Classifying English Documents by National Dialect

Marco Lui♡♣ and Paul Cook♡
♡ Department of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
♣ NICTA Victoria Research Laboratory
{mhlui,paulcook}@unimelb.edu.au

Abstract

We investigate *national dialect identification*, the task of classifying English documents according to their country of origin. We use corpora of known national origin as a proxy for national dialect. In order to identify general (as opposed to corpus-specific) characteristics of national dialects of English, we make use of a variety of corpora of different sources, with inter-corpus variation in length, topic and register. The central intuition is that features that are predictive of national origin across different data sources are features that characterize a national dialect. We examine a number of classification approaches motivated by different areas of research, and evaluate the performance of each method across 3 national dialects: Australian, British, and Canadian English. Our results demonstrate that there are lexical and syntactic characteristics of each national dialect that are consistent across data sources.

1 Introduction

The English language exhibits substantial variation in its usage throughout the world with regional differences being noted at the lexical and syntactic levels (e.g., Trudgill and Hannah, 2008) between varieties of English such as that used in Britain and the United States. Although there are many varieties of English throughout the world — including, for example, New Zealand English, African American Vernacular English, and Indian English — there are a smaller number of so-called standard Englishes. British English and American English (or North American English) are often taken to be the two main varieties of standard English (Trudgill and Hannah, 2008; Quirk, 1995), with other varieties of standard English, such as Canadian English and Australian English, viewed as more-similar to one of these main varieties.

The theme of this work is *national dialect identification*, the classification of documents as one of a closed set of candidate standard Englishes (hereafter referred to as dialects), by exploiting lexical and syntactic variation between dialects. We make use of corpora of text of known national origin as a proxy for text of each dialect. Specifically, we consider Australian English, British English, and Canadian English, three so-called "inner circle" standard Englishes (Jenkins, 2009).[1]

This preliminary work aims to establish whether standard approaches to text classification are able to accurately predict the variety of standard English in which a document is written. The notion of standard English is differentiated from other factors such as style (e.g., formality) or topic Trudgill (1999), which are expected confounding factors. A model of dialect classification built on a single text type (e.g., standard national corpora) may be classifying documents on the basis of non-dialectal differences such as topic or genre. In order to control for the confounding factors, we utilize text from a variety of sources. By drawing training and test data from different sources, the successful transfer of models from one text source to another is evidence that the classifier is indeed capturing differences between different documents that are dialectal, rather than being due to any of the aforementioned confounding factors.

The main contributions of this paper are: (1) we introduce national dialect identification as a classification task, (2) we relate national dialect identification to existing research on text classification, (3) we assemble a dataset for national dialect identification using corpora from a variety of sources,

[1] We don't consider American English because of a rather surprising lack of available resources for this national dialect, discussed in Section 4.

Marco Lui and Paul Cook. 2013. Classifying English Documents by National Dialect. In *Proceedings of Australasian Language Technology Association Workshop*, pages 5−15.

(4) we empirically evaluate a number of text classification methods for national dialect identification, and (5) we find that we can train classifiers that are able to predict the national dialect of documents across data sources.

2 Related Work

National dialect identification is conceptually related to a range of established text classification tasks. In this section, we give some background on related areas, deferring the description of the specific methods we implement to Section 3.2.

2.1 Text Categorization

Text categorization has been described as the intersection of machine learning and information retrieval (Sebastiani, 2005), and is focused on tasks such as mapping newswire documents onto the topics they discuss (Debole and Sebastiani, 2005). A large variety of methods have been examined in the literature, due to the large overlap with the machine learning community (Sebastiani, 2002). One approach that has been shown to consistently perform well is the use of Support Vector Machines (SVM, Cortes and Vapnik, 1995). Joachims (1998) argued for their use in text categorization, observing that SVMs were well suited due to their ability to handle high-dimensional input spaces with few irrelevant features. Furthermore, he observed that most text categorization problems are linearly separable, a view that has been validated in a variety of studies (e.g., Yang and Liu, 1999; Drucker et al., 1999).

2.2 Language Identification

Language identification is the task of classifying a document according to the natural language it is written in. Recent work has applied language identification techniques to the identification of Dutch dialects, with encouraging results (Trieschnigg et al., 2012).

2.3 Native Language Identification (NLI)

Authorship profiling is an umbrella term for classification tasks that involve inferring some characteristic of a document's author, such as age, gender and native language (Estival et al., 2007). Native language identification (NLI, Koppel et al., 2005) is a well established authorship profiling task. The aim of NLI is to classify a document with respect to an author's native language, where this is not the language that the document is written in. One approach to NLI is to capture grammatical errors made by authors, through the use of contrastive analysis (Wong and Dras, 2009), parse structures (Wong and Dras, 2011) or adaptor grammars (Wong et al., 2012). Brooke and Hirst (2012) test a broad array of approaches to NLI, and specifically highlight issues with in-domain evaluation thereof.

2.4 Authorship Attribution

Authorship profiling focuses on identifying features which vary between groups of authors but are fairly consistent for a given group. In contrast, authorship attribution is the task of mapping a document onto a particular author from a set of candidate authors (Stamatatos, 2009), and is sometimes incorrectly conflated with authorship profiling. Mosteller and Wallace (1964) used a set of function words to attribute papers of disputed authorship. Other stylometric features used to identify authors include average sentence and word length (Yule, 1939). Modern features used for authorship attribution include distributions over function words (Zhao and Zobel, 2005), as well as features derived from parsing and part-of-speech tagging (Hirst and Feiguina, 2007). Author-aware topic models have also been proposed for authorship attribution (Seroussi et al., 2012).

2.5 Text-based Geolocation

Social media has recently exploded in popularity, with Twitter reporting that roughly 500 million tweets are sent each day (Twitter, 2013). There is a relationship between textual content and geolocation, with for example, texts containing words such as *streetcar*, *Maple Leafs*, and *DVP* likely being related to Toronto, Canada (Han et al., 2012).

Eisenstein et al. (2010) apply techniques from topic modeling to study variation in word usage on Twitter in the United States. Of particular relevance to our work, Wing and Baldridge (2011) and Roller et al. (2012) aggregate the tweets of users to predict their physical location in grid-based representations of the continental United States. These methods consider the KL-divergence between the distribution of words in a user's aggregated tweets and that of the tweets known to originate from each grid cell, with the most-similar cell being selected as the target user's most-likely location.

2.6 Computational Dialectal Studies

Although the specific issue of English national dialect classification has not been considered to date, a small number of computational studies have examined issues related to dialects. For example, Atwell et al. (2007) consider which variety of English, British or American, is most common on the Web. Peirsman et al. (2010) use techniques based on distributional similarity to identify lectal markers — words characteristic of one dialect versus another due to differences in sense or frequency — of dialects of Dutch. Zaidan and Callison-Burch (2012) studied dialect identification in Arabic dialects using automatic classifiers, and found that classifiers using dialectal data outperformed an informed baseline, achieving near-human classification accuracy.

Of particular relevance to our work, Cook and Hirst (2012) consider whether Web corpora from top-level domains (specifically `.ca` and `.uk`, in their work) represent corresponding national dialects (Canadian English and British English, respectively). They find that the relative distribution of spelling variants (e.g., the frequency of *color* relative to that of *colour*) is quite consistent across corpora of known national dialect. Furthermore, they show that these distributions are similar for corpora of known national dialect and Web corpora from a corresponding top-level domain.

3 Methodology

National dialect identification is a classification task, where each document must be mapped onto a single national dialect from a closed set of candidate dialects. We evaluate each method by training a classifier on a set of training documents and applying it to an independent set of test documents. For each experiment, we compute per-class precision, recall and F-score, using their standard definitions. We focus our evaluation on F-score, macroaveraged over all the per-class values, in order to maintain balance across precision and recall and across individual classes.

3.1 Cross-domain classification

A key challenge in evaluating national dialect identification as a text classification task is that documents in the training data may exhibit some non-dialectal variation that the classifiers may pick up on. For example, if British English were represented by a balanced corpus such as the British National Corpus (Burnard, 2000), but a corpus of say, newspaper texts, were used for American English (e.g., The New York Times Annotated Corpus, Sandhaus, 2008) then a classifier trained to distinguish between documents of these two corpora may pick up on differences in genre and topic as opposed to national dialect. Even if more-comparable corpora than those just mentioned above were chosen, because a corpus is a sample, certain topics or words will tend to be over- or under-represented. Indeed Kilgarriff (2001) points out such issues in the context of keyword comparisons of comparable corpora of British and American English, and Brooke and Hirst (2012) specifically highlight the same issue in native language identification.

In an effort to avoid this pitfall, we utilize text of known national origin from a variety of different sources. Specifically, we collect text representing each national dialect from up to 4 different sources (Section 4). In this paper, following the terminology of Pan and Yang (2010), we refer to each source as a *domain*, and acknowledge that this does not correspond to the topical sense of the term *domain* that is more common in NLP.

We cross-validate by holding out each source in turn, training a classifier on the union of the remaining sources and then applying it to the held-out source. By carrying out *cross-domain classification*, we mitigate the risk that confounding factors such as topic, genre or document length will misleadingly give high classification accuracy.

3.2 Classification Methods

We select methods from each field (Section 2) that are promising for national dialect identification.

3.2.1 BASELINE

We use a random classifier as our baseline, eschewing majority-class as it is not applicable in the cross-source context we consider; one of the primary differences anticipated between sources is that the relative distribution of classes will vary. The random classifier maps each document onto a dialect from our dialect set independently. It represents a trivial baseline that we expect all other classifiers to exceed.

3.2.2 TEXTCATEGORIZATION

We use the general text categorization approach proposed by Joachims (1998), applying a linear SVM to a standard bag-of-words representation.

3.2.3 NATIVELID

We use part-of-speech plus function word n-grams with a maximum entropy classifier (Wong and Dras, 2009). Wong and Dras aim to exploit grammatical errors, as contrastive analysis suggests that difficulties in acquiring a new language are due to differences between the new language and the native language of the learner, implying that the types of errors made are characteristic of the native language of the author. In national dialect identification, we do not expect grammatical errors to be as salient, because English is a national language of each of the countries considered. Nevertheless, part-of-speech plus function word n-grams are of interest because they roughly capture syntax — which is known to vary amongst national dialects (Trudgill and Hannah, 2008) — and are independent of the specific lexicalization.

3.2.4 AUTHORSHIPATTRIB

Authorship attribution is about modeling the linguistic idiosyncrasies of a particular author, in terms of some markers of the individual's style. Although in national dialect identification we do not assume that each document has a single unique author, we do assume that documents from the same country share stylistic properties resulting from the national dialect. We hypothesize that this results in systematic differences in the choice of function words (Zhao and Zobel, 2005). We capture this using a distribution over function words, which is a restricted bag-of-words model, where only words on an externally specified 'whitelist' are retained. We use the same stopword list as for native language identification as a proxy for function words. As per Zhao and Zobel (2005), we apply a naive Bayes classifier.

3.2.5 LANGID

We treat each dialect as a distinct language, and apply the language identification method of Lui and Baldwin (2011) in which documents are represented using a mixture of specially-selected byte sequences. The method specifically exploits differences in data sources to learn a set of byte sequences that is representative of languages (or in our case, dialects) across all the data sources. This feature selection is done by scoring each sequence using information gain (IG, Quinlan, 1993), with respect to each dialect as well as with each data source. This representation is then combined with a multinomial naive Bayes classifier.

3.2.6 GEOLOCATION

Our geolocation classifier is a nearest-prototype classifier using K-L divergence as the distance metric on a standard bag-of-words (Wing and Baldridge, 2011). The class prototypes are calculated from the concatenation of all members of the class. For both documents and classes, probability mass is assigned to unseen terms using a pseudo-Good-Turing smoothing, the parameters of which we estimate from the training data.

3.2.7 VARIANTPAIR

Motivated by Cook and Hirst's (2012) work on comparing dialects, our variant pair classifier uses the relative frequencies of spelling variants (e.g., *color*/*colour*, *yoghurt*/*yogurt*) to distinguish between dialects. For each of a set of ~1.8k spelling variant pairs from VarCon,[2] we calculate the frequency difference in a document between the first and second variant (e.g., freq(*color*) − freq(*colour*)). A standard vector-space model of similarity is used: each dialect is modeled as the sum of the vectors of all documents for that dialect; Cosine is used to map a given document to the most similar dialect.

4 Text Sources

4.1 NATIONAL

Large corpora are available for British and Canadian English. The written portion of the British National Corpus (BNC, Burnard, 2000) consists of roughly 87 million words of a variety of genres and topics from British authors from the late twentieth century. The Strathy Corpus[3] consists of roughly 40 million words of a variety of text types by Canadian authors from a similar time period. We use these two corpora in this study.

Appropriate resources are not available for American or Australian English. The Corpus of Contemporary American English (COCA, Davies, 2009) currently consists of over 450 million words of American English, but can only be accessed through a web interface; the full text form is unavailable. The American National Corpus (ANC, Ide, 2009) is much smaller than the BNC and Strathy Corpus at approximately only 11 million words.[4] In the case of Australian English, the Aus-

[2] `http://wordlist.sourceforge.net`

[3] `http://www.queensu.ca/strathy/`

[4] This figure refers specifically to the written portion of the Open ANC, the freely-available version of this corpus.

Domain	Australia			Canada			United Kingdom		
	#	μ	σ	#	μ	σ	#	μ	σ
NATIONAL	0	–	–	10000	2415.8	2750.4	10000	2742.3	2692.9
WEB	10000	2111.7	3261.5	10000	2459.4	3839.5	10000	2098.1	3527.4
WEBGOV	10000	1237.2	2706.3	10000	3980.4	4522.4	10000	2558.1	3327.4
TWITTER	1857	12.1	6.3	3598	11.8	6.3	24047	12.0	6.5

Table 1: Characteristics of the ENDIALECT dataset. # is the document count, μ and σ are the mean and standard deviation of document length (in words).

tralian Corpus of English (Green and Peters, 1991) consists of just 1 million words.[5]

4.2 WEB

The Web has been widely used for building corpora (e.g., Baroni et al., 2009; Kilgarriff et al., 2010) with Cook and Hirst (2012) presenting preliminary results suggesting that English corpora from top-level domains might represent corresponding national dialects of English. Australia, Canada, and the United Kingdom all have corresponding top-level domains that contain a wide variety of text types — namely `.au`, `.ca`, and `.uk`, respectively — from which we can build corpora. However, the top-level domain for the United States, `.us`, is primarily used for more-specialized purposes, such as government, and so a similar Web corpus cannot easily be built for American English. Here we build English Web corpora from `.au`, `.ca`, and `.uk` which — based on the findings of Cook and Hirst (2012) — we assume to represent Australian, Canadian, and British English, respectively.

One common method for corpus construction is to issue a large number of queries to a search engine, download the resulting URLs, and post-process the documents to produce a corpus (e.g., Baroni and Bernardini, 2004; Sharoff, 2006; Kilgarriff et al., 2010). Cook and Hirst (2012) use such a method to build corpora from the `.ca` and `.uk` domains; we follow their approach here. Specifically, we select alphabetic types in the BNC with character length greater than 2 and frequency rank 1001–5000 in the BNC as *seed words*. We then use Baroni and Bernardini's (2004) BootCaT tools to form 18k random 3-tuples from these seeds. We use the BootCaT tools to issue search engine queries for these tuples in the `.au`, `.ca`, and `.uk` domains. Using the BootCaT tools we then download the resulting URLs, and eliminate duplicates. We further eliminate non-English documents using `langid.py` (Lui and Baldwin, 2012). Following Cook and Hirst we only retain up to three randomly-selected documents per domain (e.g., `www.cbc.ca`). The final corpora consist of roughly 77, 96, and 115 million tokens for the `.au`, `.ca`, and `.uk` domains, respectively.

4.3 WEBGOV (Government)

The government of each of the countries considered in this study produces an enormous number of documents which can be used to build corpora. Furthermore, because many government websites are in particular second-level domains (e.g., `.gov.uk`) it is possible to easily construct a Web corpus consisting of such documents.

To build governmental Web corpora we follow a very similar process to that in the previous subsection, this time issuing queries for each of `.gov.au`, `.gc.ca`, and `.gov.uk`.[6] The resulting Australian, British, and Canadian government corpora contain roughly 199, 161, and 148 million words, respectively.[7]

4.4 TWITTER

Twitter[8] is an enormously popular micro-blogging service which has previously been used in studies of regional linguistic variation (e.g., Eisenstein et al., 2010). Twitter allows users to post short (up to 140 characters) messages known as *tweets*, and a recent report from Twitter indicates that roughly 500 million tweets are sent each day (Twitter, 2013). Crucially for this project, roughly 1% of tweets include geolocation metadata and

[5] The Australian National Corpus (`http://www.ausnc.org.au/`) is much larger, but consists of relatively little written material from the same time period as our other corpora.

[6] In this case there is an obvious domain to use to build an American government corpus, i.e., `.gov`. However, because we did not have a general Web corpus, or an appropriate national corpus, for American English, we did not build a government corpus for this dialect.

[7] There is a small amount of overlap between WEB and WEBGOV, with 3.7% of the WEB documents coming from governmental second-level domains.

[8] `http://twitter.com/`

can be used to build corpora known to correspond to a particular geographical region.

Using the Twitter API we collected a sample of tweets from October 2011 – January 2012 with geotags indicating that they were sent from Australia, Canada, or the United Kingdom.[9] We then filtered this collection to include only English tweets (again using `langid.py`). The resulting collection includes roughly 140k, 240k, and 1.4M tweets from Australia, Canada, and the United Kingdom, respectively.

5 The EnDialect dataset

The EnDialect dataset (Table 1), consists of 109502 documents in 3 English dialects (Australian, British, and Canadian) across 4 text sources (National, Web, WebGov and Twitter, described in Section 4). We conducted a pilot study, and found that across all the methods we test, the in-domain classification accuracy did not vary significantly beyond 5000 documents per dialect. Thus, for National, Web and WebGov, we retained 10000 documents per dialect. For Web and WebGov, we randomly sampled 10000 documents (without replacement) from each dialect. For National, the documents are substantially longer, and furthermore, documents from the (Canadian) Strathy Corpus are on average twice as long as those from the (British) BNC. In order to extract documents of comparable length to the Web and WebGov, we divided each document in National into equal-sized fragments (10 fragments per document for the BNC and 20 per document for the Strathy Corpus). We then sampled 10000 fragments from each, yielding pseudo-documents of comparable length to documents from Web and WebGov.

Constructing documents from the Twitter data is more difficult because individual messages are very short; preliminary experiments indicated that trying to infer dialect from a single message is nearly impossible. For Twitter, we therefore concatenate all documents from a given user to form a single pseudo-document per user. The Twitter crawl available to us had insufficient data to extract 10000 users per country, so we opted to retain all the users that had 15 or more messages in our data, giving us a total number of user pseudo-documents comparable to the number of documents for our other data sources (albeit with a skew between dialects that is not present for the other text sources).

6 Results

The first set of experiments we perform is in a leave-one-out cross-domain learning setting over our 4 text sources (referred to interchangeably as "domains") and 7 classification methods. We train one classifier for each pair of classification method and target domain, for a total of 28 classifiers. The training data used for each classifier is leave-one-out over the set of domains. For example, for any given classification method, the classifier applied to Web is trained on the union of data from National, WebGov, and Twitter.

Table 2 summarizes the macroaveraged F-score for each classifier in the cross-domain classification setting. We find that overall, the best methods for national dialect identification are TextCategorization and NativeLID. We also find that F-score varies greatly between target domains; in general, F-score is highest for National, and lowest for Twitter.

In this work, we primarily focus on cross-domain national dialect identification, for reasons discussed in Section 3.1. However, most of the methods we consider were not developed for cross-domain application, and thus in-domain results provide an interesting point of comparison. Hence, we present results from in-domain 10-fold cross-validation in Table 3 for comparison with the cross-domain outcome.

Our in-domain results are consistent with our cross-domain findings, in that methods that perform better in-domain tend to also perform better cross-domain, and target domains that are "easier" in-domain also tend to be "easier" cross-domain, "easier" meaning that all methods tend to attain better results. For most methods, the in-domain performance is better than the cross-domain performance, which is not surprising given that it is likely that there are particular terms that are predictive of a dialect in-domain that may not generalize across domains.

Overall, the results on in-domain and cross-domain classification suggest that TextCategorization is consistently the best among the methods compared across multiple domains, and that some domains are inherently easier for national

[9] Although an abundance of geolocated tweets are available for the United States, since we do not have corpora from the other sources for this national dialect we do not consider it here.

Approach	Target Domain NATIONAL (2-way)	 WEB (3-way)	 WEBGOV (3-way)	 TWITTER (3-way)
BASELINE	0.491	0.317	0.313	0.269
TEXTCATEGORIZATION	0.911	0.656	0.788	0.447
NATIVELID	0.812	0.606	0.480	0.314
AUTHORSHIPATTRIB	0.502	0.367	0.227	0.334
LANGID	0.772	0.538	0.597	0.043
GEOLOCATION	0.432	0.347	0.312	0.369
VARIANTPAIR	0.443	0.267	0.226	0.281

Table 2: Macroaverage F-score for cross-domain learning. For each domain/method combination, a classifier is trained on the union of the 3 non-target domains.

Approach	Target Domain NATIONAL (2-way)	 WEB (3-way)	 WEBGOV (3-way)	 TWITTER (3-way)
BASELINE	0.499	0.336	0.328	0.329
TEXTCATEGORIZATION	0.975	0.762	0.870	0.773
NATIVELID	0.946	0.577	0.708	0.521
AUTHORSHIPATTRIB	0.591	0.368	0.489	0.451
LANGID	–	–	–	–
GEOLOCATION	0.861	0.532	0.544	0.316
VARIANTPAIR	0.532	0.359	0.333	0.337

Table 3: Macroaverage F-score for in-domain (supervised) classification for each domain/method combination. (We do not have in-domain LANGID results as the method of Lui and Baldwin (2011) specifically requires cross-domain training data.)

dialect identification than others. To better understand the difference between domains, we conducted a further experiment, where we trained a classifier using each method on data from only one of our domains. We then applied this classifier to every other domain. We conducted this experiment for the two best-performing methods in the cross-domain setting: TEXTCATEGORIZATION and NATIVELID. The results of this experiment are summarized in Table 4.

The performance of classifiers trained on all non-test domains is generally better than that of classifiers trained on a single domain. The only exception to this is with classifiers trained on WEB applied to WEBGOV, which could be due to the noted overlap between these domains. However, this relationship is not symmetrical: classifiers trained only on WEBGOV do not perform better on WEB than classifiers trained on WEBGOV +NATIONAL +TWITTER.

7 Discussion

The high performance of TEXTCATEGORIZATION provides strong evidence of the viability of the cross-domain approach to identifying national dialect. This can be partly attributed to the much larger feature set of this method — to which no feature selection is applied — as compared to the other methods. The total vocabulary across all the datasets amounts to over 3 million unique terms. From this, the SVM algorithm was able to learn parameter weights that were applicable across domains — this can be seen from how the cross-domain text categorization results (Table 2) comfortably exceed the baseline in all domains.

AUTHORSHIPATTRIB uses a set of ~ 400 function words, in contrast to the ~ 3 million terms in the text categorization approach. The AUTHORSHIPATTRIB results are very close to the baseline in the cross-domain setting, suggesting that stylistic variation as captured by these features is not characteristic of English dialects.

F-scores for NATIVELID comfortably exceed the baseline, which suggests that English dialects have systematic differences at the syntactic level. The results are inferior to TEXTCATEGORIZATION, indicating that there are specific words that are predictive of national dialect across domains. This suggests there are systematic differences in the topics of discussion between documents of different origin, likely due to the discussion of specific locations. For example, analysis of our results indicates that (unsurprisingly) the term *Canada* is strongly associated with documents of Canadian origin.

The relatively poor performance of LANGID

Method	Training Domain	Target Domain			
		National (2-way)	Web (3-way)	WebGov (3-way)	Twitter (3-way)
TextCategorization	National	*0.975*	0.287	0.358	0.181
	Web	0.908	*0.762*	0.811	0.355
	WebGov	0.886	0.645	*0.870*	0.415
	Twitter	0.631	0.573	0.637	*0.773*
NativeLID	National	*0.946*	0.317	0.384	0.101
	Web	0.794	*0.577*	0.623	0.325
	WebGov	0.808	0.507	*0.708*	0.259
	Twitter	0.508	0.346	0.329	*0.521*

Table 4: Macroaverage F-score for pairwise cross-domain learning. Same-domain results (Table 3) are replicated in *italics* for comparison.

may be due to the small feature set. Lui and Baldwin (2011) select the top 400 features per language over 97 languages, so their feature set consists of 7480 features. We only consider 3 dialects, with a corresponding feature set of 1058 features. Though our features are clearly informative for the task (LangID results comfortably exceed the baseline), there may be useful information that is lost when a document is mapped into this reduced feature space. LangID performs exceptionally poorly when applied to Twitter in a cross-domain setting, because the classifier predicts a minority class 'Australian' for almost all documents. This is likely due to the lack of national corpus training data for 'Australian', as Table 4 suggests that national corpus data are an especially poor proxy for Twitter (a result consistent with the findings of Baldwin et al. (2013)).

The poor performance of the Geolocation is perhaps more surprising, as like TextCategorization this approach makes use of the full bag-of-words feature set. However, in the geolocation task of Wing and Baldridge (2011), the class space is much larger, and furthermore it is structured; classes correspond to regions of the Earth's surface, and the distance of the predicted region to the goldstandard region is taken into account in evaluation. The national dialect identification task is much more coarse-grained, potentially making it a poor match for geolocation methods.

VariantPair performs poorly throughout, with results below the random baseline in the cross-domain setting. The key difference between our national dialect identification task and the work of Cook and Hirst (2012) is that they classify entire corpora, whereas we classify individual documents. Documents are much shorter than corpora, and contain less spelling variation because they typically have a single author who is unlikely to choose different spellings of a given word.

8 Conclusion

Our cross-domain classification results strongly suggest that there are characteristics of each national dialect that are consistent across multiple domains. These characteristics go beyond simple topical differences, as representations such as function word distributions, and part-of-speech plus function word bigrams, omit topical information from consideration. Even without topical information, a classifier trained using techniques from native language identification is able to comfortably surpass a random baseline.

In future work, we intend to analyze the features weighted highly by our classifiers to potentially identify previously-undocumented differences between national dialects. Additionally, work on dialect identification might benefit methods for language identification. Prager (1999) finds that modeling Norwegian dialects separately improves language identification performance. In future work, we will examine if similarly modeling English dialects improves language identification.

Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Eric Atwell, Junaid Arshad, Chien-Ming Lai, Lan Nim, Noushin Rezapour Asheghi, Josiah Wang, and Justin Washtell. 2007. Which English dominates the World Wide Web, British or American? In *Proceedings of the Corpus Linguistics Conference (CL 2007)*. Birmingham, UK.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364. Asian Federation of Natural Language Processing, Nagoya, Japan.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. In *Proceedings of COLING 2012*, pages 391–408. The COLING 2012 Organizing Committee, Mumbai, India.

Lou Burnard. 2000. *The British National Corpus Users Reference Guide*. Oxford University Computing Services.

Paul Cook and Graeme Hirst. 2012. Do Web corpora from top-level domains represent national varieties of English? In *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*, pages 281–293. Liège, Belgium.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20:273–297.

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Franca Debole and Fabrizio Sebastiani. 2005. An analysis of the relative hardness of reuters-21578 subsets. *Journal of the American Society for Information Science and Technology*, 56(6):584–596.

Harris Drucker, Vladimir Vapnik, and Dongui Wu. 1999. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10:1048–1054.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Empirical Methods in Natural Language Processing*, pages 1277–1287. Cambridge, MA, USA.

Dominique Estival, Tanja Gaustad, and Ben Hutchinson. 2007. Author profiling for english emails. In *Proccedings of the 10th Conference for the Pacific Association for Computational Linguistics*, pages 263–272. Melbourne,Australia.

Elizabeth Green and Pam Peters. 1991. The Australian corpus project and Australian English. *International Computer Archive of Modern English*, 15:37–53.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062. The COLING 2012 Organizing Committee, Mumbai, India.

Graeme Hirst and Olga Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417.

Nancy Ide. 2009. The American National Corpus: Then, now, and tomorrow. In Michael Haugh, editor, *Selected Proceedings of the 2008 HCSNet Workshop on Designing an Australian National Corpus*, pages 108–113. Cascadilla Proceedings Project, Sommerville, MA.

Jennifer Jenkins. 2009. *World Englishes: A resource book for students*. Routledge, London, second edition.

Thorsten Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142. Chemnitz, Germany.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh PVS. 2010. A corpus factory for many languages. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pages 904–910. Valletta, Malta.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous authors native language. *Intelligence and Security Informatics*, 3495:209–217.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 553–561. Chiang Mai, Thailand.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30. Jeju, Republic of Korea.

Frederick Mosteller and David L. Wallace. 1964. *Inference and disputed authorship: The Federalist Papers*. Addison-Wesley, Reading,USA.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

Yves Peirsman, Dirk Geeraerts, and Dirk Speelman. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4):469–491.

John M. Prager. 1999. Linguini: language identification for multilingual documents. In *Proceedings the 32nd Annual Hawaii International Conference on Systems Sciences (HICSS-32)*. Maui, Hawaii.

John Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, USA.

Randolph Quirk. 1995. *Grammatical and lexical variance in English*. Longman, London.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Jeju Island, Korea.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia, PA.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Fabrizio Sebastiani. 2005. *Text categorization*, pages 109–129. TEMIS Text Mining Solutions S.A., Italy.

Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. 2012. Authorship attribution with author-aware topic models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 264–269. Association for Computational Linguistics, Jeju Island, Korea.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *Wacky! Working papers on the Web as Corpus*, pages 63–98. GEDIT, Bologna, Italy.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of The American Society for Information Science and Technology*, 60:538–556.

Dolf Trieschnigg, Djoerd Hiemstra, Mariët Theune, Franciska de Jong, and Theo Meder. 2012. An exploration of language identification techniques for the dutch folktale database. In *Proceedings of the LREC workshop on the Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects*.

Peter Trudgill. 1999. Standard English: What it isnt. In Tony Bex and Richard J. Watts, editors, *Standard English: The widening debate*, pages 117–128. Routledge, London.

Peter Trudgill and Jean Hannah. 2008. *International English: A guide to varieties of Standard English*. Hodder Education, London, fifth edition.

Twitter. 2013. New tweets per second record, and how! `https://blog.twitter.com/2013/new-tweets-`

per-second-record-and-how. Retrieved 19 August 2013.

Benjamin P. Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 955–964. Association for Computational Linguistics, Portland, Oregon, USA.

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

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 (EMNLP 2011)*, pages 1600–1610. Association for Computational Linguistics, Edinburgh, Scotland, UK.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2012. Exploring adaptor grammars for native language identification. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 699–709. Association for Computational Linguistics, Jeju Island, Korea.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, pages 42–49. ACM Press, New York, USA.

G. Udny Yule. 1939. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30:363–390.

Omar F Zaidan and Chris Callison-Burch. 2012. Arabic dialect identification. *Computational Linguistics*, 52(1).

Ying Zhao and Justin Zobel. 2005. Effective and Scalable Authorship Attribution Using Function Words. In *Asia Information Retrieval Symposium*, pages 174–189.