Topic Modeling for Native Language Identification

Sze-Meng Jojo Wong Mark Dras Mark Johnson
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
{sze.wong,mark.dras,mark.johnson}@mq.edu.au

Abstract

Native language identification (NLI) is the task of determining the native language of an author writing in a second language. Several pieces of earlier work have found that features such as function words, part-of-speech n-grams and syntactic structure are helpful in NLI, perhaps representing characteristic errors of different native language speakers. This paper looks at the idea of using Latent Dirichlet Allocation as a feature clustering technique over lexical features to see whether there is any evidence that these smaller-scale features do cluster into more coherent latent factors, and investigates their effect in a classification task. We find that although (not unexpectedly) classification accuracy decreases, there is some evidence of coherent clustering, which could help with much larger syntactic feature spaces.

1 Introduction

Native language identification (NLI), the task of determining the native language of an author writing in a second language, typically English, has gained increased attention in recent years. The problem was first phrased as a text classification task by Koppel et al. (2005), using a machine learner with fundamentally lexical features — function words, character n-grams, and part-of-speech (PoS) n-grams. A number of subsequent pieces of work, such as that of Tsur and Rappoport (2007), Estival et al. (2007), Wong and Dras (2009) and Wong and Dras (2011), have taken that as a starting point, typically along with a wider range of features, such as document structure or syntactic structure.

Wong and Dras (2011) looked particularly at syntactic structure, in the form of production rules and parse reranking templates. They noted that they did not find the expected instances of clearly ungrammatical elements of syntactic structure indicating non-native speaker errors; instead there were just different distributions over regular elements of grammatical structure for different native languages. Our intuition is that it is several elements together that indicate particular kinds of indicative errors, such as incorrect noun-number agreement; and from this, that there might be coherent clusters of correlated features that are indicative of a particular native language. In this preliminary work, we investigate this using the basic lexical features of the original Koppel et al. (2005) model.

Latent Dirichlet Allocation (LDA) — a generative probabilistic model for unsupervised learning — was first introduced by Blei et al. (2003) to discover a set of latent mixture components known as *topics* which are representative of a collection of discrete data. The underlying idea of LDA is that each document from a text corpus is constructed according to a specific distribution of topics, in which words comprising the document are generated based on the word distribution for each selected topic; a topic is typically represented by a set of words such as *species*, *phylogenetic*, *evolution* and so on. Such a model allows multiple topics in one document as well as sharing of topics across documents within the corpus.

LDA can be viewed as a form of dimensionality reduction technique. In this paper, we intend to exploit LDA to discover the extent to which a lower dimension of feature space (i.e. a set of potentially

Sze-Meng Jojo Wong, Mark Dras and Mark Johnson. 2011. Topic Modeling for Native Language Identification. In *Proceedings of Australasian Language Technology Association Workshop*, pages 115−124

useful clusters of features) in each document affects classification performance. Here we are mapping clusters of features as 'topics' in typical LDA models and the posterior topic distributions inferred are to be used for classifying the native language of the authors against baseline models using the actual features themselves. We are particularly interested in whether the topics appear at all to form coherent clusters, and consequently whether they might potentially be applicable to the much larger class of syntactic features.

The remainder of this paper is structured as follows. In Section 2, we discuss some related work on the two key concepts of this paper: first relevant work in NLI, and then a brief description of LDA with its application to classification. We then describe both the topic models and the classification models used for the corpus to be examined, in Section 3. Section 4 presents classification results, and is followed by discussion in Section 5.

2 Related Work

2.1 Native Language Identification

Most of the existing work on native language identification adopts the supervised machine learning approach to classification. Koppel et al. (2005) is the earliest work in this classification paradigm using as features function words, character n-grams, and PoS bi-grams, together with some spelling mistakes. They used as their corpus the first version of *International Corpus of Learner English* (ICLE), selecting authors writing in English who have as their native language one of Bulgarian, Czech, French, Russian, or Spanish. Koppel et al. (2005) suggested that syntactic features (specifically errors) might be potentially useful, but only explored this idea at a rather shallow level by characterising ungrammatical structures with rare PoS bi-grams. This work of Koppel et al. (2005) was then investigated by Tsur and Rappoport (2007) to test their hypothesis that the choice of words in second language writing is highly influenced by the frequency of native language syllables, through measuring classification accuracy with only character bi-grams as features.

Another work with a similar goal, of developing profiles of authors, is that of Estival et al. (2007). They used a variety of lexical and document structure features over a set of three languages — English, Spanish and Arabic — also looking at predicting other demographic and psychometric author traits in addition to native language.

Wong and Dras (2009) first replicated the work of Koppel et al. (2005) with the three types of lexical feature as mentioned above and then extended the classification model with three syntactic errors commonly observed in non-native English users — subject-verb disagreement, noun-number disagreement and misuse of determiners — which had been identified as being influenced by the native language based on 'constrative analysis' (Lado, 1957). Although the overall classification did not improve over the lexical features alone, an ANOVA analysis showed that there were significant differences amongst different groups of non-native English users in terms of the errors made. In this work the classification task was carried out using the second version of ICLE (Granger et al., 2009), across seven languages (those of Koppel et al. (2005) with the two Asian languages Chinese and Japanese).

The later work of Wong and Dras (2011), on the same data, further explored the usefulness of syntactic features in a broader sense by characterising syntactic errors with cross sections of parse trees obtained from statistical parsing. More specifically, they utilised two types of parse tree substructure to use as classification features — horizontal slices of the trees as sets of CFG production rules and the feature schemas used in discriminative parse reranking (Charniak and Johnson, 2005). It was demonstrated that using these kinds of syntactic features performs significantly better than lexical features alone.

One key phenomenon observed by Wong and Dras (2011) was that there were different proportions of parse production rules indicative of particular native languages. One example is the production rule `NP` $\rightarrow$ `NN NN`, which appears to be very common amongst Chinese speakers compared with other native language groups; they claim that this is likely to reflect determiner-noun agreement errors, as that rule is used at the expense of one headed by a plural noun (`NP` $\rightarrow$ `NN NNS`). Our intuition here is that there might be coherent clusters of related features, with these clusters characterising typical errors or idiosyncrasies, that are predictive of a particular native language. In this paper we use LDA to cluster features, although in this preliminary work we use only the simpler lexical features of Wong and Dras

(2011).

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation is a Bayesian probabilistic model used to represent collections of discrete data such as text corpora, introduced by Blei et al. (2003). It addressed limitations of earlier techniques such as *probabilistic latent semantic indexing*, which is prone to overfitting and unable to generalise to unseen documents. LDA is a relaxation of classical document mixture models in which each document is associated with only a single topic, as it allows documents to be generated based on a mixture of topics with different distributions. We discuss the basic details of LDA, and our particular representation, in Section 3.1.

LDA has been applied to a wide range of tasks, such as building cluster-based models for ad hoc information retrieval (Wei and Croft, 2006) or grounded learning of semantic parsers (Börschinger et al., 2011). Relevant to this paper, it has been applied to a range of text classification tasks.

The original paper of Blei et al. (2003) used LDA as a dimensionality reduction technique over word unigrams for an SVM, for genre-based classification of Reuters news data and classification of collaborative filtering of movie review data, and found that LDA topics actually improved classification accuracy in spite of the dimensionality reduction. This same basic approach has been taken with other data, such as spam filtering of web text (Bíró et al., 2008), where LDA topics improved classification f-measure, or finding scientific topics from article abstracts (Griffiths and Steyvers, 2004), where LDA topics appear to be useful diagnostics for scientific subfields.

It has also been augmented in various ways: supervised LDA, where topic models are integrated with a response variable, was introduced by Blei and McAuliffe (2008) and applied to predicting sentiment scores from movie review data, treating it as a regression problem rather than a classification problem. Work by Wang et al. (2009) followed from that, extending it to classification problems, and applying it to the simultaneous classification and annotation of images. An alternative approach to joint models of text and response variables for sentiment classification of review texts (Titov and McDonald, 2008), with a particular focus on constructing topics related to aspects of reviews (e.g. food, decor, or service for restaurant reviews), found that LDA topics were predictively useful and seemed qualitatively intuitive.

In all of this preceding work, a document to be classified is represented by an exchangeable set of (content) words: function words are generally removed, and are not typically found in topics useful for classification. It is exactly these that are used in NLI, so the above work does guarantee that an LDA-based approach will be helpful here.

Two particularly relevant pieces of work on using LDA in classification are for the related task of authorship attribution, determining which author wrote a particular document. Rajkumar et al. (2009) claim that models with stopwords (function words) alone are sufficient to achieve high accuracy in classification, which seems to peak at 25 topics, and outperform content word-based models; the results presented in Table 2 and the discussion are, however, somewhat contradictory. Seroussi et al. (2011) also include both function words and content words in their models; they find that filtering words by frequency is almost always harmful, suggesting that function words are helping in this task.[1]

In this paper we will explore both function words and PoS n-grams, the latter of which is quite novel to our knowledge in terms of classification using LDA, to investigate whether clustering shows any potential for our task.

3 Experimental Setup

3.1 Mechanics of LDA

3.1.1 General Definition

Formally, each document is formed from a fixed set of vocabulary V and fixed set of topics T ($|T| = t$). Following the characterisation given by Griffiths and Steyvers (2004), the process of generating a corpus of m documents is as follows: first generate a set of multinomial distributions over topics $\boldsymbol{\theta}_j$ for each document D_j according to a T-dimensional Dirichlet distribution with concentration parameter $\boldsymbol{\alpha}$ (i.e. $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha})$); then generate a set of multinomial distributions $\boldsymbol{\phi}_i$ over the vocabulary V for each topic i according to a V-dimensional Dirichlet distribution with concentration parameter $\boldsymbol{\beta}$ (i.e. $\boldsymbol{\phi}_i \sim \text{Dir}(\boldsymbol{\beta})$);

[1] They note that for function words the term 'latent factor' is more appropriate than 'topic', with its connotation of semantic content.

and finally generate each of the n_j words for document D_j by selecting a random topic z according $\boldsymbol{\theta}_j$ and then drawing a word $w_{j,k}$ from $\boldsymbol{\phi}_z$ of the selected topic. The overall generative probabilistic model can be summarised as follows:

$$\begin{array}{ll} \boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha}) & j \in 1, ..., m \\ \boldsymbol{\phi}_i \sim \text{Dir}(\boldsymbol{\beta}) & i \in 1, ..., t \\ z_{j,k} \sim \boldsymbol{\theta}_j & j \in 1, ..., m, k \in 1, ..., n_j \\ w_{j,k} \sim \boldsymbol{\phi}_{z_{j,k}} & j \in 1, ..., m, k \in 1, ..., n_j \end{array}$$

From the inference perspective, given a corpus of m documents with n_j words each, the task is to estimate the posterior topic distributions $\boldsymbol{\theta}_j$ for each document D_j as well as the posterior word distributions $\boldsymbol{\phi}_i$ for each topic i that maximise the log likelihood of the corpus. As exact inference of these posterior distributions is generally intractable, there is a wide variety of means of approximate inference for LDA models which include approximation algorithms such as *Variational Bayes* (Blei et al., 2003) and expectation propagation (Minka and Lafferty, 2002) as well as *Markov Chain Monte Carlo* inference algorithm with Gibbs sampling (Griffiths and Steyvers, 2004).

3.1.2 LDA as PCFG

Johnson (2010) showed that LDA topic models can be regarded as a specific type of probabilistic context-free grammar (PCFG), and that Bayesian inference for PCFGs can be used to learn LDA models where the inferred distributions of PCFGs correspond to those distributions of LDA. A general schema used for generating PCFG rule instances for representing m documents with t topics is as follows:[2]

$$\begin{array}{ll} Sentence \rightarrow Doc'_j & j \in 1, ..., m \\ Doc'_j \rightarrow \_j & j \in 1, ..., m \\ Doc'_j \rightarrow Doc'_j\ Doc_j & j \in 1, ..., m \\ Doc_j \rightarrow Topic_i & i \in 1, ..., t; j \in 1, ..., m \\ Topic_i \rightarrow w & i \in 1, ..., t; w \in V \end{array}$$

Each of the rules in the PCFG is associated with a Bayesian inferred probability. The probabilities associated with the rules expanding $Topic_i$ correspond to the word distributions $\boldsymbol{\phi}_i$ of the LDA model, and the probabilities associated with the rules expanding Doc_j correspond to the topic distributions $\boldsymbol{\theta}_j$ of LDA. Similarly, inference on the posterior rule distributions can be approximated with Variational Bayes and Gibbs sampling. We use this PCFG formulation of LDA in this work.

[2] It should be noted that each document is given with a document identifier in which sentences in the document are prefixed with $\_j$.

3.2 Experimental Models

This section describes both the LDA models and the corresponding classification models used for our native language identification task on the ICLE corpus (Version 2) (Granger et al., 2009). Following Wong and Dras (2011), our experimental dataset consists of 490 essays written by non-native English users from seven different groups of language background — namely, Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese. There are 70 documents per native language.

Unlike the documents often inferred by LDA topic models which mostly consist of only content words, we represent our documents with function words instead, as this is typical for authorship related tasks, and does not allow unfair clues based on different distribution of domain discourses. In addition, we also experiment with documents represented by another type of lexical features for NLI, PoS bi-grams.

3.2.1 LDA Models for NLI

For each of the models we describe below, we experiment with different numbers of topics, $t = \{5, 10, 15, 20, 25\}$. In terms of the total number of PCFG rules representing each model, there are 490 of the first three rules as shown in the schema (Section 3.1.2), $490 \times t$ of the rule expanding $Doc_j \rightarrow Topic_i$, and $t \times v$ of the rule expanding $Topic_i \rightarrow w$ (see Table 1). All the inferences are performed with the PCFG-based Gibbs sampler implemented by Mark Johnson.[3]

FW-LDA Models The first LDA model is function word based. The vocabulary used for generating documents with this model is therefore a set of function words. We adopt the same set as used in Wong and Dras (2011) which consists of 398 words. An instance of the PCFG rule expanding $Topic_i \rightarrow w$ is $Topic_1 \rightarrow the$; there are 398 such rules for each topic.

[3] Software is available at `http://web.science.mq.edu.au/~mjohnson/Software.htm`.

LDA Models	$t = 5$	$t = 10$	$t = 15$	$t = 20$	$t = 25$
FW-LDA	5,910	10,350	14,790	19,230	23,670
POS-LDA	4,920	8,370	11,820	15,270	18,720
FW+POS-LDA	6,910	12,350	17,790	23,230	28,670

Table 1: Number of PCFG rules for each LDA model with different number of topics t

POS-LDA Models The second model is PoS bi-gram based. We choose bi-grams as it has been shown useful in Tsur and Rappoport (2007), and was used in Wong and Dras (2009). By tagging the 490 documents with Brill tagger (with Brown corpus tags), we extract the 200 most frequent occurring PoS bi-grams to form the vocabulary for this model. An instance of the PCFG rule expanding $Topic_i \rightarrow w$ is $Topic_1 \rightarrow$ *NN_NN*; there are 200 such rules for each topic.

FW-POS-LDA Models The third model combines the first two. We note that this is not typical of topic models: most form topics only over single types, such as content words.[4] The vocabulary then consists of both function words and PoS bi-grams with 598 terms in total. Thus, there are 598 instances of the rule expanding $Topic_i \rightarrow w$ for each topic.

3.2.2 Classification Models for NLI

Here we exploit LDA as a form of feature space dimension reduction to discover clusters of features as represented by 'topics' for classification. Based on each of the LDA models inferred, we take the posterior topic distributions to use as features for classifying into one of the seven native language classes. All the classifications are performed with a maximum entropy learner — MegaM (fifth release) by Hal Daumé III.[5]

Baselines Each LDA classification model (as described in the following) is compared against a corresponding baseline model. These sets of model use the actual features themselves for classification without feature reduction. There are three baselines: function word based with 398 features (FW-BASELINE), PoS bi-gram based with 200 features (POS-BASELINE), and the combination of the first two set of features (FW+POS-BASELINE). For each of these models, we examine two types of feature value — relative frequency and binary.

Function Words Features used in this model (FW-LDA) are the topic distributions inferred from the first LDA model. There are five variants of this based on number of topics (Section 3.2.1). The feature values are the posterior probabilities associated with the PCFG rules expanding $Doc_j \rightarrow Topic_i$ which correspond to the topic distributions $\boldsymbol{\theta_j}$ of the LDA representation.

PoS Bi-grams Similarly, this set of classification models (POS-LDA) uses the topic probabilities inferred from the second LDA model as features. Five variants of this with respect to the different topic numbers are examined as well.

Combined Features The last set of models combine both the function words and PoS bi-grams as classification features. The feature values are then the topic probabilities extracted from the last LDA model (the combined FW+POS-LDA model).

3.3 Evaluation

Often, LDA models are evaluated in terms of goodness of fit of the model to new data, by estimating the *perplexity* or similar of unseen held-out documents given some training documents (Blei et al., 2003; Griffiths and Steyvers, 2004). However, there are issues with all such proposed measures so far, such as importance sampling, harmonic mean, Chib-style estimation, and others; see Wallach et al. (2009) for a discussion. Alternatively, LDA models can be evaluated by measuring performance of some specific applications such as information retrieval and document classification (Titov and McDonald, 2008; Wang et al., 2009; Seroussi et al., 2011). We take this approach here, and adopt the standard measure for classification models — *classification accuracy* — as an indirect evaluation on our LDA models. The evaluation uses 5-fold cross-validation.

4 Classification Results

4.1 Baseline Models

Table 2 presents the classification accuracies achieved by the three baseline models mentioned above (i.e. using the actual features themselves without feature space reduction). These results are aligned with the results presented by Wong and Dras (2009) in their earlier work where binary feature

[4] Those that include multiple types typically treat them in different ways, such as in the separate treatment of content words and movie review ratings of Blei and McAuliffe (2008).

[5] MegaM is available at `http://www.cs.utah.edu/~hal/megam/`.

Baselines	Relative Freq	Binary
FW-BASELINE	33.26	62.45
POS-BASELINE	45.92	53.87
FW+POS-BASELINE	42.65	64.08

Table 2: Classification performance (%) of each baseline model – feature types of relative frequency and binary

values perform much better in general, although the results are lower because the calculation was made under cross-validation rather than on a separate held-out test set (hence with an effectively smaller amount of training data). Combining both the function words and PoS bi-grams yield a higher accuracy as compared to individual features alone. It seems that both features are capturing different useful cues that are predictive of individual native languages.

4.2 LDA Models

The classification performance for each of the LDA models is presented in Tables 3 to 5. Three sets of concentration parameters (Dirichlet priors) were tested on each of the three models to find the best fitted topic model: Table 3 contains results for uniform priors $\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta} = 1$ (the default); Table 4 is for $\boldsymbol{\alpha} = 50/t$ and $\boldsymbol{\beta} = 0.01$ (as per Steyvers and Griffiths (2007)); and Table 5 is for $\boldsymbol{\alpha} = 5/t$ and $\boldsymbol{\beta} = 0.01$ (since for us, with a small number of topics, the $\boldsymbol{\alpha} = 50/t$ of Steyvers and Griffiths (2007) gives much larger values of $\boldsymbol{\alpha}$ than was the case in Steyvers and Griffiths (2007)). On the whole, weaker priors ($\boldsymbol{\alpha} = 5/t$ and $\boldsymbol{\beta} = 0.01$) lead to a better model as evidenced by the accuracy scores.

As observed in Table 3, the model with 10 topics is the best model under uniform priors for both the individual feature-based models (FW-LDA and POS-LDA) with accuracies of 50.61% and 51.02% respectively, while the combined model (FW+POS-LDA) performs best at 55.51% with 15 topics. It should be noted that these are the outcomes of using the topic probabilities as feature value. (We also investigated the extent to which binary feature values could be useful by setting a probability threshold at 0.1; however, the results are consistently lower.)

By setting a stronger $\boldsymbol{\alpha} = 50/t$ and a much weaker $\boldsymbol{\beta} = 0.01$, the resulting models perform no better than those with uniform priors (see Table 4). The best performing models under this setting are with 25 topics for the individual feature-based models but with 20 topics for the combined

LDA Models	$t = 5$	$t = 10$	$t = 15$	$t = 20$	$t = 25$
FW-LDA	44.89	*50.61*	44.29	47.14	49.59
POS-LDA	47.35	*51.02*	50.00	50.61	49.79
FW+POS-LDA	49.79	54.08	*55.51*	52.86	53.26

Table 3: Classification performance (%) of each LDA-induced model ($\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta} = 1$); feature values of topic probabilities

LDA Models	$t = 5$	$t = 10$	$t = 15$	$t = 20$	$t = 25$
FW-LDA	32.45	42.45	44.29	45.71	*47.35*
POS-LDA	44.29	46.53	*50.82*	48.76	*50.82*
FW+POS-LDA	47.75	49.39	51.02	*54.49*	50.81

Table 4: Classification performance (%) of each LDA-induced model ($\boldsymbol{\alpha} = 50/t$ and $\boldsymbol{\beta} = 0.01$); feature values of topic probabilities

model. This setting of priors was found to work well for most of the text collections as suggested in Steyvers and Griffiths (2007). However, given that our topic sizes are just within the range of 5 to 25, we also tried $\boldsymbol{\alpha} = 5/t$. The classification results based on $\boldsymbol{\alpha} = 5/t$ and $\boldsymbol{\beta} = 0.01$ are showed in Table 5. This setting leads to the best accuracy (thus far) for each of the models with 25 topics — FW-LDA (52.45%), POS-LDA (53.47%), FW+POS-LDA (56.94%). The overall trajectory suggests that more than 25 topics might be useful.

Overall, the classification performance for each of the LDA-induced models (regardless of the parameter settings) performs worse than the baseline models (Section 4.1) where the actual features were used, contra the experience of Rajkumar et al. (2009) in authorship attribution. The drop is, however, only small in the case of PoS tags; the overall result is dragged down by the drop in function word model accuracies. And comparatively, they are still well above the majority baseline of 14.29% ($70/490$), so the LDA models are detecting something. On the one hand it is not surprising that reducing a relatively small feature space reduces performance; on the other hand, other work (as discussed in Section 2.2) had found that this had actually helped. While these results are not conclusive — a more systematic search might find better values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ — the results of the POS-LDA model suggests some promise for applying the method to a much larger feature space of similar terms: this could either be the unrestricted set of PoS bi-grams, or of syntactic structure features. We investigate this further by looking more deeply in Section 5 at some of the 'topics' (latent factors) found.

LDA Models	$t = 5$	$t = 10$	$t = 15$	$t = 20$	$t = 25$
FW-LDA	41.63	47.14	48.76	45.51	*52.45*
POS-LDA	43.47	49.79	51.22	52.86	*53.47*
FW+POS-LDA	51.84	50.61	53.88	52.62	*56.94*

Table 5: Classification performance (%) of each LDA-induced model ($\boldsymbol{\alpha} = 5/t$ and $\boldsymbol{\beta} = 0.01$); feature values of topic probabilities

5 Discussion

Despite the fact that all the LDA-induced models had lower accuracy scores than the baseline models, the inferred topics (clusters of related features) did demonstrate some useful cues that appear to be indicative of a particular native language. Here we present a discussion of three of these.

Analysis of FW-LDA It is often noted in the literature on second language errors that a typical error of Chinese speakers of English is with articles such as *a*, *an*, and *the*, as Chinese does not have these. Looking at the best performing FW-LDA model (weak priors of $\boldsymbol{\alpha} = 5/t$ and $\boldsymbol{\beta} = 0.01$; 25 topics), we observed that for the three topics — $Topic_8$ (the 8th feature), $Topic_{19}$ (the 19th feature) and $Topic_{20}$ (the 20th feature) — each of these is associated with a much higher feature weight for Chinese as compared to other native language groups (Table 6 shows the analysis on $Topic_8$). As for the function words clustered under these topics, *the* appears to be the most probable one with the highest probabilities of around 0.188, 0.181, and 0.146 for each respectively (i.e. the PCFG rules of $Topic_8 \rightarrow the$, $Topic_{19} \rightarrow the$, and $Topic_{20} \rightarrow the$); this is a higher weighting than for any other word in any topic. To verify that the topic model accurately reflects the data, we found that the relative frequency of *the* in the documents produced by Chinese learners is the highest in comparison with other languages in our corpus. It seems that Chinese learners have a tendency to misuse this kind of word in their English constructions, overusing *the*: this parallels the example given in Wong and Dras (2011), noted in Section 2.1, of the overuse of rules like NP → NN NN (rather than specifically ungrammatical constructions) characterising Chinese texts. However, there is no obvious pattern to the clustering (at least, that is evident to the authors)—if the clusters were to be grouping features in a way representative of errors, one of these topics might reflect misuse of determiners. But, none of these appear to: in $Topic_8$, for example, *a* appears only in 5th place, and no other determiners appear at all in the upper end of the distribution.

Language	Feature Weight	Relative Freq of *the*
Bulgarian	(relative to Bulgarian)	0.0814
Czech	-0.0457	0.0648
French	0.2124	0.0952
Russian	0.0133	0.0764
Spanish	-0.0016	0.0903
Chinese	*3.2409*	*0.1256*
Japanese	0.4485	0.0661

Table 6: Analysis on FW-LDA for $Topic_8$

Language	Feature Weight	Relative Freq of *NN_NN*
Bulgarian	(relative to Bulgarian)	0.0126
Czech	0.7777	0.0157
French	0.2566	0.0148
Russian	0.0015	0.0129
Spanish	0.0015	0.0142
Chinese	*2.4843*	*0.0403*
Japanese	0.4422	0.0202

Table 7: Analysis on POS-LDA for $Topic_1$

Analysis of POS-LDA However, there is a different story for POS-LDA, in terms of Chinese error phenomena. As shown in Table 7, Chinese has the highest feature weight for the first feature, $Topic_1$ (and also for $Topic_4$). To characterise this, we note that the PoS bi-gram NN_NN appears as the top bi-gram under $Topic_1$ (∼0.18) (and also occurs most frequently among Chinese learners as compared to other native language groups). Further, the next four bi-grams are NN_IN, AT_IN, IN_NN and NN_NNS, the last of which appears to be in complementary distribution in Chinese errors with NN_NN (i.e. Chinese speakers tend to use the singular more often in compound nouns, when a plural might be more appropriate). This observation also seems to be consistent with the finding of Wong and Dras (2011) in which the production rule NP → NN NN, reflecting determiner-noun disagreement, appears to be very common amongst Chinese learners. $Topic_1$ thus seems to be somehow connected with noun-related errors.

Our second instance to look at in some detail is

Language	Feature Weight	Relative Freq of *PPSS_VB*
Bulgarian	(relative to Bulgarian)	0.0111
Czech	0.7515	0.0137
French	-0.7080	0.0074
Russian	-0.2097	0.0116
Spanish	-0.3394	0.0117
Chinese	-0.1987	0.0059
Japanese	*2.0707*	*0.0224*

Table 8: Analysis on POS-LDA for $Topic_8$

Native Languages	Absolute Frequency										
	I	**They**	**Thou**	**We**	**You**	**it**	**she**	**they**	**we**	**you**	**Total**
Bulgarian	229	66	0	52	38	1	0	297	338	219	1240
Czech	483	188	0	166	34	1	0	459	348	202	1881
French	161	55	0	71	4	2	0	282	261	90	926
Russian	355	100	1	76	28	1	0	332	286	110	1289
Spanish	157	52	0	49	6	2	1	361	360	107	1095
Chinese	143	52	0	9	2	2	0	259	66	30	563
Japanese	*1062*	104	0	115	13	4	0	310	473	71	*2152*

Table 9: Pronoun usage across seven native language groups (absolute frequency of words tagged with PPSS)

for Japanese. Our expectation is that there are likely to be errors related to pronouns, as Japanese often omits them. In his comprehensive survey of second language acquisition, Ellis (2008) describes four measures of crosslinguistic influence: error (negative transfer), where differences between the languages lead to errors; facilitation (positive transfer), where similarities between the languages lead to a reduction in errors (relative to learners of other languages); avoidance, where constructions that are absent in the native language are avoided in the second language; and overuse, where constructions are used more frequently in an incorrect way in the second language, because of overgeneralisation.

A priori, it is difficult to predict which of these types of influence might be the case. The classic study of avoidance by Schachter (1974) examines Persian, Arab, Chinese, and Japanese learners of English, and their performance on using relative clauses. It found that even though Persian and Arabic have similar (right-branching) relative clauses to English, and Japanese and Chinese have different (left-branching) ones, the Japanese and Chinese learners made fewer errors; but that that was because they avoided using the construction. On the other hand, for a grammatically less complex phenomenon such as article use, several studies such as those of Liu and Gleason (2002) show that there can be a developmental aspect to crosslinguistic influence, with initial errors or avoidance turning to overuse because of overgeneralisation, which is later corrected; intermediate learners thus show the greatest level of overuse.

Looking at $Topic_8$ and $Topic_{20}$ under the POS-LDA model, relative to other topics inferred, top-ranking PoS bi-grams are mostly related to pronouns (such as PPSS_VB, PPSS_MD, and PPSS_VBD). Much higher feature weights are associated to these two topics for Japanese (as seen in Table 8 the analysis on $Topic_8$). Bi-grams of PPSS_VB and PPSS_MD occur much more often in Japanese learners' writings, and they are the first and the fifth terms under $Topic_8$, which seems to capture some of these phenomena.

To understand what these were saying about Japanese pronoun usage, we looked at a breakdown of pronoun use (see Table 9). Most apparently, the texts by Japanese speakers use more pronouns than any others. As the texts in the ICLE corpus are written by intermediate speakers, this could indicate a very strong instance of overuse. Looking at the distribution of pronouns, the Japanese speakers make much more use of the pronoun *I* than others: this has been noted elsewhere by Ishikawa (2011) on different corpora, particularly in the use of phrases such as *I think*. (The phrase *I think* is over-represented among Japanese speakers in our data also.)

Overall, then, POS-LDA seems to provide useful clustering of terms, while FW-LDA does not. This accords with the classification accuracies seen.

Analysis of FW+POS-LDA One question about the combined models was whether topics split along feature type — if that were the case, for a rough 2:1 ratio of function words to PoS bi-grams under 15 topics, there might be 10 topics whose upper rankings are dominated by function words, and 5 by PoS bi-grams. However, they are relatively evenly spread: for the top 20 words in each topic (uniform priors; 15 topics), the proportion of function words varied from 0.22 to 0.44, mean 0.339 and standard deviation 0.063. The topics thus appear to be quite mixed.

Looking into the combined model, $Topic_3$ and $Topic_{11}$ inferred by this model are amongst the features that associated with high feature weights for Chinese. Coinciding with our expectation, the two potential terms indicative of Chinese — NN_NN and *the* — topped the lists of $Topic_3$ and $Topic_{11}$ re-

spectively (where *the* also appears as the second most probable in $Topic_3$).

6 Conclusion

Although the LDA-induced classification models with feature space reduction somewhat underperformed in relation to the full feature-based models (the baselines), the 'topics' (latent factors) found appear in fact to be capturing useful information for individual native languages. Given the performance of POS-LDA, and the fact that the clustering seems more intuitive here, it seems promising to explore LDAs further with larger class of unrestricted PoS bi-grams, or of syntactic features such as the parse tree substructures used in Wong and Dras (2011). This could be complemented by using the adaptor grammars of Johnson (2010) to capture collocational pairings. Another potential approach that could be combined with this is to deploy the supervised LDA proposed by Blei and McAuliffe (2008), which might produce feature clusters that are more closely aligned to native language identification cues.

Acknowledgments

We acknowledge the support of ARC grant LP0776267. We also thank the anonymous reviewers, particularly for insightful critiques of the analysis of the topic models.

References

István Bíró, Jácint Szabó, and András A. Benczúr. 2008. Latent dirichlet allocation in web spam filtering. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*, pages 29–32, Beijing, China, April.

David Blei and Jon McAuliffe. 2008. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. MIT Press, Cambridge, MA.

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

Benjamin Börschinger, Bevan K. Jones, and Mark Johnson. 2011. Reducing grounded learning tasks to grammatical inference. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1416–1425, Edinburgh, Scotland, July.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180, Ann Arbor, Michigan, June.

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

Dominique Estival, Tanja Gaustad, Son-Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 263–272.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English (Version 2)*. Presses Universitaires de Louvain, Louvian-la-Neuve.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl. 1):5228–5235.

Shun'ichiro Ishikawa. 2011. A New Horizon in Learner Corpus Studies: The Aim of the ICNALE Project. In G. Weir, S. Ishikawa, and K. Poonpon, editors, *Corpora and Language Technologies in Teaching, Learning and Research*, pages 3–11. University of Strathclyde Press, Glasgow, UK.

Mark Johnson. 2010. PCFGs, Topic Models, Adaptor Grammars and Learning Topical Collocations and the Structure of Proper Names. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1157, Uppsala, Sweden, July.

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

Robert Lado. 1957. *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. University of Michigan Press, Ann Arbor, MI, US.

Dilin Liu and Johanna L. Gleason. 2002. Acquisition of the Article *the* by Nonnative Speakers of English: An Analysis of Four Nongeneric Uses. *Studies in Second Language Acquisition*, 24:1–26.

Thomas Minka and John Lafferty. 2002. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 352–359.

Arun Rajkumar, Saradha Ravi, Venkatasubramanian Suresh, M. Narasimha Murty, and C. E. Veni Madhavan. 2009. Stopwords and stylometry: A latent Dirichlet allocation approach. In *Proceedings of the NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond (Poster Session)*, Whistler, Canada, December.

J. Schachter. 1974. An error in error analysis. *Language Learning*, 27:205–214.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2011. Authorship attribution with latent Dirichlet allocation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 181–189, Portland, Oregon, June.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic Topic Models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*, chapter 21, pages 427–448. Lawrence Erlbaum Associates.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio, June.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16.

Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation Methods for Topic Models. In *Proceedings of the 26 th International Conference on Machine Learning*, Montreal, Canada.

Chong Wang, David Blei, and Fei-Fei Li. 2009. Simultaneous image classification and annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1903–1910, June.

Xing Wei and W. Bruce Croft. 2006. LDA-Based Document Models for Ad-hoc Retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference (SIGIR'06)*, pages 178–185.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia, December.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland, July.