sum_{w_j \in S_{i,p}} \sum_{w_m \in S_{i,n}} I(h_\theta(b_{ij}) < h_\theta(b_{im}))$$

Here, $S_{i,p}$ and $S_{i,n}$ refer to the positive and negative examples of style i, respectively, h_θ is the linear regression function, and I is an indicator function equal to 1 if the statement is true, and 0 otherwise.

Using such a loss function discourages standard approaches to linear regression, but in this context (a small feature space and training set), it is reasonably practical to search the space exhaustively for weights which provide a (near-)optimal result (on the training data).[2] Starting with full weight (1) on the feature corresponding to the dimension being derived and 0 on all others, we search the range -1 to 1 at 0.001 intervals for the other dimensions, proceeding in order based on the greatest difference across positive and negative examples of each style. We found that one such iteration across each element of the vector was sufficient, resulting in a stable model. This method can be applied on the initial vector, or on a vector that has already been refined by some other method, i.e. the output of label propagation.

[2] At the suggestion of a reviewer, we also tried applying SVMrank to this regression; it was much faster but performance was worse.

5 Evaluation

5.1 Setup

Our evaluation is based on the pairwise comparison of words which are known (from our annotation) to differ relevant to a certain style. Accuracy for a test set S_i (of a style i) is defined as the number of instances where the expected inequality exists between a pair of opposing words, divided by the total number of such pairings:

$$Accuracy(S_i) = \frac{\sum_{w_j \in S_{i,p}} \sum_{w_m \in S_{i,n}} I(z_{ij} > z_{im})}{|S_{i,p}| \cdot |S_{i,n}|}$$

Here z can refer to any of the metrics for style discussed in the previous section. The major advantage of this definition of accuracy is that it does not require an arbitrary cutoff point, but 100% accuracy nonetheless indicates that the two sets are perfectly separable. Also, it does not assume anything about the degree of difference between two words, e.g. that more is better, since for any given pair of words we cannot be certain what an ideal difference would be.

We evaluate using 3-fold crossvalidation, using the original 150-per-style annotation of our 900 words for the purposes of stratifying the data, which allows for balanced sets of 600 for training and 300 for testing. All seeding, training, and evaluation use the majority annotation of the 5 annotators, discussed in Section 3. Since the initial splits add a significant random factor, all results here are averaged over 5 runs, with the same 5 runs (i.e. same splits) used for all evaluated conditions.

5.2 Comparison of models

Table 3 shows a comparison of the performance of various models, organized by the method of corpus analysis. First, we note that most of these numbers are quite high, almost all are above 80% and most are above 90%. It is worth mentioning that if only direct opposites are considered (e.g. colloquial versus literary, concrete versus abstract), most dimensions reach results above 99%; our multi-style evaluation here offers a more realistic view. Among individual styles, colloquial words seem the most distinct, which is consistent with the results of human annotation. Acquisition of subjectivity, on the other hand, is strikingly more difficult than the other styles.

Based only on average accuracy, we could conclude that LSA $>$ LDA $>$ NPMI with respect

Model	By Style						Average
	Lit.	Abs.	Obj.	Coll.	Conc.	Subj.	
guessing baseline	50.0	50.0	50.0	50.0	50.0	50.0	50.0
NPMI							
base (Normalized)	68.4	91.2	94.4	95.6	73.4	77.1	83.0
LP-cos	90.1	91.5	95.1	94.4	90.0	80.0	90.2
LP-L2	88.2	88.9	94.1	94.1	89.4	76.6	88.5
base+LP-cos	90.2	92.8	95.6	96.0	90.6	80.9	91.0
base, LR	89.8	93.6	94.2	96.5	85.5	79.7	89.9
base+LP-cos, LR	90.2	93.6	95.5	95.9	90.5	81.0	91.1
LDA							
base	67.3	93.3	96.5	96.2	93.2	83.5	88.3
LP-cos	86.0	92.9	96.0	93.6	94.8	86.5	91.6
LP-L2	78.1	91.1	95.0	92.5	94.2	83.2	89.0
base+LP-cos	86.4	93.5	96.6	96.3	95.5	86.7	92.5
base, LR	84.3	93.9	96.5	96.4	94.7	85.7	91.8
base+LP-cos, LR	87.2	93.9	96.5	96.3	**95.8**	**87.0**	92.8
LSA							
k=20, base	89.1	93.5	95.6	94.4	90.8	76.0	89.9
k=500, base	91.2	93.7	96.5	96.5	93.7	83.5	92.6
k=500, LP-cos	92.4	91.7	96.0	96.8	94.3	85.2	92.8
k=500, LP-L2	92.1	92.1	96.5	96.5	94.3	85.0	92.8
k=500, base+LP-cos	92.5	93.6	96.8	97.5	94.8	85.9	93.5
k=500, base, LR	**92.7**	**94.0**	**97.2**	97.2	94.9	86.5	**93.7**
k=500, base+LP-cos, LR	**92.7**	93.8	97.0	**97.7**	94.9	86.4	**93.7**

Table 3: Model performance in lexical induction of seeds, % pairwise accuracy. LP = label propagation, cos = cosine similarity, L2 = inverse Euclidean distance, LR = linear regression. Bold is best in column.

to extracting relevant stylistic information from the corpus. That NPMI is the worst performing method is not surprising, since it relies only on direct co-occurrence between seeds and test words, and is not able to take advantage of larger patterns in the data; we would expect similar results for other simple relatedness measures. Though LSA is better overall, the distinction between LSA and LDA is more subtle, since in fact LDA is the higher performing model for two of the six styles, and its poorer overall performance can be attributed to a rather dismal showing for literary words, worse than NPMI. This is interesting because subjective and concrete words, where LDA does well, are the most common in the corpus, whereas literary words are consistently the least common. We posit, based on this and our earlier research focused on the LDA method, that successful low-dimensional seeded LDA requires styles (topics) that are reasonably well-represented in the corpus; when that condition is met, LDA will likely do better than LSA because it will distinguish rather than collapse correlated styles. LSA, on the other hand, is robust against the scarcity problem because it requires only that a set of words have a reasonably distinct k-dimensional profile to form a coherent style.

Based on the results in Table 3, we can conclude decisively that both of our optimization techniques are effective. The effects are particularly marked for NPMI, but is reasonably consistent across all three corpus analysis techniques and the various individual styles. With regards to the similarity function in label propagation, we found that cosine similarity, a less common choice for building graphs, was generally as good as, and often better than, Euclidean distance. The vector resulting from label propagation also consistently benefited from being combined with the base vector, the result being better than either alone. It is not entirely clear which of the two optimization methods is to be preferred (their effects seem roughly similar), though linear regression seems to have edge when using LSA. Combining the two methods seems a

good strategy, particularly for LDA.

The LSA results presented here mostly use $k = 500$, a fairly standard choice. However, we tested other values, in particular extremely low values ($k = 20$) to see if we could confirm our supposition (Brooke et al., 2010) that much stylistic information is contained with the first few dimensions of LSA. Our results suggest that the basic supposition is valid, since the difference between the two conditions for most dimensions is not large, but the identification of subjectivity (not considered by Brooke et al. 2010) does seem to benefit greatly from a higher-dimensional vector.

6 Qualitative analysis

To investigate further the successes and failures of our method, we carried out two qualitative examinations of the output of our model. First, we looked at those words within our annotated set of words which consistently caused the most errors across the various splits and runs. Second, we ran a high-performing LSA model built from the entire seed set on a subset of our vocabulary (we excluded words of document frequency less than 100), creating lexicons for each style; we manually inspected non-seed words that were ranked highest on each dimension.

The clearest result from the inspection of the seed output was that many of the false negatives involve words that are strong on some other dimension, typically on the other side of the oral/literate divide. For example, the most difficult-to-identify literary and abstract terms are strongly subjective (e.g. *loathe* and *obscene*), while the most difficult objective word, *translucent*, is very concrete. The most difficult concrete words are literary (*yoke*, *raiment*) or objective (*conflagration*), and the most difficult subjective words are also somewhat objective (*eminent*) or abstract (*autocratic*). Interestingly, a manual inspection of the weights for linear regression suggests that our optimization is correcting for just this kind of situation: we generally see negative weights on (what we would predict to be) positively correlated styles, and vice versa. However, in certain cases where one style has a much larger role in determining the co-occurrence pattern in the corpus, this correction may be insufficient.

Most of the false positives, by contrast, involve overextension of each category in predictable ways. For example, our highest ranking literary words from the general vocabulary were mostly very good, but contained a few words that are obvious over generalizations into biblical and fantasy texts, e.g. *locust* and *sorcerers*, while among the objective words there were a number of academia-relevant words that are really more abstract than objective, e.g. *coauthors* and *peer-review*. Our derived colloquial words contained many (sometimes purposeful) misspellings (*wayy*, *annnnd*) which we could argue are genuinely colloquial; less clear are the many lower-case celebrity names (e.g. *miley*), but the fact that the bloggers used lower case does make them non-standard. Consistent with our qualitative results, subjective was the most problematic in the general vocabulary: though there were many good subjective words, there were a lot of other words which suggest topics that people tend to express opinions about, e.g. *sitcoms*, *entertainer*, or *flick*; movie-related words are particularly common, which might be a reflection the lexicon we took our subjective seeds from.

7 Conclusion

We have presented a methodology for deriving high-quality stylistic lexicons from corpora. A key aspect of our approach its hybrid nature: information is first extracted (using efficient, well-established methods) in a semi-supervised fashion from large corpora, and then refined using fully-supervised techniques. We argue that there are clear benefits in looking at multiple styles simultaneously, not only in terms of improving performance but also in taking our evaluation beyond 'toy' situations where we ignore the complexities and interactions among styles, drawing connections with broader insights from linguistics.

One possible criticism of our method is that we use only co-occurrence information, and not other information (e.g. word morphology) which could be relevant to particular styles in English; this option should be explored further, particularly in the optimization phase where we can easily add other features, though we stress that our ultimate goal is to derive methods that are easily extensible to more styles and more languages. We have also not considered word senses or multiword expressions, but both can and should be added to the model.

Acknowledgements

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

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC'10*, Valletta, Malta.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.

Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Julian Brooke and Graeme Hirst. 2013. A multi-dimensional Bayesian approach to lexical style. In *Proceedings of NAACL '13*, Atlanta.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of COLING '10*, Beijing.

Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of ICWSM '09*, San Jose.

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

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science Technology*, 56(13):1448–1462.

Max Coltheart. 1980. *MRC Psycholinguistic Database User Manual: Version 1*. Birkbeck College.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

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

Martin Joos. 1961. *The Five Clocks*. Harcourt, Brace and World, New York.

Thomas S. Kane. 1983. *The Oxford Guide to Writing*. Oxford Univeristy Press.

Feng Kang, Rong Jin, and Rahul Sukthankar. 2006. Correlated label propagation with application to multi-label learning. In *Proceedings of CVPR '06*, New York.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of ACL '97*, Madrid.

Paul Kidwell, Guy Lebanon, and Kevyn Collins-Thompson. 2009. Statistical estimation of word acquisition with application to readability prediction. In *Proceedings of EMNLP'09*, Singapore.

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 and Context: A Functional Linguistic Theory of Register*. Pinter.

Ching-Fen Pan and Shu-Kai Hsieh. 2010. Word space modeling for measuring semantic specificity in Chinese. In *Proceedings of COLING '10*, Beijing.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the Enron corpus. In *Proceedings of ACL '11*, Portland.

Delip Rao and Deepak Ravichandra. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of EACL '09*, Athens.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of ACL '11*, Portland.

William Strunk and E.B. White. 1979. *The Elements of Style*. Macmillan, 3rd edition.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

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

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP '11*, Edinburgh, United Kingdom.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Proceedings of NAACL '10*, Los Angeles.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP '05*, Vancouver.