A Multi-Dimensional Bayesian Approach to Lexical Style

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

We adapt the popular LDA topic model (Blei et al., 2003) to the representation of stylistic lexical information, evaluating our model on the basis of human-interpretability at the word and text level. We show, in particular, that this model can be applied to the task of inducing stylistic lexicons, and that a multi-dimensional approach is warranted given the correlations among stylistic dimensions.

1 Introduction

In language, stylistic variation is a reflection of various contextual factors, including the backgrounds of and relationship between the parties involved. Although in the context of prescriptive linguistics (Strunk and White, 1979), style is often assumed to be a matter of aesthetics, the stylistic intuitions of language users are inextricably linked to the conventions of register and genre (Biber and Conrad, 2009). Intentional or not, stylistic differences play a role in numerous NLP tasks. Examples include genre classification (Kessler et al., 1997), author profiling (Garera and Yarowsky, 2009; 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; Inkpen and Hirst, 2006). 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. When content words are considered, they are often limited to manually-constructed lists (Argamon et al., 2007), or used as individual features for supervised classification, which can be confounded by topic (Petrenz and Webber, 2011) or fail in the face of lexical variety. Our interest is models that offer broad lexical coverage of human-identifiable stylistic variation.

Research most similar to ours has focused on classifying the lexicon in terms of individual aspects relevant to style (e.g. formality, specificity, readability) (Brooke et al., 2010; Pan and Yang, 2010; Kidwell et al., 2009) and a large body of research on the induction of polarity lexicons, in particular from large corpora (Turney, 2002; Kaji and Kitsuregawa, 2007; Velikovich et al., 2010). Our work is the first to represent multiple dimensions of style in a single statistical model, adapting latent Dirichlet allocation (Blei et al., 2003), a Bayesian 'topic' model, to our stylistic purposes; as such, our approach also follows on recent interest in the interpretability of topic-model topics (Chang et al., 2009; Newman et al., 2011). We show that our model can be used for acquisition of stylistic lexicons, and we also evaluate the model relative to theories of register variation and the expected stylistic character of particular genres.

2 Model

2.1 Linguistic foundations

In English manuals of style and other prescriptivist texts (Fowler and Fowler, 1906; Gunning, 1952; Follett, 1966; Strunk and White, 1979; Kane, 1983; Hayakawa, 1994), writers are urged to pay attention to various aspects of lexical style, including elements such as clarity, familiarity, readability, for-

Proceedings of NAACL-HLT 2013, pages 673–679,
Atlanta, Georgia, 9–14 June 2013.

mality, fanciness, colloquialness, specificity, concreteness, objectivity, and naturalness; these stylistic categories reflect common aesthetic judgments about language. In descriptive studies of register, some researchers have posited a few fixed styles (Joos, 1961) or a small, discrete set of situational constraints which determine style and register (Crystal and Davy, 1969; 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 full 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 (Birch, 1995).

For the present study, we have chosen 3 dimensions (6 styles) which are clearly represented in the lexicon, which are discussed often in the relevant literature, and which fit well into the Leckie-Tarry conception of related subclines: colloquial vs. literary, concrete vs. abstract, and subjective vs. objective. In addition to a negative correlation between opposing styles, we also expect a positive correlation between stylistic aspects that tend toward the same main pole, situational (i.e. colloquial, concrete, subjective) or cultural (i.e. literary, abstract, objective). These correlations can potentially interfere with accurate lexical acquisition.

2.2 Implementation

Our main model is an adaption of the popular latent Dirichlet allocation topic model (Blei et al., 2003), with each of the 6 styles corresponding to a topic. Briefly, latent Dirichlet allocation (LDA) is a generative Bayesian model: for each document d, a distribution of topics θ_d is drawn from a Dirichlet prior (with parameter α). For each topic z, there is a probability distribution β_z[1] corresponding to the probability of that topic generating any given word in the vocabulary. Words in document d are generated by first selecting a topic z randomly according to θ_d, and then randomly selecting a word w according to β_z. An extension of LDA, the correlated topic model (CTM) (Blei and Lafferty, 2007), supposes a more complex representation of topics: given a matrix Σ representing the covariance between topics and μ representing the means, for each document a topic distribution η (analogous to θ) is drawn from the logistic normal distribution. Given a corpus, good estimates for the relevant parameters can be derived using Bayesian inference.

For both LDA and CTM we use the original variational Bayes implementation of Blei. Variational Bayes (VB) works by approximating the true posterior with a simpler distribution, minimizing the Kullback-Leibler divergence between the two through iterative updates of specially-introduced free variables. The mathematical and algorithmic details are omitted here; see Blei et al. (2003; 2007). Our early investigations used an online, batch version of LDA (Hoffman et al., 2010), which is more appropriate for large corpora because it requires only a single iteration over the dataset. We discovered, however, that batch models were markedly inferior to more traditional models for our purposes because the influence of the initial model diminishes too quickly; here, we need particular topics in the model to correspond to particular styles, and we accomplish this by seeding the model with known instances of each style (see Section 3). Specifically, our initial β consists of distributions where the entire probability mass is divided amongst the seeds for each corresponding topic, and a full iteration over the corpus occurs before β is updated. Typically, LDA iterates over the corpus until a convergence requirement is met, but in this case this is neither practical (due to the size of our corpus) nor necessarily desirable; the diminishing effects of the initial seeding means that the model may not stabilize, in terms of its likelihood, until after it has shifted away from our desired stylistic dimensions towards some other

[1] Some versions of LDA smooth this distribution using a Dirichlet prior; here, though, we use the original formulation from Blei (2003), which does not.

variation in the data. Therefore, we treat the optimal number of iterations as a variable to investigate.

The model is trained on a 1 million text portion of the 2009 version of the ICWSM Spinn3r dataset (Burton et al., 2009), a corpus of blogs we have previously used for formality lexicon induction (Brooke et al., 2010). Since our method relies on co-occurrence, we followed our earlier work in using only texts with at least 100 different word types. All words were tokenized and converted to lower-case, with no further lemmatization. Following Hoffman et al. (2010), we initialized the α of our models to $1/k$ where k is the number of topics. Otherwise we used the default settings; when they overlap they were identical for the LDA and CTM models.

3 Lexicon Induction

Our primary evaluation is based on the stylistic induction of held-out seed words. The words were collected from various sources by the first author and further reviewed by the second; we are both native speakers of English with significant experience in English linguistics. Included words had to be clear, extreme members of their stylistic category, with little or no ambiguity with respect to their style. The 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 objects and actions strongly rooted in the physical world, e.g. *shove* and *lamppost*, while the abstract seeds all involve concepts which require significant human psychological or cultural knowledge to grasp, for instance *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 alternative, e.g. *residence* (for *home*), *jocular* (for *funny*) and *communicable* (for *contagious*). We filtered initial lists to 150 of each type, removing words which did not appear in the corpus or which occurred in multiple lists. For evaluation we used stratified 3-fold crossvalidation, averaged over 5 different (3-way) splits of the seeds, with the same splits used for all evaluated conditions.

Given two sets of opposing seeds, we follow our earlier work in evaluating our performance in terms of the number of pairings of seeds from each set which have the expected stylistic relationship relative to each other (the guessing baseline is 0.5). Given a word w and two opposing styles (topics) p and n, we place w on the *PN* dimension according to the β of our trained model as follows:

$$PN_w = \frac{\beta_{pw} - \beta_{nw}}{\beta_{pw} + \beta_{nw}}$$

The normalization is important because otherwise more-common words would tend to have higher *PN*'s, when in fact the opposite is true (rare words tend to be more stylistically prominent). We then calculate pairwise accuracy as the percentage of pairs $\langle w_p, w_n \rangle$ ($w_p \in P_{seeds}$ and $w_n \in N_{seeds}$) where $PN_{w_p} > PN_{w_n}$. However, this metric does not address the case where the degree of a word in one stylistic dimension is overestimated because of its status on a parallel dimension. Two more-holistic alternatives are total accuracy, the percentage of seeds for which the highest β_{tw} is the topic t for which w is a seed (guessing baseline is 0.17), and the average rank of the correct t as ordered by β_{tw} (in the range 1–6, guessing baseline is 3.5); the latter is more forgiving of near misses.

We tested a few options which involved straightforward modifications to model training. Standard LDA produces all tokens in the document, but when dealing with style rather than topic, the number of times a word appears is much less relevant (Brooke et al., 2010). Our binary model assumes an LDA that generates types, not tokens.[2] A key comparison

[2] At the theoretical level, this move is admittedly problematic, since our LDA model is thus being trained under the assumption that texts with multiple instances of the same type can be generated, when of course such texts cannot by definition exist. We might address this by moving to Bayesian models with very different generative assumptions, e.g. the spherical topic model (Reisinger et al., 2010), but these methods involve a significant increase of computational complexity and we believe that on a practical level there are no real negatives associated with directly using a binary representation as input to LDA; in fact, we are avoiding what appears to be a much more serious problem, burstiness (Doyle and Elkan, 2009), i.e. the fact that

Model	Pairwise Accuracy (%)				Total Acc. (%)	Avg. Rank
	Lit/Col	Abs/Con	Obj/Sub	All		
guessing baseline	50.0	50.0	50.0	50.0	16.6.	3.50
basic LDA (iter 2)	94.3	98.8	93.0	95.4	55.0	1.79
binary LDA (iter 2)	**96.2**	98.9	**93.5**	**96.2**	**57.7**	**1.74**
combo binary LDA (iter 1)	95.4	**99.2**	93.3	96.0	53.1	1.86
binary CTM (iter 1)	96.3	99.0	89.6	95.0	53.0	1.87

Table 1: Model performance in lexical induction of seeds. Bold indicates best in column.

here is with a combined LDA model (*combo*), an amalgamation of three independently trained 2-topic models, one for each dimension; this tests our key hypothesis that training dimensions of style together is beneficial. Finally, we test against the correlated topic model (CTM), which offers an explicit representation of style correlation, but which has done poorly with respect to interpretability, despite offering better perplexity (Chang et al., 2009).

The results of the lexicon induction evaluation are in Table 1. Since the number of optimal iterations varies, we report the result from the best of the first five iterations, as measured by total accuracy; the best iteration is shown in parenthesis. In general, all the results are high enough—we are reliably above 90% for the pairwise task, and above 50% for the 6-way task—for us to conclude with some confidence that our model is capturing a significant amount of stylistic variation. As predicted, using words as boolean features had a net positive gain, consistent across all of our metrics, though this effect was not as marked as we have seen previously. The model with independent training of each dimension (*combo*) did noticeably worse, supporting our conclusion that a multidimensional approach is warranted here. Particularly striking is the much larger drop in overall accuracy as compared to pairwise accuracy, which suggests that the combo model is capturing the general trends but not distinguishing correlated styles as well. However, the most complex model, the CTM, actually does slightly worse than the combo, which was contrary to our expectations but nonetheless consistent with previous work on the interpretability of topic models. The performance of the full LDA models benefited from a second iteration, but this was not true of combo LDA or CTM, and the performance of all models dropped after the second iteration.

An analysis of individual errors reveals, unsurprisingly, that most of the errors occur across styles on the same pole; by far the largest single common misclassification is objective words to abstract. Of the words that consistently show this misclassification across the runs, many of them, e.g. *animate*, *aperture*, *encircle*, and *constrain* are clearly errors (if anything, these words tend towards concreteness), but in other cases the word in question is arguably also fairly abstract, e.g. *categorize* and *predominant*, and might not be labeled an error at all. Other signs that our model might be doing better than our total accuracy metric gives it credit for: many of the subjective words that are consistently mislabeled as literary have an exaggerated, literary feel, e.g. *jubilant*, *grievous*, and *malevolent*.

4 Text-level Analysis

Our secondary analysis involved evaluating the θ's of our best configuration (based on average pairwise and total accuracy) on other texts. After training, we carried out inference on the BNC corpus, averaging the resulting θ's to see which styles are associated with which genres. Appearances of the seed terms for each model were disregarded during this process; only the induced part of the lexicon was used. The average differences relative to the mean across the various stylistic dimensions (as measured by the probabilities in θ) are given for a selection of genres in Table 2.

The most obvious pattern in table 2 is the dominance of the medium: all written genres are positive for our styles on the 'cultural' pole and negative for styles on the 'situational' pole and the opposite is

traditional LDA is influenced too much by multiple instances of the same word.

Genre	Styles					
	Literary	**Abstract**	**Objective**	**Colloquial**	**Concrete**	**Subjective**
News	+0.67	+0.50	+0.43	−0.31	−0.72	−0.57
Religious texts	+0.38	+0.38	+0.28	−0.27	−0.44	−0.32
Academic	+0.18	+0.29	+0.26	−0.20	−0.36	−0.18
Fiction	+0.31	+0.09	+0.02	−0.05	−0.12	−0.25
Meeting	−0.61	−0.54	−0.42	+0.35	+0.69	+0.55
Courtroom	−0.63	−0.53	−0.41	+0.32	+0.69	+0.57
Conversation	−0.56	−0.63	−0.54	+0.43	+0.80	+0.50

Table 2: Average differences from corpus mean of LDA-derived stylistic dimension probabilities for various genres in the BNC, in hundredths.

true for spoken genres. The magnitude of this effect is more difficult to interpret: though it is clear why fiction should sit on the boundary (since it contains spoken dialogue), the appearance of news at the written extreme is odd, though it might be due to the fact that news blogs are the most prevalent formal genre in the training corpus.

However, if we ignore magnitude and focus on the relative ratios of the stylistic differences for styles on the same pole, we can identify some individual stylistic effects among genres within the same medium. Relative to the other written genres, for instance, fiction is, sensibly, more literary and much less objective, while academic texts are much more abstract and objective; for the other two written genres, the spread is more even, though relative to religious texts, news is more objective. At the situational pole, fiction also stands out, being much more colloquial and concrete than other written genres. Predictably, if we consider again the ratios across styles, conversation is the most colloquial genre here, though the difference is subtle.

We carried out a correlation analysis of the LDA-reduced styles of all texts in the BNC and, consistent with the genre results in Table 2, found a strong positive correlation for all styles on the same main pole, averaging 0.83. The average negative correlation between opposing poles is even higher, −0.88. This supports the Leckie-Tarry formulation. The independence assumptions of the LDA model did not prevent strong correlations from forming between these distinct yet clearly interrelated dimensions; if anything, the correlations are stronger than we would have predicted.

5 Conclusion

We have introduced a Bayesian model of stylistic variation. Topic models like LDA are often evaluated using information-theoretic measures, but our emphasis has been on interpretibility: at the word level we can use the model to induce stylistic lexicons which correspond to human judgement, and at the text level we can use it distinguish genres in expected ways. Another theme has been to offer evidence that indeed a multi-dimensional approach is strongly warranted: importantly, our results indicate that separate unidimensional models of style are inferior for identifying the core stylistic character of each word, and in our secondary analysis we found strong correlations among styles attributable to the situational/cultural dichotomy. However, an off-the-shelf model that integrates correlation among topics did not outperform basic LDA.

One advantage of a Bayesian approach is in the flexibility of the model: there are any number of other interesting possible extensions at both the θ and β levels of the model, including alternative approaches to correlation (Li and McCallum, 2006). Beyond Bayesian models, vector space and graphical approaches should be compared. More work is clearly needed to improve evaluation: some of our seeds could fall into multiple stylistic categories, so a more detailed annotation would be useful.

Acknowledgements

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

References

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 7:91–109.

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 Birch. 1995. Introduction. In Helen Leckie-Tarry, editor, *Language and Context: A Functional Linguistic Theory of Register*. Pinter.

David M. Blei and John D. Lafferty. 2007. Correlated topic models. *Annals of Applied Statistics*, 1(1):17–35.

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, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, Beijing.

Kevin Burton, 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.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of Neural Information Processing Systems (NIPS '09)*.

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.

David Crystal and Derek Davy. 1969. *Investigating English Style*. Indiana University Press.

Gabriel Doyle and Charles Elkan. 2009. Accounting for burstiness in topic models. In *International Conference on Machine Learning (ICML '09)*.

Wilson Follett. 1966. *Modern American Usage*. Hill & Wang, New York.

H. W. Fowler and F. G. Fowler. 1906. *The King's English*. Clarendon Press, Oxford, 2nd edition.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP '09)*, pages 710–718, Singapore.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill, New York.

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

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

Matthew D. Hoffman, David M. Blei, and Francis R. Bach. 2010. Online learning for latent Dirichlet allocation. In *Neural Information Processing Systems (NIPS '10)*, pages 856–864.

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.

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

Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*.

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

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 (ACL '97)*, pages 32–38, Madrid, Spain.

Paul Kidwell, Guy Lebanon, and Kevyn Collins-Thompson. 2009. Statistical estimation of word acquisition with application to readability prediction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 900–909, Singapore.

Helen Leckie-Tarry. 1995. *Language and Context: A Functional Linguistic Theory of Register*. Pinter.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 577–584.

David Newman, Edwin V. Bonilla, and Wray Buntine. 2011. Improving topic coherence with regularized topic models. In *Proceedings of Advances in Neural Information Processing Systems (NIPS '11)*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10).

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on

the Enron corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. *Computational Linguistics*, 37(2):385–393, June.

J. Reisinger, A. Waters, B. Silverthorn, and R. Mooney. 2010. Spherical topic models. In *International Conference on Machine Learning (ICML '10)*.

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 the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.

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 D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Philadelphia, Pennsylvania.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 777–785, Los Angeles, California.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT/EMNLP '05, pages 347–354.