Unsupervised Text Segmentation Based on Native Language Characteristics

Shervin Malmasi[1,2] **Mark Dras**[2] **Mark Johnson**[2] **Lan Du**[3] **Magdalena Wolska**[4]
[1]Harvard Medical School, Harvard University
smalmasi@bwh.harvard.edu
[2]Department of Computing, Macquarie University
{ shervin.malmasi, mark.dras, mark.johnson }@mq.edu.au
[3]Faculty of IT, Monash University
lan.du@monash.edu
[4]LEAD Graduate School, Universität Tübingen
magdalena.wolska@uni-tuebingen.de

Abstract

Most work on segmenting text does so on the basis of topic changes, but it can be of interest to segment by other, stylistically expressed characteristics such as change of authorship or native language. We propose a Bayesian unsupervised text segmentation approach to the latter. While baseline models achieve essentially random segmentation on our task, indicating its difficulty, a Bayesian model that incorporates appropriately compact language models and alternating asymmetric priors can achieve scores on the standard metrics around halfway to perfect segmentation.

1 Introduction

Most work on automatically segmenting text has been on the basis of topic: segment boundaries correspond to topic changes (Hearst, 1997). There are various contexts, however, in which it is of interest to identify changes in other characteristics; for example, there has been work on identifying changes in authorship (Koppel et al., 2011) and poetic voice (Brooke et al., 2012). In this paper we investigate text segmentation on the basis of change in the native language of the writer.

Two illustrative contexts where this task might be of interest are patchwriting detection and literary analysis. Patchwriting is the heavy use of text from a different source with some modification and insertion of additional words and sentences to form a new text. Pecorari (2003) notes that this is a kind of textual plagiarism, but is a strategy for learning to write in an appropriate language and style, rather than for deception. Keck (2006), Gilmore et al. (2010) and Vieyra et al. (2013) found that non-native speakers, not surprisingly in situations of imperfect mastery of a language, are strongly over-represented in this kind of textual plagiarism. In these cases the boundaries between the writer's original text and (near-)copied native text are often quite apparent to the reader, as in this short example from Li and Casanave (2012) (copied text italicised): "Nevertheless, doubtfulness can be cleared reasonably by the experiments conducted upon the 'split-brain patients', *in whom intra-hemispheric communication is no longer possible*. To illustrate, *one experiment has the patient sit at a table with a* non-transparent *screen blocking the objects behind*, who is then asked to *reach* the objects with different hand respectively." Because patchwriting can indicate imperfect comprehension of the source (Jamieson and Howard, 2013), identifying it and supporting novice writers to improve it has become a focus of programmes like the Citation Project.[1]

For the second, perhaps more speculative context of literary analysis, consider Joseph Conrad, known for having written a number of famous English-language novels, such as *Heart of Darkness*; he was born in Poland and moved to England at the age of 21. His writings have been the subject of much manual analysis, with one particular direction of such research being the identification of likely influences on his English writing, including his native Polish language and the French he learnt before English. Morzinski (1994), for instance, notes aspects of his writing that exhibit Polish-like syntax, verb inflection, or other linguistic characteristics (*e.g.* "Several had still their staves in their hands" where the awkwardly placed adverb *still* is typical of Polish). These appear both in isolated sentences and in larger chunks of text, and part of

[1]http://citationproject.net/

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1457–1469
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1134

an analysis can involve identifying these chunks.

This raises the question: Can NLP and computational models identify the points in a text where native language changes? Treating this as an unsupervised text segmentation problem, we present the first Bayesian model of text segmentation based on authorial characteristics, applied to native language.

2 Related Work

Topic Segmentation The most widely-researched text segmentation task has as its goal to divide a text into topically coherent segments. *Lexical cohesion* (Halliday and Hasan, 1976) is an important concept here: the principle that text is not formed by a random set of words and sentences but rather logically ordered sets of related words that together form a topic. In addition to the semantic relation between words, other methods such as back-references and conjunctions also help achieve cohesion. Based on this, Morris and Hirst (1991) proposed the use of *lexical chains*, sequences of related words (defined via thesaurus), to break up a text into topical segments: breaks in lexical chains indicate breaks in topic. The *TextTiling* algorithm (Hearst, 1994, 1997) took a related approach, defining a function over lexical frequency and distribution information to determine topic boundaries, and assuming that each topic has its own vocabulary and that large shifts in this vocabulary usage correspond to topic shifts.

There have been many approaches since that time. A key one, which is the basis for our own work, is the unsupervised Bayesian technique BAYESSEG of Eisenstein and Barzilay (2008), based on a generative model that assumes that each segment has its own language model. Under this assumption the task can be framed as predicting boundaries at points which maximize the probability of a text being generated by a given language model. Their method is based on lexical cohesion — expressed in this context as topic segments having compact and consistent lexical distributions — and implements this within a probabilistic framework by modelling words within each segment as draws from a multinomial language model associated with that segment.

Much other subsequent work either uses this as a baseline, or extends it in some way: Jeong and Titov (2010), for example, who propose a model for joint discourse segmentation and alignment for documents with parallel structures, such as a text with commentaries or presenting alternative views on the same topic; or Du et al. (2013), who use hierarchical topic structure to improve the linear segmentation.

Bible Authorship Koppel et al. (2011) consider the task of decomposing a document into its authorial components based on their stylistic properties and propose an unsupervised method for doing so. The authors use as their data two biblical books, Jeremiah and Ezekiel, that are generally believed to be single-authored: their task was to segment a single artificial text constructed by interleaving chapters of the two books. Their most successful method used work in biblical scholarship on lexical choice: they give as an example the case that in Hebrew there are seven synonyms for the word *fear*, and that different authors may choose consistently from among them. Then, having constructed their own synsets using available biblical resources and annotations, they represent texts by vectors of synonyms and apply a modified cosine similarity measure to compare and cluster these vectors. While the general task is relevant to this paper, the particular notion of synonymy here means the approach is specific to this problem, although their approach is extended to other kinds of text in Akiva and Koppel (2013). Aldebei et al. (2015) proposed a new approach motivated by this work, similarly clustering sentences, then using a Naive Bayes classifier with modified prior probabilities to classify sentences.

Poetry Voice Detection Brooke et al. (2012) perform stylistic segmentation of a well-known poem, *The Waste Land* by T.S. Eliot. This poem is renowned for the great number of voices that appear throughout the text and has been the subject of much literary analysis (Bedient and Eliot, 1986; Cooper, 1987). These distinct voices, conceived of as representing different characters, have differing tones, lexis and grammatical styles (*e.g.* reflecting the level of formality). The transitions between the voices are not explicitly marked in the poem and the task here is to predict the breaks where these voice changes occur. The authors argue that the use of generative models is not feasible for this task, noting: "Generative models, which use a bag-of-words assumption, have a very different problem: in their standard form, they can capture only lexical cohesion, which is not the (primary) focus of stylistic analysis."

They instead present a method based on a curve that captures stylistic change, similar to the TextTiling approach but generalised to use a range of features. The local maxima in this change curve represent potential breaks in the text. The features are both internal to the poem (*e.g.* word length, syllable count, POS tag) as well as external (*e.g.* average unigram counts in the 1T Corpus or sentiment polarity from a lexicon). Results on an artificially constructed mixed-style poem achieve a P_k of 0.25. Brooke et al. (2013) extend this by considering clustering following an initial segmentation.

Native Language Identification (NLI) NLI casts the detecting of native language (L1) influence in writing in a non-native (L2) language as a classification task: the framing of the task in this way comes from Koppel et al. (2005). There has been much activity on it in the last few years, with Tetreault et al. (2012) providing a comprehensive analysis of features that had been used up until that point, and a shared task in 2013 (Tetreault et al., 2013) that attracted 29 entrants. The shared task introduced a new, now-standard dataset, TOEFL11, and work has continued on improving classification results, *e.g.* by Bykh and Meurers (2014) and Ionescu et al. (2014).

In addition to work on the classification task itself, there have also been investigations of the features used, and how they might be employed elsewhere. Malmasi and Cahill (2015) examine the effectiveness of individual feature types used in the shared task and the diversity of those features. Of relevance to the present paper, simple part-of-speech n-grams alone are fairly effective, with classification accuracies of between about 40% and 65%; higher-order n-grams are more effective than lower, and the more fine-grained CLAWS2 tagset more effective than the Penn Treebank tagset. An area for application of these features is in Second Language Acquisition (SLA), as a data-driven approach to finding L1-related characteristics that might be a result of cross-linguistic influence and consequently a possible starting for an SLA hypothesis (Ellis, 2008); Swanson and Charniak (2013) and Malmasi and Dras (2014) propose methods for this.

Tying It Together Contra Brooke et al. (2012), we show that it is possible to develop effective generative models for segmentation on stylistic factors, of the sort used for topic segmentation. To apply it specifically to segmentation based on a writer's L1, we draw on work in NLI.

3 Experimental Setup

We investigate the task of L1-based segmentation in three stages:

1. Can we define any models that do better than random, in a best case scenario? For this best case scenario, we determine results over a devset with the best prior found by a grid search, for a single language pair likely to be relatively easily distinguishable. Note that as this is *unsupervised* segmentation, it is a devset in the sense that it is used to find the best prior, and also in a sense that some models as described in §4 use information from a related NLI task on the underlying data.

2. If the above is true, do the results also hold for test data, using priors derived from the devset?

3. Further, do the results also hold for all language pairs available in our dataset, not just a single easily distinguishable pair?

We first describe the evaluation data — artificial texts generated from learner essays, similar to the artificially constructed texts of previously described work on Bible authorship and poetry segmentation — and evaluation metrics, followed in §4 by the definitions of our Bayesian models.

3.1 Source Data

We use as the source of data the TOEFL11 dataset used for the NLI shared task (Blanchard et al., 2013) noted in §2. The data consists of 12100 essays by writers with 11 different L1s, taken from TOEFL tests where the test-taker is given a *prompt*[2] as the topic for the essay. The corpus is balanced across L1s and prompts (which allows us to verify that segmentation isn't occurring on the basis of topic), and is split into standard training, dev and test sets.

3.2 Document Generation

As the main task is to segment texts by the author's L1, we want to ensure that we are not segmenting by topic and thus use texts written by authors from different L1 backgrounds on the same topic (prompt). We will also create one dataset to verify that segmentation by topic works in this domain; for this we use texts written by authors from the same L1 background on different topics.

For our L1-varying datasets, we construct composite documents to be segmented as alternat-

[2]For example, prompt P7 is: "Do you agree or disagree with the following statement? It is more important for students to understand ideas and concepts than it is for them to learn facts. Use reasons and examples to support your answer."

ing segments drawn from TOEFL11 from two different L1s holding the topic (prompt) constant, broadly following a standard approach (Brooke et al., 2012, for example) (see Appendix A.1 for details). We follow the same process for our topic-varying datasets, but hold the L1 constant while alternating the topic (prompt). For our single pair of L1s, we choose German and Italian. German is the class with the highest NLI accuracy in the TOEFL11 corpus across the shared task results and Italian also performs very well. Additionally, there is very little confusion between the two; a binary NLI classifier we trained on the language pair achieved 97% accuracy. For our all-pairs results, given the 11 languages in the TOEFL11 corpus, we have 55 sets of documents of alternating L1s (one of which is German–Italian).

We generate four distinct types of datasets for our experiments using the above methodology. The documents in these datasets, as described below, differ in the parameters used to select the essays for each segment and what type of tokens are used. Tokens (words) can be represented in their original form and used for performing segmentation. Alternatively, using an insight from Wong et al. (2012), we can represent the documents at a level other than lexical: the text could consist of the POS tags corresponding to all of the tokens, or n-grams over those POS tags. The POS representation is motivated by the usefulness of POS-based features for capturing L1-based stylistic differences as noted in §2. Our method for encoding n-grams is described in Appendix A.2.

TOPICSEG-TOKENS This data is generated by keeping the L1 class constant and alternating segments between two topics. We chose Italian for the L1 class and essays from the prompts "P7" and "P8" are used. The dataset, constructed from TOEFL11-TRAIN and TOEFL11-DEV, contains a total of 53 artificial documents, and will be used to verify that topic segmentation as discussed in Eisenstein and Barzilay (2008) functions as expected for data from this domain: that is, that topic change is detectable.

TOPICSEG-PTB Here the tokens in each text are replaced with their POS tags or n-grams over those tags, and the segmentation is performed over this data. In this dataset the tags are obtained via the Stanford Tagger and use the Penn Treebank (PTB) tagset. The same source data (TOEFL11-TRAIN and TOEFL11-DEV), L1 and topics as TOPICSEG-TOKENS are used for a total of 53 documents. This dataset will be used to investigate, *inter alia*, whether segmentation over these stylistically related features could take advantage of topic cues. We would expect not.

L1SEG-PTB This dataset is used for segmentation based on native language, also using (n-grams over) the PTB POS tags. We choose a specific topic and then retrieve all essays from the corpus that match this; here we chose prompt "P7", since it had the largest number of essays for our chosen single L1 pair, German–Italian. For the dataset constructed from TOEFL11-TRAIN and TOEFL11-DEV (which we will refer to as L1SEG-PTB-GI-DEV), this resulted in 57 documents. Documents that are composites of two L1s are then generated as described above. For investigating questions 2 and 3 above, we similarly have datasets constructed from the the smaller TOEFL11-TEST data (L1SEG-PTB-GI-TEST), which consist of 5 documents of 5 segments each for the single L1 pair, and from all language pairs (L1SEG-PTB-ALL-DEV, L1SEG-PTB-ALL-TEST). We would expect that these datasets should not be segmentable by topic, as all the segments are on the same topic; the segments should however, differ in stylistic characteristics related to the L1.

L1SEG-CLAWS2 This dataset is generated using the same methodology as L1SEG-PTB, with the exception that the essays are tagged using the RASP tagger which uses the more fine-grained CLAWS2 tagset, noting that the CLAWS2 tagset performed better in the NLI classification task (Malmasi and Cahill, 2015).

3.3 Evaluation

We use the standard P_k (Beeferman et al., 1999) and WindowDiff (WD) (Pevzner and Hearst, 2002) metrics, which (broadly speaking) select sentences using a moving window of size k and determines whether these sentences correctly or incorrectly fall into the same or different reference segmentations. P_k and WD scores range between 0 and 1, with a lower score indicating better performance, and 0 a perfect segmentation. It has been noted that some "degenerate" algorithms — such as placing boundaries randomly or at every possible position — can score 0.5 (Pevzner and Hearst, 2002). WD scores are typically similar to P_k, correcting for differential penalties between false positive boundaries and false negatives implicit in P_k. P_k and WD scores reported in §5 are

averages across all documents in a dataset. Formal definitions are given in Appendix A.3.

4 Segmentation Models

For all of our segmentation we use as a starting point the unsupervised Bayesian method of Eisenstein and Barzilay (2008); see §2.[3] We recap the important technical definitions here.

In Equation 1 of their work they define the observation likelihood as,

$$p(\mathbf{X} \mid \mathbf{z}, \Theta) = \prod_{t}^{T} p(\mathbf{x}_t \mid \theta_{z_t}), \quad (1)$$

where $\mathbf{X}$ is the set of all T sentences, $\mathbf{z}$ is the vector of segment assignments for each sentence, $\mathbf{x}_t$ is the bag of words drawn from the language model and Θ is the set of all K language models $\Theta_1 \ldots \Theta_K$. As is standard in segmentation work, K is assumed to be fixed and known (Malioutov and Barzilay, 2006); it is set to the actual number of segments. The authors also impose an additional constraint, that z_t must be equal to either z_{t-1} (the previous sentence's segment) or $z_{t-1}+1$ (the next segment), in order to ensure a linear segmentation.

This segmentation model has two parameters: the set of language models Θ and the segment assignment indexes $\mathbf{z}$. The authors note that since this task is only concerned with the segment assignments, searching in the space of language models is not desirable. They offer two alternatives to overcome this: (1) taking point estimates of the language models, which is considered to be theoretically unsatisfying and (2) marginalizing them out, which yields better performance. Equation 7 of Eisenstein and Barzilay (2008), reproduced here, shows how they marginalize over the language models, supposing that each language model is drawn from a symmetric Dirichlet prior (*i.e.* $\theta_j \sim \mathrm{Dir}(\theta_0)$):

$$p(\mathbf{X} \mid \mathbf{z}, \theta_0) = \prod_{j}^{K} p_{dcm}(\{\mathbf{x}_t : z_t = j\} \mid \theta_0) \quad (2)$$

The Dirichlet compound multinomial distribution p_{dcm} expresses the expectation over all the multinomial language models, when conditioned on the symmetric Dirichlet prior θ_0:

$$p_{dcm}(\{\mathbf{x}_t : z_t = j\} \mid \theta_0) = \frac{\Gamma(W\theta_0)}{\Gamma(N_j + W\theta_o)} \prod_{i}^{W} \frac{\Gamma(n_{j,i} + \theta_0)}{\Gamma(\theta_0)} \quad (3)$$

where W is the number of words in the vocabulary and $N_j = \sum_i^W n_{j,i}$, the total number of words in the segment j. They then observe that the optimal segmentation maximizes the joint probability

$$p(\mathbf{X}, \mathbf{z} \mid \theta_0) = p(\mathbf{X} \mid \mathbf{z}, \theta_0) p(\mathbf{z})$$

and assume a uniform $p(\mathbf{z})$ over valid segmentations with no probability mass assigned to invalid segmentations. The hyperparameter θ_0 can be chosen, or can be learned via an Expectation-Maximization process.

Inference Eisenstein and Barzilay (2008) defined two methods of inference, a dynamic programming (DP) one and one using Metropolis-Hastings (MH). Only MH is applicable where shifting a boundary will affect the probability of every segment, not just adjacent segments, as in their model incorporating cue phrases. Where this is not the case, they use DP inference. Their DP inference algorithm is suitable for all of our models, so we also use that.

Priors For our priors, we carry out a grid search on the devsets (that is, the datasets derived from TOEFL11-TRAIN and TOEFL11-DEV) in the interval $[0.1, 3.0]$, partitioned into 30 evenly spaced values; this includes both weak and strong priors.[4]

4.1 TOPICSEG

Our first model is exactly the one proposed by Eisenstein and Barzilay (2008) described above. The aim here is to look at how we perform at segmenting learner essays by topic in order to confirm that topic segmentation works for this domain and these types of topics. We apply this model to the TOPICSEG-TOKENS and TOPICSEG-PTB datasets where the texts have the same L1 and boundaries are placed between essays of differing topics (prompts).

4.2 L1SEG

Our second model modifies that of Eisenstein and Barzilay (2008) by revising the generative story.

[3] An open-source implementation of the method, called BAYESSEG, is made available by the authors at `http://groups.csail.mit.edu/rbg/code/bayesseg/`

[4] The Eisenstein and Barzilay (2008) code does implement an EM method for finding priors in the symmetric case, but we found that perhaps surprisingly the grid search almost always found better ones.

Where they assume a standard generative model over words with constraints on topic change between sentences, we make minor modifications to adapt the model for our task. The standard generative story (Blei, 2012) — an account of how a model generates the observed data — usually generates words in a two-stage process: (1) For each document, randomly choose a distribution of topics. (2) For each word in the document: (a) Assign a topic from those chosen in step 1. (b) Randomly choose a word from that topic's vocabulary.

Here we modify this story to be over part-of-speech data instead of lexical items. By using this representation (which as noted in §2 is useful for NLI classification) we aim to segment our texts based on the L1 of the author for each segment. For this model we only make use of the L1SEG-PTB-GI-DEV dataset.[5]

4.3 L1SEG-COMP

It is not obvious that the same properties that produce compact distributions in standard lexical chains would also be the case for POS data, particularly if extended to POS n-grams which can result in a very large number of potential tokens. In this regard Eisenstein and Barzilay (2008) note: "To obtain a high likelihood, the language models associated with each segment should concentrate their probability mass on a compact subset of words. Language models that spread their probability mass over a broad set of words will induce a lower likelihood. This is consistent with the principle of lexical cohesion."

Eisenstein and Barzilay (2008) discuss this within the context of topic segmentation.[6] However, it is unclear if this would also would happen for POS tags; there is no syntactic analogue for the sort of lexical chains important in topic segmentation. It may then turn out that using all POS tags or n-grams over them as in the previous model would not achieve a strong performance. We thus use knowledge from the NLI classification task to help.

Discarding Non-Discriminative Features One approach that could possibly overcome these limitations is the removal of features from the input space that have been found to be non-discriminative in NLI classification. This would allow us to encode POS sequence information via n-grams while also keeping the model's vocabulary sufficiently small. Doing this requires the use of extrinsic information for filtering the n-grams. The use of such extrinsic information has proven to be useful for other similar tasks such as the poetry style change segmentation work of Brooke et al. (2012), as noted in §2.

We perform this filtering using the discriminative feature lists derived from the NLI classification task using the system and method described in Malmasi and Dras (2014), also noted in §2. We extract the top 300 most discriminative POS n-gram features for each L1 from TOEFL11-TRAIN and TOEFL11-DEV, resulting in two lists of 600 POS bigrams and trigrams; these are thus independent of our test datasets. (We illustrate a text with respect to these discriminative features in Appendix A.4.) Note that discriminative n-grams can overlap with each other within the same class and also between two classes. We resolve such conflicts by using the weights of the features from the classification task as described in Malmasi and Dras (2014) and choosing the feature with the higher weight.

4.4 L1SEG-ASYMP

Looking at the distribution of discriminative features in our documents, one idea is that incorporating knowledge about which features are associated with which L1 could potentially help improve the results. One approach to do this is the use of asymmetric priors. We note that features associated with an L1 often dominate in a segment. Accordingly, priors can represent evidence external to the data that some some aspect should be weighted more strongly: for us, this is evidence from the NLI classification task. The segmentation models discussed so far only make use of a symmetric prior but later work mentions that it would be possible to modify this to use an asymmetric prior (Eisenstein, 2009).

Given that priors are effective for incorporating external information, recent work has highlighted the importance of optimizing over such priors, and in particular, the use of asymmetric priors. Key work on this is by Wallach et al. (2009) on LDA, who report that "an asymmetric Dirichlet prior over the document-topic distributions has substantial advantages over a symmetric prior",

[5] We also looked at including words. The results of these models were always worse, and we do not discuss them in this paper.

[6] For example, a topic segment related to the previously mentioned essay prompt P7 might concentrate its probability mass on the following set of words: {`education`, `learning`, `understanding`, `fact`, `theory`, `idea`, `concept`, `knowledge`}.

with prior values being determined through hyperparameter optimization. Such methods have since been applied in other tasks such as sentiment analysis (Lin and He, 2009; Lin et al., 2012) to achieve substantial improvements. For sentiment analysis, Lin and He (2009) incorporate external information from a subjectivity lexicon. In applying LDA, instead of using a uniform Dirichlet prior for the document–sentiment distribution, they use asymmetric priors for positive and negative sentiment, determined empirically.

For our task, we assign a prior to each of two languages in a document, one corresponding to $\mathrm{L1}_a$ and the other to $\mathrm{L1}_b$. Given this, we can assume that segments will alternate between $\mathrm{L1}_a$ and $\mathrm{L1}_b$. And instead of a single θ_0, we have two asymmetric priors that we call θ_a, θ_b corresponding to $\mathrm{L1}_a$ and $\mathrm{L1}_b$ respectively. This will require reworking the definition of p_{dcm} in Equation 3. First adapting Equation 2,

$$p(\mathbf{X} \mid \mathbf{z}, \theta_a, \theta_b) = \prod_{\{j_o\}} p_{dcm}(\{\mathbf{x}_t : z_t = j_o\} \mid \theta_a) \cdot \prod_{\{j_e\}} p_{dcm}(\{\mathbf{x}_t : z_t = j_e\} \mid \theta_b), \quad (4)$$

with $\{j_o\} = \{j \mid j \mod 2 = 1, 1 \leq j \leq K\}$ the set of indices over odd segments and $\{j_e\} = \{j \mid j \mod 2 = 0, 1 \leq j \leq K\}$ the set over evens. K is the (usual) total number of segments. Then

$$p_{dcm}(\{\mathbf{x}_t : z_t = j_o\} \mid \theta_a) = \frac{\Gamma(\sum_k^W \theta_a[k])}{\Gamma(N_{j_o} + \sum_k^W \theta_a[k])} \prod_i^W \frac{\Gamma(n_{j,i} + \theta_a[i])}{\Gamma(\theta_a[i])} \quad (5)$$

W is now more generally the number of items in our vocabulary (whether words or POS n-grams). A notational addition here is $\theta_a[k]$ which refers to the $\mathrm{L1}_a$ prior for the kth word or POS n-gram. There is an analogous p_{dcm} for θ_b.

The next issue is how to construct the θ_a and θ_b. The simplest scenario would require a single constant value for all elements in one L1 and another for all elements in the other L1. Specifically, using $\mathrm{discrim}(\mathrm{L1}_x)$ to denote "the ranked list of discriminative n-grams for $\mathrm{L1}_x$", we define

$$\theta_a[i] = \begin{cases} c_1 & \text{if } \theta_a[i] \in \mathrm{discrim}(\mathrm{L1}_a) \\ c_2 & \text{if } \theta_a[i] \in \mathrm{discrim}(\mathrm{L1}_b) \end{cases}$$

and analogously for $\theta_b[i]$. We would expect that $c_1 > c_2$ (*i.e.* the prior is stronger for elements that come from the appropriate ranked list of discriminative features), but these values will be learned.

In principle we would calculate versions of $p(\mathbf{X} \mid \mathbf{z}, \theta_a, \theta_b)$ twice: once where we assign θ_a to segment 1, and the second time where we assign θ_b. We'd then compare the two $p(\mathbf{X} \mid \mathbf{z}, \theta_a, \theta_b)$, and see which one fits better. In this work, however, we will fix the initial L1: segment 1 corresponds to $\mathrm{L1}_a$ and consequently has prior θ_a.[7]

5 Results

5.1 Segmenting by Topic

We begin by testing the TOPICSEG model to ensure that the Bayesian segmentation methodology can achieve reasonable results for segmenting learner essays by topic. The results on the TOPICSEG-TOKENS dataset (Table 1) show that content words are very effective at segmenting the writings by topic, achieving $\mathrm{P_k}$ values in the range 0.19–0.21. These values are similar to those reported for segmenting *Wall Street Journal* text (Beeferman et al., 1999). On the other hand, using the PTB POS tag version of the data in the TOPICSEG-PTB dataset results in very poor segmentation results, with $\mathrm{P_k}$ values around 0.45. This is essentially the same as the performance of degenerate algorithms (noted in §3.3) of 0.5. This demonstrates that, as expected, POS unigrams do not provide enough information for topic segmentation; it is not possible to construct even an approximation to lexical chains using them.

5.2 L1-based Segmentation

Having verified that the Bayesian segmentation approach is effective for topic segmentation on this data, we now turn to the L1SEG model for segmenting by the native language.

From the results in Table 1 we see very poor performance with a $\mathrm{P_k}$ value of 0.466 for segmenting the texts in L1SEG-PTB-GI-DEV using the unigrams as is. This was a somewhat unexpected result given than we know POS unigram distributions are able to capture differences between L1-groups (Malmasi and Cahill, 2015), albeit with limited accuracy. Moreover, neither bigram nor trigram encodings, which perform better at NLI, resulted in any improvement in our results.

[7] This requires an extension of the BAYESSEG software to support asymmetric priors. We will make this extended version of the code available under the same conditions as BAYESSEG. Please contact the first or second author for this.

Model	Dataset	Prior(s)	P_k	WD
TopicSeg	TopicSeg-Tokens	0.1	0.203	0.205
TopicSeg	TopicSeg-PTB	0.8	0.444	0.480
L1Seg	L1Seg-PTB-GI-Dev unigrams	0.1	0.466	0.489
L1Seg	L1Seg-PTB-GI-Dev bigrams	0.8	0.466	0.487
L1Seg	L1Seg-PTB-GI-Dev trigrams	0.8	0.480	0.489
L1Seg-Comp	L1Seg-PTB-GI-Dev bigrams	0.1	0.476	0.490
L1Seg-Comp	L1Seg-PTB-GI-Dev trigrams	0.4	0.393	0.398
L1Seg-Comp	L1Seg-CLAWS2-GI-Dev bigrams	0.4	0.387	0.400
L1Seg-Comp	L1Seg-CLAWS2-GI-Dev trigrams	0.4	0.370	0.373
L1Seg-AsymP	L1Seg-CLAWS2-GI-Dev trigrams	(0.6,0.3)	**0.316**	**0.318**

Table 1: Results on devsets for single L1 pair (German–Italian).

Model	P_k	WD
L1Seg-Comp	0.358	0.360
L1Seg-AsymP	**0.266**	**0.271**

Table 2: Results on testset L1Seg-CLAWS2-GI-Test for single L1 pair (German–Italian). Priors are the ones from the corresponding devsets in Table 1.

Model	P_k	WD
L1Seg-Comp	0.365 (0.014)	0.369 (0.019)
L1Seg-AsymP	**0.299** (0.022)	**0.312** (0.027)
L1Seg-Comp	0.376 (0.032)	0.381 (0.033)
L1Seg-AsymP	**0.314** (0.043)	**0.319** (0.045)

Table 3: Results on dev and test datasets (upper: L1Seg-CLAWS2-All-Dev, lower: L1Seg-CLAWS2-All-Test): means and standard deviations (in parentheses) across datasets for all 55 L1 pairs.

5.3 Incorporating Discriminative Features

Filtering the bigrams results in some minor improvements over the best results from the L1Seg model. However, there are substantial improvements when using the filtered POS trigrams, with a P_k value of 0.393. We did not test unigrams as they were the weakest NLI feature of the three.

This improvement is, we believe, because the Bayesian modelling of lexical cohesion over the input tokens requires that each segment concentrates its probability mass on a compact subset of words. In the context of the n-gram tokenization method tested in the previous section, the L1Seg model with n-grams would most likely exacerbate the issue by substantially increasing the number of tokens in the language model: while the unigrams do not capture enough information to distinguish non-lexical shifts, the n-grams provide too many features.

We also see that using the CLAWS2 tagset outperforms the PTB tagset. The results achieved for bigrams are much higher, while the trigram results are also better, with $P_k = 0.370$. NLI experiments using different POS tagsets have established that more fine-grained tagsets (*i.e.* those with more tag categories) provide greater classification accuracy when used as n-gram features for classification.[8] Results here comport with the previous findings.

As one of the two best models, we run it on the held-out test data, using the best priors found from the grid search on the devset data (Table 2); we find the P_k and WD values are comparable (and in fact slightly better), so the model still works if the filtering uses discriminative NLI features from the devset. Looking at results across all 55 L1 pairs (Table 3), we also see similar mean P_k and WD values with only a small standard deviation, indicating the approach works just as well across all language pairs. Priors here are all also weak, in the range $[0.1, 0.9]$.

In sum, the results here demonstrate the importance of inducing a compact distribution, which we did here by reducing the vocabulary size by stripping non-informative features.

5.4 Applying Two Asymmetric Priors

Our final model, L1Seg-AsymP, assesses whether setting different priors for each L1 can improve performance. Our grid search over two priors gives 900 possible prior combinations. These combinations also include cases where θ_a and θ_b are symmetric, which is equivalent to the L1Seg-Comp model. We observe (Table 1) that

[8] In §2 we noted the comparison of PTB and CLAWS2 tagsets in Malmasi and Cahill (2015); also, Gyawali et al. (2013) compared Penn Treebank and Universal POS tagsets and found that the more fine-grained PTB ones did better.

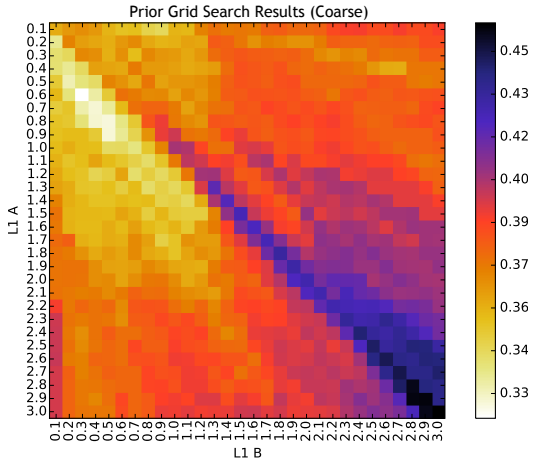

Figure 1: Heatmap over asymmetric priors on L1SEG-CLAWS2-ALL-DEV

the prior pair of $(0.6, 0.3)$ achieves a P_k value of 0.321, a substantial improvement over the previous best result of 0.370. Inspecting priors (see Figure 1 for a heatmap over priors) shows that the best results are in the region of weak priors for both values, which is consistent with the emphasis on compactness since weak priors result in more compact models (noted by *e.g.* Wang and Blei (2009)). Moreover, they are away from the diagonal, *i.e.* the L1SEG-COMP model will not produce the best results. A more fine-grained grid search, focusing on the range that provided the best results in the coarse search, can improve the results further still: over the interval $[0.3, 0.9]$, partitioned into 60 evenly spaced values, finds a prior pair of $(0.64, 0.32)$ that provides a slight improvement of the P_k value to 0.316.

As with L1SEG-COMP, we also evaluate this on the same held-out test set (Table 2). Applying the best asymmetric prior from the devset grid search, this improves to 0.266. Again, results across all 55 L1 pairs (Table 3) show the same pattern, and much as for L1SEG-COMP, priors are all weak or neutral (range $[0.1, 1.0]$). These results thus demonstrate that setting an asymmetric prior gives the best performance on this task.

6 Conclusion and Future Work

Applying the approach to our two illustrative applications of §1, patchwriting and literary analysis, would require development of relevant corpora. In both cases the distinction would be between native writing and writing that shows characteristics of a non-native speaker, rather than between two non-native L1s. There isn't yet a topic-balanced corpus like TOEFL11 which includes native speaker writing for evaluation, although we expect (given recent results on distinguishing native from non-native text in Malmasi and Dras (2015)) that the techniques should carry over. For the literary analysis, as well, to bridge the gap between work like Morzinski (1994) and a computational application, it remains to be seen how precise an annotation is possible for this task. Additionally, the granularity of segmentation may need to be finer than sentence-level, as suggested by the examples in §1; this level of granularity hasn't previously been tackled in unsupervised segmentation.

In terms of possible developments for the models presented for the task here, previous NLI work has shown that other, syntactic features can be useful for capturing L1-based differences. The incorporation of these features for this segmentation task could be a potentially fruitful avenue for future work. We have taken a fairly straightforward approach which modifies the generative story. A more sophisticated approach would be to incorporate features into the unsupervised model. One such example is the work of Berg-Kirkpatrick et al. (2010) which demonstrates that each component multinomial of a generative model can be turned into a miniature logistic regression model with the use of a modified EM algorithm. Their results showed that the feature-enhanced unsupervised models which incorporate linguistically-motivated features achieve substantial improvements for tasks such as POS induction and word segmentation. We note also that the models are potentially applicable to other stylistic segmentation tasks beyond L1 influence.

As far as this initial work is concerned we have shown that, framed as a segmentation task, it is possible to identify units of text that differ stylistically in their L1 influence. We demonstrated that it is possible to define a generative story and associated Bayesian models for stylistic segmentation, and further that segmentation results improve substantially by compacting the n-gram distributions, achieved by incorporating knowledge about discriminative features extracted from NLI models. Our best results come from a model that uses alternating asymmetric priors for each L1, with the priors selected using a grid search and then evaluated on a held-out test set.

Acknowledgements

The authors thank John Pate for very helpful discussions in the early stages of the paper, and the three anonymous referees for useful suggestions.

A Details on Dataset Generation and Evaluation

A.1 Document Generation

For our L1-varying datasets, we construct composite documents to be segmented as alternating segments drawn from TOEFL11 from two different L1s. Broadly following a standard approach (Brooke et al., 2012, for example), to generate such a document, we randomly draw TOEFL11 essays — each of which constitutes a segment — from the appropriate L1s and concatenate them, alternating the L1 class after each segment. This is repeated until the maximum number of segments per document, s, is reached. We generate multiple composite documents until all TOEFL11 have been used. In this work we use datasets generated with $s = 5$.[9] We follow the same process for our topic-varying datasets, but hold the L1 constant while alternating the topic (prompt).

A.2 Encoding n-gram information

Lau et al. (2013) investigated the importance of n-grams within topic models over lexical items. They note that in topic modelling each token receives a topic label and that the words in a collocation — *e.g. stock market*, *White House* or *health care* — may receive different topic assignments despite forming a single semantic unit. They found that identifying collocations (via a t-test) and preprocessing the text to turn these into single tokens provides a notable improvement over a unigram bag of words.

We implement a similar preprocessing step that converts each sentence within each document to a set of bigrams or trigrams using a sliding window, where each n-gram is represented by a single token. So, for example, the trigram `DT JJ NN` becomes a single token: `DT-JJ-NN`.

A.3 Evaluation: Metric Definitions

Given two segmentations r (reference) and h (hypothesis) for a corpus of N sentences,

[9] This is the average number of segments per chapter in the written text used by Eisenstein and Barzilay (2008). However, we have also successfully replicated our results using $s = 7, 9$.

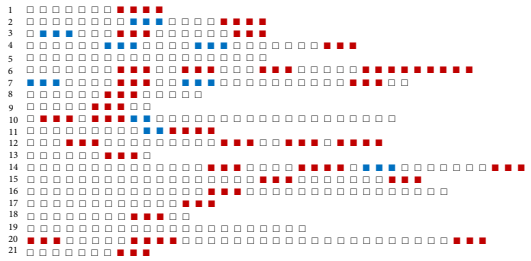

Figure 2: A visualization of sentences from a single segment. Each row represents a sentence and each token is represented by a square. Token trigrams considered discriminative for either of our two L1 classes are shown in blue or red, with the rest being considered non-discriminative.

$$P_D(r,h) = \sum_{1 \le i \le j \le N} D(i,j)(\delta_r(i,j) \bar{\oplus} \delta_h(i,j)) \quad (6)$$

where $\delta_r(i,j)$ is an indicator function specifying whether i and j lie in the same reference segment, $\delta_h(i,j)$ similarly for a hypothesised segment, $\bar{\oplus}$ is the XNOR function, and D is a distance probability distribution over the set of possible distances between sentences. For P_k, this D is defined by a fixed window of size k which contains all the probability mass, and k is set to be half the average reference segment length. The WD definition is:

$$WD(r,h) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(r_i, r_{i+k}) - b(h_i, h_{i+k})| > 0) \quad (7)$$

where $b(r_i, r_j)$ represents the number of boundaries between positions i and j in the reference text (similarly, the hypothesis text).

A.4 Visualisation of Discriminative Features

Figure 2 shows a visualization of the discriminative features of a single segment where each row represents a sentence and each token is represented by a square. Tokens that are part of a trigram which is considered discriminative for either of our two L1 classes are shown in blue or red. Note that discriminative n-grams can overlap with each other within the same class (*e.g.* on lines 1 and 2 where two overlapping trigrams form a group of four consecutive tokens) and also between two classes (*e.g.* on lines 10 and 11).

References

Navot Akiva and Moshe Koppel. 2013. A Generic Unsupervised Method for Decomposing Multi-Author Documents. *Journal of the American Society for Information Science and Technology (JASIST)* 64(11):2256–2264.

Khaled Aldebei, Xiangjian He, and Jie Yang. 2015. Unsupervised decomposition of a multi-author document based on naive-bayesian model. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 501–505. http://www.aclweb.org/anthology/P15-2082.

Calvin Bedient and Thomas Stearns Eliot. 1986. *He Do the Police in Different Voices: The Waste Land and its protagonist*. University of Chicago Press.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning* 34(1-3):177–210. https://doi.org/10.1023/A:1007506220214.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, pages 582–590. http://www.aclweb.org/anthology/N10-1083.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A Corpus of Non-Native English. Technical report, Educational Testing Service.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2012. Unsupervised Stylistic Segmentation of Poetry with Change Curves and Extrinsic Features. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, Montréal, Canada, pages 26–35. http://www.aclweb.org/anthology/W12-2504.

Julian Brooke, Graeme Hirst, and Adam Hammond. 2013. Clustering voices in the waste land. In *Proceedings of the Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, Atlanta, Georgia, pages 41–46. http://www.aclweb.org/anthology/W13-1406.

Serhiy Bykh and Detmar Meurers. 2014. Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* pages 1962–1973.

John Xiros Cooper. 1987. *TS Eliot and the politics of voice: The argument of The Waste Land*. 79. UMI Research Press.

Lan Du, Wray Buntine, and Mark Johnson. 2013. Topic segmentation with a structured topic model. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 190–200. http://www.aclweb.org/anthology/N13-1019.

Jacob Eisenstein. 2009. Hierarchical text segmentation from multi-scale lexical cohesion. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Boulder, CO, pages 353–361. www.aclweb.org/anthology/N09-1040.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pages 334–343. http://www.aclweb.org/anthology/D08-1035.

Rod Ellis. 2008. *The Study of Second Language Acquisition, 2nd edition*. Oxford University Press, Oxford, UK.

Joanna Gilmore, Denise Strickland, Briana Timmerman, Michelle Maher, and David Feldon. 2010. Weeds in the flower garden: An exploration of plagiarism in graduate students research proposals and its connection to enculturation, ESL, and contextual factors. *International Journal for Educational Integrity* 6(1):13–28.

Binod Gyawali, Gabriela Ramirez, and Thamar Solorio. 2013. Native Language Identification: a Simple n-gram Based Approach. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Atlanta, Georgia, pages 224–231. http://www.aclweb.org/anthology/W13-1729.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman Publishing Group.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Las Cruces, New Mexico, USA, pages 9–16. https://doi.org/10.3115/981732.981734.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Lingustics* 23(1):33–64. http://www.aclweb.org/anthology/J97-1003.

Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2014. Can characters reveal your native language? A language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1363–1373. http://www.aclweb.org/anthology/D14-1142.

Sandra Jamieson and Rebecca Moore Howard. 2013. Sentence-Mining: Uncovering the Amount Of Reading and Reading Comprehension In College Writers' Researched Writing. In Randall McClure and James P. Purdy, editors, *The New Digital Scholar: Exploring and Enriching the Research and Writing Practices of NextGen Students*, American Society for Information Science and Technology, Medford, NJ, pages 111–133.

Minwoo Jeong and Ivan Titov. 2010. Unsupervised discourse segmentation of documents with inherently parallel structure. In *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, Uppsala, Sweden, pages 151–155. http://www.aclweb.org/anthology/P10-2028.

Casey Keck. 2006. The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing* 15(4):261–278.

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 1356–1364. http://www.aclweb.org/anthology/P11-1136.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*. Springer-Verlag, volume 3495 of *LNCS*, pages 209–217.

Jey Han Lau, Timothy Baldwin, and David Newman. 2013. On Collocations and Topic Models. *ACM Transactions on Speech and Language Processing (TSLP)* 10(3).

Yongyan Li and Christine Pearson Casanave. 2012. Two first-year students strategies for writing from sources: Patchwriting or plagiarism? *Journal of Second Language Writing* 21:165–180.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. ACM, New York, NY, USA, pages 375–384. http://doi.acm.org/10.1145/1645953.1646003.

Chenghua Lin, Yulan He, Richard Everson, and Stefan Rüger. 2012. Weakly supervised joint sentiment-topic detection from text. *Knowledge and Data Engineering, IEEE Transactions on* 24(6):1134–1145.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, pages 25–32. https://doi.org/10.3115/1220175.1220179.

Shervin Malmasi and Aoife Cahill. 2015. Measuring feature diversity in native language identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Denver, Colorado, pages 49–55. http://www.aclweb.org/anthology/W15-0606.

Shervin Malmasi and Mark Dras. 2014. Language Transfer Hypotheses with Linear SVM Weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1385–1390. http://aclweb.org/anthology/D14-1144.

Shervin Malmasi and Mark Dras. 2015. Multilingual Native Language Identification. *Natural Language Engineering* https://doi.org/10.1017/S1351324915000406.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics* 17(1):21–48.

Mary Morzinski. 1994. *The Linguistic influence of Polish on Joseph Conrad's style*. Columbia University Press, New York, NY.

Diane Pecorari. 2003. Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing* 12(4):317–345.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics* 28(1):19–36. https://doi.org/10.1162/089120102317341756.

Ben Swanson and Eugene Charniak. 2013. Extracting the Native Language Signal for Second Language Acquisition. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 85–94. http://www.aclweb.org/anthology/N13-1009.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth*

Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, Atlanta, Georgia, pages 48–57. http://www.aclweb.org/anthology/W13-1706.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, pages 2585–2602. http://www.aclweb.org/anthology/C12-1158.

Michelle Vieyra, Denise Strickland, and Brianna Timmerman. 2013. Patterns in plagiarism and patchwriting in science and engineering graduate students' research proposals. *International Journal for Educational Integrity* 9(1):35–49.

Hanna M. Wallach, David M. Mimno, and Andrew McCallum. 2009. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, Curran Associates, Inc., pages 1973–1981.

Chong Wang and David M Blei. 2009. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in Neural Information Processing Systems*. pages 1982–1989.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2012. Exploring Adaptor Grammars for Native Language Identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, pages 699–709. http://www.aclweb.org/anthology/D12-1064.