Exploring Adaptor Grammars 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

The task of inferring the native language of an author based on texts written in a second language has generally been tackled as a classification problem, typically using as features a mix of n-grams over characters and part of speech tags (for small and fixed n) and unigram function words. To capture arbitrarily long n-grams that syntax-based approaches have suggested are useful, adaptor grammars have some promise. In this work we investigate their extension to identifying n-gram collocations of arbitrary length over a mix of PoS tags and words, using both maxent and induced syntactic language model approaches to classification. After presenting a new, simple baseline, we show that learned collocations used as features in a maxent model perform better still, but that the story is more mixed for the syntactic language model.

1 Introduction

The task of inferring the native language of an author based on texts written in a second language — *native language identification* (NLI) — has, since the seminal work of Koppel et al. (2005), been primarily tackled as a text classification task using supervised machine learning techniques. Lexical features, such as function words, character n-grams, and part-of-speech (PoS) n-grams, have been proven to be useful in NLI (Koppel et al., 2005; Tsur and Rappoport, 2007; Estival et al., 2007). The recent work of Wong and Dras (2011), motivated by ideas from Second Language Acquisition (SLA), has shown that syntactic features — potentially capturing syntactic errors characteristic of a particular native language — improve performance over purely lexical ones.

PoS n-grams can be leveraged to characterise surface syntactic structures: in Koppel et al. (2005), for example, ungrammatical structures were approximated by rare PoS bigrams. For the purpose of NLI, small n-gram sizes like bigram or trigram might not suffice to capture sequences that are characteristic of a particular native language. On the other hand, an attempt to represent these with larger n-grams would not just lead to feature sparsity problems, but also computational efficiency issues. Some form of feature selection should then come into play.

Adaptor grammars (Johnson, 2010), a hierarchical non-parametric extension of PCFGs (and also interpretable as an extension of LDA-based topic models), hold out some promise here. In that initial work, Johnson's model learnt collocations of arbitrary length such as *gradient descent* and *cost function*, under a topic associated with machine learning. Hardisty et al. (2010) applied this idea to perspective classification, learning collocations such as *palestinian violence* and *palestinian freedom*, the use of which as features was demonstrated to help the classification of texts from the Bitter Lemons corpus as either Palestinian or Israeli perspective.

Typically in NLI and other authorship attribution tasks, the feature sets exclude content words, to avoid unfair cues due to potentially different domains of discourse. In our context, then, what we are interested in are 'quasi-syntactic collocations' of either pure PoS (e.g. `NN IN NN`) or a mixture of PoS with function words (e.g. `NN of NN`). The particular question of interest for this paper, then, is to

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 699–709, Jeju Island, Korea, 12–14 July 2012.

investigate whether the power of adaptor grammars to discover collocations — specifically, ones of arbitrary length that are useful for classification — extends to features beyond the purely lexical.

We examine two different approaches in this paper. We first utilise adaptor grammars for discovery of high performing 'quasi-syntactic collocations' of arbitrary length as mentioned above and use them as classification features in a conventional maximum entropy (maxent) model for identifying the author's native language. In the second approach, we adopt a grammar induction technique to learn a grammar-based language model in a Bayesian setting. The grammar learned can then be used to infer the most probable native language that a given text written in a second language is associated with. The latter approach is actually closer to the work of Hardisty et al. (2010) using adaptor grammars for perspective modeling, which inspired our general approach. This alternative approach is also similar in nature to the work of Börschinger et al. (2011) in which grounded learning of semantic parsers was reduced to a grammatical inference task.

The structure of the paper is as follows. In Section 2, we review the existing work of NLI as well as the mechanics of adaptor grammars along with their applications to classification. Section 3 details the supervised maxent classification of NLI with collocation (n-gram) features discovered by adaptor grammars. The language model-based classifier is described in Section 4. Finally, we present a discussion in Section 5 and follow with concluding remarks.

2 Related Work

2.1 Native Language Identification

Most of the existing research treats the task of native language identification as a form of text classification deploying supervised machine learning approaches.

The earliest notable work in this classification paradigm is that of Koppel et al. (2005) using as features function words, character n-grams, and PoS bigrams, together with some spelling errors. Their experiments were conducted on English essays written by authors whose native language one of Bulgarian, Czech, French, Russian, or Spanish. The corpus used is the first version of *International Corpus of Learner English* (ICLE). Apart from investigating lexical features, syntactic features (errors in particular) were highlighted by Koppel et al. (2005) as potentially useful features, but they only explored this by characterising ungrammatical structures with rare PoS bigrams: they chose 250 rare bigrams from the Brown corpus.

Features for this task can include content words or not: Koppel et al. (2009), in reviewing work in the general area of authorship attribution (including NLI), discuss the (perhaps unreasonable) advantage that content word features can provide, and comment that consequently they "are careful ... to distinguish results that exploit content-based features from those that do not". We will not be using content words as features; we therefore note only approaches to NLI that similarly do not use them.

Following Koppel et al. (2005), Tsur and Rappoport (2007) replicated their work and hypothesised that word choices in second language writing is highly influenced by the frequency of native language syllables. They investigated this through measuring classification performance with only character bigrams as features.

Estival et al. (2007) tackled the broader task of developing profiles of authors, including native language and various other demographic and psychometric author traits, across a smaller set of languages (English, Spanish and Arabic). To this end, they deployed various lexical and document structure features.

Wong and Dras (2011), starting from the Koppel et al. (2005) approach, explored the usefulness of syntactic features in a broader sense in which they characterised syntactic errors with cross sections of parse trees obtained from statistical parsers, both horizontal slices of the parse trees in the form of CFG production rules, and the feature schemata used in discriminative parse reranking (Charniak and Johnson, 2005); they also found that using the top 200 PoS bigrams helped. Their results on the second version of the ICLE corpus, across seven languages (those of Koppel et al., plus two Oriental languages, Chinese and Japanese) demonstrated that syntactic features of these kinds lead to significantly better performance than the Koppel et al. features alone, with a top accuracy (on 5-fold cross-validation) of 77.75%.

Subsequently, Wong et al. (2011) explored Bayesian *topic modeling* (Blei et al., 2003; Griffiths and Steyvers, 2004) as a form of feature dimensionality reduction technique to discover coherent latent factors ('topics') that might capture predictive features for individual native languages. Their topics, rather than the typical word n-grams, consisted of bigrams over (only) PoS. However, while there was some evidence of topic cluster coherence, this did not improve classification performance.

The work of the present paper differs in that it uses Bayesian techniques to discover collocations of arbitrary length for use in classification, over a mix of both PoS and function words, rather than for use as feature dimensionality reduction.

2.2 Adaptor Grammars

Adaptor Grammars are a non-parametric extension to PCFGs that are associated with a Bayesian inference procedure. Here we provide an informal introduction to Adaptor Grammars; Johnson et al. (2007) provide a definition of Adaptor Grammars as a hierarchy of mixtures of Dirichlet (or 2-parameter Poisson-Dirichlet) Processes to which the reader should turn for further details.

Adaptor Grammars can be viewed as extending PCFGs by permitting the grammar to contain an unbounded number of productions; they are non-parametric in the sense that the particular productions used to analyse a corpus depends on the corpus itself. Because the set of possible productions is unbounded, they cannot be specified by simply enumerating them, as is standard with PCFGs. Instead, the productions used in an adaptor grammar are specified indirectly using a *base grammar*: the subtrees of the base grammar's "adapted non-terminals" serve as the possible productions of the adaptor grammar (Johnson et al., 2007), much in the way that subtrees function as productions in Tree Substitution Grammars .[1]

Another way to view Adaptor Grammars is that they relax the independence assumptions associated with PCFGs. In a PCFG productions are generated independently conditioned on the parent non-terminal, while in an Adaptor Grammar the probability of generating a subtree rooted in an adapted non-terminal is roughly proportional to the number of times it has been previously generated (a certain amount of mass is reserved to generate "new" subtrees). This means that the distribution generated by an Adaptor Grammar "adapts" based on the corpus being generated.

2.2.1 Mechanics of adaptor grammars

Adaptor Grammars are specified by a PCFG G, plus a subset of G's non-terminals that are called the *adapted non-terminals*, as well as a *discount parameter* a_A, where $0 \leq a_A < 1$ and a *concentration parameter* b_A, where $b > -a$, for each adapted non-terminal A. An adaptor grammar defines a two-parameter Poisson-Dirichlet Process for each adapted non-terminal A governed by the parameters a_A and b_A. For computational purposes it is convenient to integrate out the Poisson-Dirichlet Process, resulting in a predictive distribution specified by a Pitman-Yor Process (PYP). A PYP can be understood in terms of a "Chinese Restaurant" metaphor in which "customers" (observations) are seated at "tables", each of which is labelled with a sample from a "base distribution" (Pitman and Yor, 1997).

In an Adaptor Grammar, unadapted non-terminals expand just as they do in a PCFG; a production r expanding the non-terminal is selected according to the multinomial distribution θ_r over productions specified in the grammar. Each adapted non-terminal A is associated with its own Chinese Restaurant, where the tables are labelled with subtrees generated by the grammar rooted in A. In the Chinese Restaurant metaphor, the customers are expansions of A, each table corresponds to a particular subtree expanding A, and the PCFG specifies the base distribution for each of the adapted non-terminals. An adapted non-terminal A expands as follows. A expands to a subtree t with probability proportional to n_t, where n_t is the number of times t has been previously generated. In addition, A expands using a PCFG rule r expanding A with probability proportional to $(m_A\, a_A + b_A)\, \theta_r$, where m_A is the number of subtrees expanding A (i.e., the number of tables in A's restaurant). Because the underlying Pitman-Yor Processes have a "rich get richer" property, they generate power-law distributions over the subtrees for adapted non-terminals.

[1] For computational efficiency reasons Adaptor Grammars require the subtrees to completely expand to terminals. The *Fragment Grammars* of O'Donnell (2011) lift this restriction.

2.2.2 Adaptor grammars as LDA extension

With the ability to rewrite non-terminals to entire subtrees, adaptor grammars have been used to extend unigram-based LDA topic models (Johnson, 2010). This allows topic models to capture sequences of words with abitrary length rather than just unigrams of word. It has also been shown that it is crucial to go beyond the bag-of-words assumption as topical collocations capture more meaning information and represent more interpretable topics (Wang et al., 2007).

Taking the PCFG formulation for the LDA topic models, it can be modified such that each topic $Topic_i$ generates sequences of words by adapting each of the $Topic_i$ non-terminals (usually indicated with an *underline* in an adaptor grammar). The overall schema for capturing topical collocations with an adaptor grammar is as follows:

$$
\begin{array}{ll}
Sentence \rightarrow Doc_j & j \in 1, \ldots, m \\
Doc_j \rightarrow \_j & j \in 1, \ldots, m \\
Doc_j \rightarrow Doc_j\ Topic_i & i \in 1, \ldots, t; \\
 & j \in 1, \ldots, m \\
\underline{Topic_i} \rightarrow Words & i \in 1, \ldots, t \\
\overline{Words} \rightarrow Word & \\
Words \rightarrow Words\ Word & \\
Word \rightarrow w & w \in V
\end{array}
$$

There is a non-grammar-based approach to finding topical collocations as demonstrated by Wang et al. (2007). Both of these approaches learned useful collocations: for instance, as mentioned in Section 1, Johnson (2010) found collocations such *gradient descent* and *cost function* associated with the topic of machine learning; Wang et al. (2007) found the topic of human receptive system comprises of collocations such as *visual cortex* and *motion detector*.

Adaptor grammars have also been deployed as a form of feature selection in discovering useful collocations for perspective classification. Hardisty et al. (2010) argued that indicators of perspectives are often beyond the length of bigrams and demonstrated that the use of the adaptor grammar inferred n-grams of arbitrary length as features establishes the start-of-the-art performance for perspective classification on the Bitter Lemons corpus, depicting two different perspectives of Israeli and Pelestinian. We are adopting a similar approach in this paper for classifying texts with respect to the author's native language; but the key difference with Hardisty et al. (2010)'s approach is that our focus is on collocations that mix PoS and lexical elements, rather than being purely lexical.

3 Maxent Classification

In this section, we first explain the procedures taken to set up the conventional supervised classification task for NLI through the deployment of adaptor grammars for discovery of 'quasi-syntactic collocations' of arbitrary length. We then present the classification results attained based on these selected sets of n-gram features. In all of our experiments, we investigate two sets of collocations: pure PoS and a mixture of PoS and function words. The idea of examining the latter set is motivated by the results of Wong and Dras (2011) where inclusion of parse production rules lexicalised with function words as features had shown to improve the classification performance relative to unlexicalised ones.

3.1 Experimental Setup

3.1.1 Data and evaluation

The classification experiments are conducted on the second version of ICLE (Granger et al., 2009).[2] Following our earlier NLI work in Wong and Dras (2011), our data set consists of 490 texts written in English by authors of seven different native language groups: Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese. Each native language contributes 70 out of the 490 texts. As we are using a relative small data set, we perform k-fold cross-validation, choosing $k = 5$.

3.1.2 Adaptor grammars for supervised classification

We derive two adaptor grammars for the maxent classification setting, where each is associated with a different set of vocabulary (i.e. either pure PoS or the mixture of PoS and function words). We use

[2] Joel Tetreault and Daniel Blanchard from ETS have pointed out (personal communication) that there is a subtle issue with ICLE that could have an impact on the classification performance of NLI tasks; in particular, when character n-grams are used as features, some special characters used in some ICLE texts might affect performance. For our case, this should not be of much issue since they will not appear in our collocations.

the grammar of Johnson (2010) as presented in Section 2.2.2, except that the vocabulary differs: either $w \in V_{pos}$ or $w \in V_{pos+fw}$. For V_{pos}, there are 119 distinct PoS tags based on the Brown tagset. V_{pos+fw} is extended with 398 function words as per Wong and Dras (2011). $m = 490$ is the number of documents, and $t = 25$ the number of topics (chosen as the best performing one from Wong et al. (2011)).

Rules of the form $\texttt{Doc}_j \rightarrow \texttt{Doc}_j\ \texttt{Topic}_i$ that encode the possible topics that are associated with a document j are given similar $\boldsymbol{\alpha}$ priors as used in LDA ($\boldsymbol{\alpha} = 5/t$ where $t = 25$ in our experiments). Likewise, similar $\boldsymbol{\beta}$ priors from LDA are placed on the adapted rules expanding from $\underline{\texttt{Topic}_i} \rightarrow \texttt{Words}$, representing the possible sequences of words that each topic comprises ($\boldsymbol{\beta} = 0.01$).[3] The inference algorithm for the adaptor grammars are based on the Markov Chain Monte Carlo technique made available online by Johnson (2010).[4]

3.1.3 Classification models with n-gram features

Based on the two adaptor grammars inferred, the resulting collocations (n-grams) are extracted as features for the classification task of identifying authors' native language. These n-grams found by the adaptor grammars are only a (not necessarily proper) subset of those n-grams that are strongly characteristic of a particular native language. In principle, one could find all strongly characteristic n-grams by enumerating all the possible instances of n-grams up to a given length if the vocabulary is of a small enough closed set, such as for PoS tags, but this is infeasible when the set is extended to PoS plus function words. The use of adaptor grammars here can be viewed as a form of feature selection, as in Hardisty et al. (2010).

Baseline models To serve as a baseline, we take the commonly used PoS bigrams as per the previous work of NLI (Koppel et al., 2005). A set of 200 PoS bigrams is selected in two ways: the 200 most frequent in the training data (as in Wong and Dras (2011)) and the 200 with the highest information gain (IG) values in the training data (not evaluated in other work).

Enumerated n-gram models Here, we enumerate all the possible n-grams up to a fixed length and select the best of these according to IG, as a generalisation of the baseline. The first motivation for this feature set is that, in a sense, this should give a rough upper bound for the adaptor grammar's PoS-alone n-grams, as these latter should most often be a subset of the former. The second motivation is that it gives a robust comparison for the mixed PoS and function word n-grams, where it is infeasible to enumerate all of them.

ENUM-POS We enumerate all possible n-grams up to the length of 5, and select those that actually occur (i.e. of the $\sum_{i=1}^{5} 119^i$ possible n-grams, this is 218,042 based on the average of 5 folds). We look at the top n-grams up to length 5 selected by IG: the top 2,800 and the top 6,500 (for comparability with adaptor grammar feature sets, below), as well as the top 10,000 and the top 20,000 (to study the effect of larger feature space).

Adaptor grammar n-gram models The classification features are the two sets of selected collocations inferred by the adaptor grammars which are the main interest of this paper.

AG-POS This first set of the adaptor grammar-inferred features comprise of pure PoS n-grams (i.e. V_{pos}). The largest length of n-gram found is 17, but about 97% of the collocations are of length between 2 to 5. We investigate three variants of this feature set: top 200 n-grams of all lengths (based on IG), all n-grams of all lengths ($n = 2,795$ on average), and all n-grams up to the length of 5 ($n = 2,710$ on average).

AG-POS+FW This second set of the adaptor grammar-inferred features are mixtures of PoS and function words (i.e. V_{pos+fw}). The largest length of n-gram found for this set is 19 and the total number of different collocations found is much higher. For the purpose of comparability with the first set of adaptor grammar features, we investigate the following five variants for this feature set: top 200 n-grams of all lengths, all n-grams of all lengths ($n = 6,490$ on average), all n-grams up to the length of 5 ($n = 6,417$ on average), top 2,800 n-grams of all different lengths,

[3] The values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are also based on the established values presented in Wong et al. (2011).

[4] Adaptor grammar software is available on `http://web.science.mq.edu.au/~mjohnson/Software.htm`.

Features (n-grams)	Accuracy
BASELINE-POS [top200 MOST-FREQ]	53.87
BASELINE-POS [top200 IG]	56.12
AG-POS [top200 IG]	61.02
AG-POS [all $\leq$17-gram] ($n \approx 2800$)	68.37
AG-POS [all $\leq$ 5-gram] ($n \approx 2700$)	68.57
AG-POS+FW [top200 IG]	58.16
AG-POS+FW [all $\leq$19-gram] ($n \approx 6500$)	**74.49**
AG-POS+FW [all $\leq$5-gram] ($n \approx 6400$)	**74.49**
AG-POS+FW [top2800 IG $\leq$ 19-gram]	71.84
AG-POS+FW [top2800 IG $\leq$ 5-gram]	71.84
ENUM-POS [top2800 IG $\leq$ 5-gram]	69.79
ENUM-POS [top6500 IG $\leq$ 5-gram]	72.44
ENUM-POS [top10K IG $\leq$ 5-gram]	71.02
ENUM-POS [top20K IG $\leq$ 5-gram]	71.43

Table 1: Maxent classification results for individual feature sets (with 5-fold cross validation).

and top 2,800 n-grams up to the length of 5. (All the selections are based on IG).

In our models, all feature values are of binary type. For the classifier, we employ a maximum entropy (MaxEnt) machine learner — MegaM (fifth release) by Hal Daumé III.[5]

3.2 Classification results

Table 1 presents all the classification results for the individual feature sets, along with the baselines. On the whole, both sets of the collocations inferred by the adaptor grammars perform better than the two baselines. We make the following observations:

- Regarding ENUM-POS as a (rough) upper bound, the adaptor grammar AG-POS with a comparable number of features performs almost as well. However, because it is possible to enumerate many more n-grams than are found during the sampling process, ENUM-POS opens up a gap over AG-POS of around 4%.

- Collocations with a mix of PoS and function words do in fact lead to higher accuracy as compared to those of pure PoS (except for the top 200 n-grams); for instance, compare the 2,800 n-grams up to length 5 from the two corresponding sets (71.84 vs. 68.57).

- Furthermore, the adaptor grammar-inferred collocations with mixtures of PoS and function words (AG-POS+FW) in general perform better than our rough upper bound of PoS collocations, i.e. the enumerated PoS n-grams (ENUM-POS): the overall best results of the two feature sets are 74.49 and 72.44 respectively.

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

Features (n-grams)	Accuracy
AG-POS [all $\leq$ 5-gram] & FW	72.04
ENUM-POS [top2800 $\leq$ 5-gram] & FW	73.67
AG-POS+FW & AG-POS [a]	**75.71**
AG-POS+FW & AG-POS [b]	74.90
AG-POS+FW & ENUM-POS [top2800] [a]	73.88
AG-POS+FW & ENUM-POS [top2800] [b]	74.69
AG-POS+FW & ENUM-POS [top10K] [b]	74.90
AG-POS+FW & ENUM-POS [top20K] [b]	75.10

Table 2: Maxent classification results for combined feature sets (with 5-fold cross validation). [a]Features from the two sets are selected based on the overall top 3700 with highest IG; [b]features from the two sets are just linearly concatenated.

Given that the AG-POS+FW n-grams are capturing different sorts of document characteristics, they could potentially usefully be combined with the PoS-alone features. We thus combined them with both AG-POS and ENUM-POS feature sets, and the classification results are presented in Table 2. We tried two ways of integrating the feature sets: one way is to take the overall top 2,800 of the two sets based on IG; the other way is to just combine the two sets of features by concatenation of feature vectors (as indicated by *a* and *b* respectively in the result table). For comparability purposes, we considered only n-grams up to the length of 5. A baseline approach to this is just to add in function words as unigram features by feature vector concatenation, giving two further models, AG-POS [all $\leq$ 5-gram] & FW and ENUM-POS [top2800 $\leq$ 5-gram] & FW.

Overall, the classification accuracies attained by the combined feature sets are higher than the individual feature sets. The best performing of all the models is achieved by combining the mixed PoS and function word collocations with the adaptor grammar-inferred PoS, producing the best accuracy thus far of 75.71. This demonstrates that features inferred by adaptor grammars do capture some useful information and function words are playing a role. The way of integrating the two feature sets has different effects on the types of combination. As seen in Table 2, method *a* works better for the com-

bination of the two adaptor grammar feature sets; whereas method *b* works better for combining adaptor grammar features with enumerated n-gram features.

Using adaptor grammar collocations also outperforms the alternative baseline of adding in function words as unigrams. For instance, the best performing combined feature set of both AG-POS and AG-POS+FW does result in higher accuracy as compared to the two alternative baseline models, comparing 75.71 with 72.04 (and 75.71 with 73.67). This demonstrates that our more general PoS plus function word collocations derived from adaptor grammars are indeed useful, and supports the argument of Wang et al. (2007) that they are a useful technique for looking into features beyond just the bag of words.

4 Language Model-based Classification

In this section, we take a language modeling approach to native language identification; the idea here is to adopt grammatical inference to learn a grammar-based language model to represent the texts written by non-English native users. The grammar learned is then used to predict the most probable native language that a document (a sentence) is associated with.

In a sense, we are using a parser-based language model to rank the documents with respect to native language. We draw on the work of Börschinger et al. (2011) for this section. In that work, the task was grounded learning of a semantic parser. Training examples there consisted of natural language strings (descriptions of a robot soccer game) and a set of candidate meanings (actions in the robot soccer game world) for the string; each was tagged with a context identifier reflecting the actual action of the game. A grammar was then induced that would parse the examples, and was used on test data (where the context identifier was absent) to predict the context. We take a similar approach to developing an grammatical induction technique, although where they used a standard LDA topic model-based PCFG, we use an adaptor grammar. We expect that the results will likely to be lower than for the discriminative approach of Section 3. However, the approach is of interest for a few reasons: because, whereas the adaptor grammar plays an ancillary, feature selection role in Section 3, here the feature selection is an organic part of the approach as per the actual implementation of Hardisty et al. (2010); because adaptor grammars can potentially be extended in a natural way with unlabelled data; and because, for the purposes of this paper, it constitutes a second, quite different way to evaluate the use of n-gram collocations.

4.1 Language Models

We derive two adaptor grammar-based language models. One consists of only unigrams and bigrams, and the other finds n-gram collocations, in both cases over either PoS or the mix of PoS and function words. The assumption that we make is that each document (each sentence) is a mixture of two sets of topics: one is the native language-specific topic (i.e. characteristic of the native language) and the other is the generic topic (i.e. characteristic of the second language — English in our case). The generic topic is thus shared across all languages, and will behave quite differently from a language-specific topic, which is not shared. In other words, there are eight topics, representing seven native language groups that are of interest (Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese) and the second language English itself.[6]

Bigram models The following rule schema is applicable to both vocabulary types of PoS and the mixture of PoS and function words.

$$
\begin{aligned}
&Root \rightarrow \_lang\ langTopics \\
&langTopics \rightarrow langTopics\ langTopic \\
&langTopics \rightarrow langTopics\ nullTopic \\
&langTopics \rightarrow langTopic \\
&langTopics \rightarrow nullTopic \\
&\underline{langTopic} \rightarrow Words \\
&\underline{nullTopic} \rightarrow Words \\
&\underline{Words} \rightarrow Word\ Word \\
&Words \rightarrow Word \\
&Word \rightarrow w \qquad\qquad w \in V_{pos}; w \in V_{pos+fw}
\end{aligned}
$$

N-gram models The grammar is the same as the above with the exception that the non-terminal $Words$ is now rewritten as follows in order to

[6] We could just induce a regular PCFG here, rather than an adaptor grammar, by taking as terminals all pairs of PoS tags. We use the adaptor grammar formulation for comparability.

capture n-gram collocations of arbitrary length.

$Words \rightarrow Words\ Word$
$Words \rightarrow Word$

It should be noted that the two grammars above can in theory be applied to an entire document or on individual sentences. For this present work, we work on the sentence level as the run-time of the current implementation of the adaptor grammars grows proportional to the cube of the sentence length. For each grammar we try both sparse and uniform Dirichlet priors ($\boldsymbol{\alpha} = \{0.01, 0.1, 1.0\}$). The sparse priors encourage only a minority of the rules to be associated with high probabilities.

4.2 Training and Evaluation

As we are using the same data set as per the previous approach, we perform 5-fold cross validation as well. However, the training for each fold is conducted with a different grammar consisting of only the vocabulary that occur in each training fold. The reason is that we are now having a form of *supervised* topic models where the learning process is guided by the native languages. Hence, each of the training sentences are prefixed with the (native) language identifiers $\_lang$, as seen in the $Root$ rules of the grammar presented above.

To evaluate the grammars learned, as in Börschinger et al. (2011) we need to slightly modify the grammars above by removing the language identifiers ($\_lang$) from the $Root$ rules and then parse the *unlabeled* sentences using a publicly available CKY parser.[7] The predicted native language is inferred from the parse output by reading off the $langTopics$ that the $Root$ is rewritten to. We take that as the most probable native language for a particular test sentence. At the document level, we select as the class the language predicted for the largest number of sentences in that document.

4.3 Parsing Results

Tables 3 and 4 present the parsing results at the sentence level and the document level, respectively. On the whole, the results at the sentence level are much poorer as compared to those at the document level. In light of the results of Section 3.2, it is surprising that bigram models appear to perform better than n-gram models for both types of vocabulary, with the exception of AG-POS+FW at the document level. In fact, one would expect n-gram models to perform better in general as it is a generalisation that would contain all the potential bigrams. Nonetheless, the language models over the mixture of PoS and function words appear to be a more suitable representative of our learner corpus as compared to those over purely PoS, confirming the usefulness of integrated function words for the NLI classification task.

It should also be noted that sparse priors generally appear to be more appriopriate; except that for AG-POS+FW n-grams, uniform priors are indeed better and resulted in the highest parsing result of 50.15. (Although all the parsing results are much weaker as compared to the results presented in Section 3.2, they are all higher than the majority baseline of 14.29% i.e. $70/490$).

Features (n-grams)	Accuracy		
	($\alpha = 0.01$)	($\alpha = 0.1$)	($\alpha = 1.0$)
AG-POS [bigrams]	26.84	27.03	26.77
AG-POS [n-grams]	25.85	25.78	25.62
AG-POS+FW [bigrams]	28.58	28.40	27.43
AG-POS+FW [n-grams]	26.64	27.64	28.75

Table 3: Language modeling-based classification results based on parsing (at the sentence level).

Features (n-grams)	Accuracy		
	($\alpha = 0.01$)	($\alpha = 0.1$)	($\alpha = 1.0$)
AG-POS [bigrams]	41.22	38.88	39.69
AG-POS [n-grams]	36.12	34.90	35.20
AG-POS+FW [bigrams]	47.45	46.94	44.64
AG-POS+FW [n-grams]	43.97	49.39	**50.15**

Table 4: Language modeling-based classification results based on parsing (at the document level).

5 Discussion

Here we take a closer look at how well each approach does in identifying the individual native languages. The confusion matrix for the best model of two approaches are presented in Table 5 and Table 6. Both approaches perform reasonably well for the two Oriental languages (Chinese in particular); this is not a major surprise, as the two languages are not part of the language family that the rest of the languages come from (i.e. Indo-European). Under the supervised maxent classification, misclassifications largely are observed in the Romance ones (French and Spanish) as well as Russian; for the language model-based approach, Bulgarian is identi-

[7] CKY parser by Mark Johnson is available on `http://web.science.mq.edu.au/~mjohnson/Software.htm`.

	BL	CZ	RU	FR	SP	CN	JP
BL	[52]	5	7	4	2	-	-
CZ	5	[50]	5	3	4	-	3
RU	6	8	[46]	5	1	-	4
FR	7	3	5	[43]	8	-	4
SP	7	2	4	9	[47]	-	1
CN	-	-	-	-	-	[70]	-
JP	-	-	2	2	1	2	[63]

Table 5: Confusion matrix based on the best performing model under maxent setting (BL:Bulgarian, CZ:Czech, RU:Russian, FR:French, SP:Spanish, CN:Chinese, JP:Japanese).

	BL	CZ	RU	FR	SP	CN	JP
BL	[20]	32	9	6	-	1	2
CZ	2	[59]	3	1	-	-	5
RU	3	41	[19]	2	1	-	4
FR	8	20	4	[31]	4	-	3
SP	7	27	11	12	[9]	-	4
CN	-	2	-	2	-	[62]	4
JP	-	19	1	2	-	1	[47]

Table 6: Confusion matrix based on the best performing model under language modeling setting (BL:Bulgarian, CZ:Czech, RU:Russian, FR:French, SP:Spanish, CN:Chinese, JP:Japanese).

fied poorly, and Spanish moreso. However, the latter approach appears to be better in identifying Czech. On the whole, the maxent approach results in much fewer misclassifications compared to its counterpart.

In fact, there is a subtle difference in the experimental setting of the models derived from the two approaches with respect to the adaptor grammar: the number of topics. Under the maxent setting, the number of topics t was set to 25, while we restricted the models with the language modeling approach to only eight topics (seven for the individual native languages and one for the common second language, English). Looking more deeply into the topics themselves reveals that there appears to be at least two out of the 25 topics (from the supervised models) associated with n-grams that are indicative of the native languages, taking Chinese and Japanese as examples (see the associated topics in Table 7).[8] Perhaps associating each native language with only one generalised topic is not sufficient.

Furthermore, the distribution of n-grams among the topics (i.e. subtrees of collocations derived from the adaptor grammars) are quite different between the two approaches although the total number of n-grams inferred by each approach is about the same. For the language modeling ones, a high number of n-grams were associated with the generic topic $nullTopic$[9] and each language-specific topic $langTopic$ has a lower number of n-grams relative to bi-grams (Table 8) associated with it. For the maxent models, in contrast, the majority of the topics were associated with a higher number of n-grams (Table 9). The smaller number of n-grams to be used as features — and the fact that their extra length means that they will occur more sparsely in the documents — seems to be the core of the problem.

Top 10 Mixture N-grams			
Japanese		**Chinese**	
topic$_2$	**topic**$_{23}$	**topic**$_9$	**topic**$_{17}$
.	.	NN	.
we VB	PPSS VB	a NN	NN NN
our NNS	my NN	NN NN	NNS
our NN	CC	VBN by	NN
NN	VBG	NP .	RB ,
PPSS VB	PPSS think	NP	of NN
about	NN	:	JJ NN
because	PPSS VBD	(	NN .
it .	RB	as	VBG NN
we are	PPSS ’ NN	NN NN NN	NN NN NN

Table 7: Top mixture n-grams (collocations) for 4 out of the 25 topics representative of Japanese and Chinese (under maxent setting). N-grams of pronoun with verb are found at the upper end of $Topic_2$ and $Topic_{23}$ reflecting the frequent usage of Japanese; n-grams of noun are top n-grams under $Topic_9$ and $Topic_{17}$ indicating Chinese's common error of determiner-noun disagreement.

Nonetheless, the language models inferred discover relevant n-grams that are representative of individual native languages. For instance, the bigram NN NN, which Wong and Dras (2011) claim may reflect the error of determiner-noun disagreement commonly found amongst Chinese learners, was found under the Chinese topic at the top-2 position with a probability of 0.052 as compared to the other languages at the probability range of 0.0005-0.003. Similarly, one example for Japanese, the mixture bigram PPSS think, indicating frequent usage of pronouns within Japanese was seen under the Japanese topic at the top-9 position with a probability of 0.025 in relation to other languages within the range of 0.0002-0.006: this phenomenon as char-

[8] Taking the examples from Wong et al. (2011) as reference, we found similar n-grams that are indicative of Japanese and Chinese.

[9] This is quite plausible as there should be quite a number of structures that are representative of native English speakers that are shared by non-native speakers.

Model Types	N-gram Frequency															
	BGTopic		CZTopic		FRTopic		RUTopic		SPTopic		CNTopic		JPTopic		NullTopic	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
Bigrams	374	187	352	219	426	165	350	211	351	156	397	351	394	194	867	6169
N-grams	177	159	226	217	151	152	148	202	128	147	357	255	209	226	3089	7794

Table 8: Distribution of n-grams (collocations) for each topic under language modeling setting. (a) subcolumns are for n-grams of pure PoS and (b) subcolumns are for n-grams of mixtures of PoS and function words.

N-gram Frequency																				
Topic$_1$		**Topic**$_2$		**Topic**$_3$		**Topic**$_4$		**Topic**$_5$		**Topic**$_6$		**Topic**$_7$		**Topic**$_8$		**Topic**$_9$		**Topic**$_{10}$	
(a)	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**
174	443	145	441	136	245	141	341	236	519	169	748	127	340	182	473	109	339	190	236
Topic$_{11}$		**Topic**$_{12}$		**Topic**$_{13}$		**Topic**$_{14}$		**Topic**$_{15}$		**Topic**$_{16}$		**Topic**$_{17}$		**Topic**$_{18}$		**Topic**$_{19}$		**Topic**$_{20}$	
(a)	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**
57	259	126	455	103	543	211	225	170	459	81	309	238	207	152	475	119	452	333	423
Topic$_{21}$				**Topic**$_{22}$				**Topic**$_{23}$				**Topic**$_{24}$				**Topic**$_{25}$			
(a)		**(b)**		**(a)**		**(b)**		**(a)**		**(b)**		**(a)**		**(b)**		**(a)**		**(b)**	
245		341		168		492		194		472		201		366		195		190	

Table 9: Distribution of n-grams (collocations) for each topic under maxent setting. (a) subcolumns are for n-grams of pure PoS and (b) subcolumns are for n-grams of mixtures of PoS and function words.

Languages	Excerpts from ICLE
Chinese	... the overpopulation problem in urban area ... ... The development of country park can directly ... ... when it comes to urban renewal project ... ... As developing new town in ... ... and reserve some country park as ...
Japanese	... I think many people will ... ... I think governments should not ... ... I think culture is the most significant ... ... I think the state should not ... ... I really think we must live ...

Table 10: Excerpts from ICLE illustrating the common phenomena observed amongst Chinese and Japanese.

acteristic of Japanese speakers has also been noted for different corpora by Ishikawa (2011). (Note that this collocation as well as its pure PoS counterpart `PPSS VB` are amongst the top n-grams discovered under the maxent setting as seen in Table 7.) Table 10 presents some excerpts extracted from the corpus that illustrate these two common phenomena.

To investigate further the issue associated with the number of topics under the language modeling setting, we attempted to extend the adaptor grammar with three additional topics that represent the language family of the seven native languages of interest: Slavic, Romance, and Oriental. (The resulting grammar is presented as below.) However, the parsing result does not improve over the initial setting with eight topics in total.

$$Root \rightarrow \_lang\ langTopics$$
$$langTopics \rightarrow langTopics\ langTopic$$
$$langTopics \rightarrow langTopics\ familyTopic$$
$$langTopics \rightarrow langTopics\ nullTopic$$
$$langTopics \rightarrow langTopic$$
$$langTopics \rightarrow familyTopic$$
$$langTopics \rightarrow nullTopic$$
$$\underline{langTopic} \rightarrow Words$$
$$\underline{familyTopic} \rightarrow Words$$
$$\underline{nullTopic} \rightarrow Words$$
$$\underline{Words \rightarrow Words\ Word}$$
$$Words \rightarrow Word$$
$$Word \rightarrow w \qquad w \in V_{pos}; w \in V_{pos+fw}$$

6 Conclusion and Future Work

This paper has shown that the extension of adaptor grammars to discovering collocations beyond the lexical, in particular a mix of PoS tags and function words, can produce features useful in the NLI classification problem. More specifically, when added to a new baseline presented in this paper, the combined feature set of both types of adaptor grammar inferred collocations produces the best result in the context of using n-grams for NLI. The usefulness of the collocations does vary, however, with the technique used for classification.

Future work will involve a broader exploration of the parameter space of the adaptor grammars, in particular the number of topics and the value of α; a look at other non-parametric extensions of PCFGs, such as infinite PCFGs (Liang et al., 2007) for finding a set of non-terminals permitting more fine-grained topics; and an investigation of how the approach can be extended to semi-supervised learning to take advantage of the vast quantity of texts with errors available on the Web.

Acknowledgments

We would like to acknowledge the support of ARC Linkage Grant LP0776267. We also thank the anonymous reviewers for useful feedback. Much gratitude is due to Benjamin Börschinger for his help with the language modeling implementation.

References

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.

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.

Eric A. Hardisty, Jordan Boyd-Graber, and Philip Resnik. 2010. Modeling perspective using adaptor grammars. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 284–292.

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, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Advances in Neural Information Processing Systems 19: Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, pages 641–648.

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.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Percy Liang, Slav Petrov, Michael I. Jordan, and Dan Klein. 2007. The infinite pcfg using hierarchical dirichlet processes. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 688–697, Prague, Czech Republic, June.

Timothy O'Donnell. 2011. *Productivity and reuse in language*. Ph.D. thesis, Harvard University.

Jim Pitman and Marc Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900.

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.

Hans van Halteren. 2008. Source language markers in EUROPARL translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 937–944.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 697–702.

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.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2011. Topic modeling for native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 115–124, Canberra, Australia, December.