Exploring Methods and Resources for Discriminating Similar Languages

Marco Lui[♡♣], Ned Letcher[♡], Oliver Adams[♡], Long Duong[♡♣], Paul Cook[♡] and Timothy Baldwin[♡♣]

[♡] Department of Computing and Information Systems
The University of Melbourne
[♣] NICTA Victoria

mhlui@unimelb.edu.au, ned@nedletcher.net, oadams@student.unimelb.edu.au,
lduong@student.unimelb.edu.au, paulcook@unimelb.edu.au, tb@ldwin.net

Abstract

The *Discriminating between Similar Languages (DSL)* shared task at VarDial challenged participants to build an automatic language identification system to discriminate between 13 languages in 6 groups of highly-similar languages (or national varieties of the same language). In this paper, we describe the submissions made by team UniMelb-NLP, which took part in both the closed and open categories. We present the text representations and modeling techniques used, including cross-lingual POS tagging as well as fine-grained tags extracted from a deep grammar of English, and discuss additional data we collected for the open submissions, utilizing custom-built web corpora based on top-level domains as well as existing corpora.

1 Introduction

Language identification (LangID) is the problem of determining what natural language a document is written in. Studies in the area often report high accuracy (Cavnar and Trenkle, 1994; Dunning, 1994; Grefenstette, 1995; Prager, 1999; Teahan, 2000). However, recent work has shown that high accuracy is only achieved under ideal conditions (Baldwin and Lui, 2010), and one area that needs further work is accurate discrimination between closely-related languages (Ljubešić et al., 2007; Tiedemann and Ljubešić, 2012). The problem has been explored for specific groups of confusable languages, such as Malay/Indonesian (Ranaivo-Malancon, 2006), South-Eastern European languages (Tiedemann and Ljubešić, 2012), as well as varieties of English (Lui and Cook, 2013), Portuguese (Zampieri and Gebre, 2012), and Spanish (Zampieri et al., 2013). The *Discriminating Similar Language (DSL)* shared task (Zampieri et al., 2014) was hosted at the VarDial workshop at COLING 2014, and brings together the work on these various language groups by proposing a task on a single dataset containing text from 13 languages in 6 groups, drawn from a variety of news text datasets (Tan et al., 2014).

In this paper, we describe the entries made by team UniMelb NLP to the DSL shared task. We took part in both the closed and the open categories, submitting to the main component (Groups A-E) as well as the separate English component (Group F). For our closed submissions, we focused on comparing a conventional LangID methodology based on individual words and language-indicative letter sequences (Section 2.1) to a methodology that uses a de-lexicalized representation of language (Section 2.3). For Groups A-E we use cross-lingual POS-tagger adaptation (Section 2.3.1) to convert the raw text to a POS stream using a per-group tagger, and use n-grams of POS tags as our de-lexicalized representation. For English, we also use a de-lexicalized representation based on lexical types extracted from a deep grammar (Section 2.3.2), which can be thought of as a very fine-grained tagset. For the open submissions, we constructed new web-based corpora using a standard methodology, targeting per-language top-level domains (Section 2.4.2). We also compiled additional training data from existing corpora (Section 2.4.1).

Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, pages 129–138,
Dublin, Ireland, August 23 2014.

2 Overview

Our main focus was to explore novel methods and sources of training data for discriminating similar languages. In this section, we describe techniques and text representations that we tested, as well as the external data sources that we used to build language identifiers for this task.

2.1 Language-Indicative Byte Sequences

Lui and Baldwin (2011) introduced the $\mathcal{LD}$ feature set, a document representation for LangID that is robust to variation in languages across different sources of text. The $\mathcal{LD}$ feature set can be thought of as language-indicative byte sequences, i.e. sequences of 1 to 4 bytes that have been selected to be strongly characteristic of a particular language or set of languages regardless of the text source. Lui and Baldwin (2012) present `langid.py`,[1] an off-the-shelf LangID system that utilizes the $\mathcal{LD}$ feature set. In this work, we re-train `langid.py` using the training data provided by the shared task organizers, and use this as a baseline result representative of the state-of-the-art in LangID.

2.2 Hierarchical LangID

In LangID research to date, systems generally do not take into account any form of structure in the class space. In this shared task, languages are explicitly grouped into 6 disjoint groups. We make use of this structure by introducing a two-level LangID model. The first level implements a single group-level classifier, which takes an input sentence and identifies the language group (A–F) that the sentence is from. The output of this group-level classifier is used to select a corresponding per-group classifier, that is trained only on data for languages in the group. This per-group classifier is applied to the input sentence and the output thereof is the final label for the sentence.

2.3 De-Lexicalized Text Representation for DSL

One of the challenges in a machine learning approach to discriminating similar languages is to learn differences between languages that are truly representative of the distinction between varieties, rather than differences that are merely representative of peculiarities of the training data (Kilgarriff, 2001). One possible confounding factor is the topicality of the training data — if the data for each variety is drawn from different datasets, it is possible that a classifier will simply learn the topical differences between datasets. Diwersy et al. (2014) carried out a study of colligations in French varieties, where the variation in the grammatical function of noun lemmas was studied across French-language newspapers from six countries. In their initial analysis the found that the characteristic features of each country included the name of the country and other country-specific proper nouns, which resulted in near 100% classification accuracy but do not provide any insight into national varieties from a linguistic perspective.

One strategy that has been proposed to mitigate the effect of such topical differences is the use of a de-lexicalized text representation (Lui and Cook, 2013). The de-lexicalization is achieved through the use of a Part-Of-Speech tagger, which labels each word in a sentence according to its word class (such as Noun, Verb, Adjective etc). De-lexicalized text representations through POS tagging were first considered for native language identification (NLI), where they were used as a proxy for syntax in order to capture certain types of grammatical errors (Wong and Dras, 2009). Syntactic structure is known to vary across national dialects (Trudgill and Hannah, 2008), so Lui and Cook (2013) investigated POS plus function word *n*-grams as a proxy for syntactic structure, and used this representation to build classifiers to discriminate between Canadian, British and American English. They found that classifiers using such a representation achieved above-baseline results, indicating some systematic differences between varieties could be captured through the use of such a de-lexicalized representation. In this work, we explore this idea further — in particular, we examine (1) the applicability of de-lexicalized text representations to other languages using automatically-induced crosslingual POS taggers, and (2) the difference in accuracy for discriminating English varieties between representations based on a coarse-grained universal tagset (Section 2.3.1) as compared to a very fine-grained tagset used in deep parsing (Section 2.3.2).

[1] `http://github.com/saffsd/langid.py`

	Sandy	quit	on	Tuesday	Sandy	quit	Tuesday
UT	NOUN	VERB	ADP	NOUN	NOUN	VERB	NOUN
LTT	n_-_pn	v_np*	p_np_i-tmp	n_-_c-dow	n_-_pn	v_np*	n_-_c-dow
	British English				American English		

Table 1: Example of tags assigned with coarse-grained Universal Tagset (UT) and fine-grained lexical type tagset (LTT).

2.3.1 Crosslingual POS Tagging

A key issue in generating de-lexicalized text representations based on POS tags is the lack of availability of POS taggers for many languages. While some languages have some tools available for POS tagging (e.g. Treaties (Schmid, 1994) has parameter files for Spanish and Portuguese), the availability of POS taggers is far from universal. To address this problem for the purposes of discriminating similar languages, we draw on previous work in unsupervised cross-lingual POS tagging (Duong et al., 2013) to build a POS tagger for each group of languages, a method which we will refer to hereafter as "UMPOS".

UMPOS employs a 12-tag Universal Tagset introduced by Petrov et al. (2012), which consists of the tags *NOUN, VERB, ADJ* (adjective), *ADV* (adverb), *PRON* (pronoun), *DET* (determiner or article), *ADP* (preposition or postposition), *NUM* (numeral), *CONJ* (conjunction), *PRT* (particle), *PUNCT* (punctuation), and *X* (all other categories, e.g., foreign words or abbreviations). These twelve basic tags constitute a "universal" tagset in that they can be used to describe the morphosyntax of any language at a coarse level.

UMPOS generates POS taggers for new languages in an unsupervised fashion, by making use of parallel data and an existing POS tagger. The input for UMPOS is: (1) parallel data between the source and target languages; and (2) a supervised POS tagger for the source language. The output will be the tagger for the target language. The parallel data acts as a bridge to transfer POS annotation information from the source language to the target language.

The steps used in UMPOS are as follow. First, we collect parallel data which has English as the source language, drawing from Europarl (Koehn, 2005) and EUbookshop (Skadiņš et al., 2014). UMPOS word-aligns the parallel data using the Giza++ alignment tool (Och and Ney, 2003). The English side is POS-tagged using the Stanford POS tagger (Toutanova et al., 2003), and the POS tags are then projected from English to the target language based solely on one-to-one mappings. Using the sentence alignment score, UMPOS ranks the "goodness" of projected sentences and builds a seed model for the target language on a subset of the parallel data. To further improve accuracy, UMPOS builds the final model by applying self-training with revision to the rest of the data as follows: (1) the parallel corpus data is divided into different blocks; (2) the first block is tagged using the seed model; (3) the block is revised based on alignment confidence; (4) a new tagger is trained on the first block and then used to tag the second block. This process continues until all blocks are tagged. In experiments on a set of 8 languages, Duong et al. (2013) report accuracy of 83.4%, which is state-of-the-art for unsupervised POS tagging.

2.3.2 English Tagging Using ERG Lexical Types

Focusing specifically on language Group F — British English and American English — we leveraged linguistic information from the analyses produced by the English Resource Grammar (ERG: Flickinger (2002)), a broad-coverage, handcrafted grammar of English in the HPSG framework (Pollard and Sag, 1994) and developed within the DELPH-IN[2] research initiative. In particular, we extracted the lexical types assigned to tokens by the parser for the best analysis of each input string. In accordance with the heavily lexicalized nature of HPSG, lexical types are the primary means of distinguishing between different morphosyntactic contexts in which a given lexical entry can occur. They can be thought of as fine-grained POS tags, containing subcategorisation information in addition to part of speech information, and semantic information in cases that it directly impacts on morphosyntax. The version of the ERG we used (the "1212" release) has almost 1000 lexical types.

Table 1 illustrates an example of the type of syntactic variation that can be captured with the finer-

[2] http://www.delph-in.net

Group	Language	Code	Web Corpora		Existing Corpora	
			TLD	# words	# datasets	# words
A	Bosnian	bs	.ba	817383	4	715602
A	Croatian	hr	.hr	43307311	5	1536623
A	Serbian	sr	.rs	1374787	4	1204684
B	Indonesian	id	.id	23812382	3	564824
B	Malaysian	my	.my	2596378	3	535221
C	Czech	cz	.cz	17103140	8	2181486
C	Slovakian	sk	.sk	17253001	8	2308083
D	Brazilian Portuguese	pt-BR	.br	27369673	4	860065
D	European Portuguese	pt-PT	.pt	22620401	8	2860321
E	Argentine Spanish	es-AR	.ar	45913651	2	619500
E	Peninsular Spanish	es-ES	.es	30965338	9	3458462
F	British English	en-GB	.uk	20375047	1	523653
F	American English	en-US	.us	21298230	1	527915

Table 2: Word count of training data used for open submissions.

grained lexical types, that would be missed with the coarse-grained universal tagset. In American English, both *Sandy resigned on Tuesday* and *Sandy resigned on Tuesday* are acceptable whereas British English does not permit the omission of the preposition before dates. In the coarse-grained tagset, the American English form results in a sequence VERB : NOUN, which is not particularly interesting as we expect this to occur in both English varieties, whereas the fine-grained lexical types allow us to capture the sequence v_np*_ntr : n_-_c-dow (verb followed by count noun [day of week]), which we expect to see in American English but not in British English.

Since the ERG models a sharp notion of grammaticality, not all inputs receive an analysis — whether due to gaps in the coverage of the grammar or genuinely ungrammatical input. The ERG achieved a coverage of 86% over the training data across both British English and American English. Sentences which failed to parse were excluded from use as input into the classifier. However the inability to classify any sentence which we cannot parse is unsatisfactory. We solved this problem by generating lexical type features for sentences which failed to parse using the ERG-trained übertagger of Dridan (2013), which performs both tokenisation and supertagging of lexical types and improves parser efficiency by reducing ambiguity in the input lattice to the parser.

2.4 External Corpora

The DSL shared task invited two categories of participation: (1) Closed, using only training data provided by the organizers (Tan et al., 2014); and (2) Open, using any training data available to participants. To participate in the latter category, we sourced additional training data through: (1) collection of data relevant to this task from existing text corpora; and (2) automatic construction of web corpora. The information about the additional training data is shown in Table 2.

2.4.1 Existing Corpora

We collected training data from a number of existing corpora, as shown in Table 3. Many of the corpora that we used are part of OPUS (Tiedemann, 2012), which is a collection of sentence-aligned text corpora commonly used for research in machine translation. The exceptions are: (1) debian, which was constructed using translations of message strings from the Debian operating system,[3]; (2) BNC — the British National Corpus (Burnard, 2000); (3) OANC — the open component of the Second Release of the American National Corpus (Ide and Macleod, 2001), and (4) Reuters Corpus Volume 2 (RCV2);[4] a corpus of news stories by local reporters in 13 languages. We sampled approximately 19000 sentences from each of the BNC and OANC, which we used as training data to generate ERG lextype features (Section 2.3.2) for British English (en-GB) and American English (en-US), respectively. From RCV2 we

[3] http://www.debian.org

[4] http://trec.nist.gov/data/reuters/reuters.html

	bs	hr	sr	pt-PT	pt-BR	id	my	cz	sk	es-ES	es-AR	en-US	en-GB
BNC													✓
debian	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
ECB				✓				✓	✓	✓			
EMEA				✓				✓	✓	✓			
EUconst				✓				✓	✓	✓			
Europarl				✓				✓	✓	✓			
hrenWaC		✓											
KDE4		✓	✓	✓	✓	✓	✓	✓	✓	✓			
KDEdoc				✓	✓				✓	✓			
OANC												✓	
OpenSubtitles	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
RCV2										✓	✓		
SETIMES2	✓	✓	✓										
Tatoeba	✓							✓					

Table 3: Training data compiled from existing corpora.

used the Latin American Spanish news stories as a proxy for Argentine Spanish (es-AR). Note that, for a given text source, we didn't necessarily use data for all available languages. For example, `debian` contains British English and American English translations, which we did not use.

2.4.2 Web Corpus Construction

Each existing corpus we describe in Section 2.4.1 provides incomplete coverage over the set of languages in the shared task dataset. In order to have a resource that covers all the languages in the shared task drawn from a single source, we constructed web corpora for each language. Our approach was strongly inspired by the approach used to create ukWaC (Ferraresi et al., 2008), and the creation of each sub-language's corpus involved crawling the top level domains of the primary countries associated with those sub-languages. Based on the findings of Cook and Hirst (2012), the assumption underlying this approach is that text found in the top-level domains (TLDs) of those countries will primarily be of the sub-language dominant in that country. For instance, we assume that Portuguese text found when crawling the `.pt` TLD will primarily be European Portuguese, while the Portuguese found in `.br` will be primarily Brazilian Portuguese.

The process of creating a corpus for each sub-language involved translating a sample of 200 of the original ukWaC queries into each language using Panlex (Baldwin et al., 2010).[5] These queries were then submitted to the Bing Search API using the BootCaT tools (Baroni and Bernardini, 2004), constraining results to the relevant TLD. For each query, we took the first 10 URLs yielded by Bing and appended them to a list of seed URLs for that language. After deduplication, the seed URLs were then fed to a Heritrix 3.1.1[6] instance with default settings other than constraining the crawled content to the relevant TLD.

Corpora were then created from the data gathered by Heritrix. Following the ukWaC approach, only documents with a MIME type of HTML and size between 5k and 200k bytes were used. Justext (Pomikálek, 2011) was used to extract text from the selected documents. `langid.py` (Lui and Baldwin, 2012) was then used to discard documents whose text was not in the relevant language or language group. The corpus was then refined through deduplication. First, near-deduplication was done at the paragraph level using Onion (Pomikálek, 2011) with its default settings. Then, exact-match sentence-level deduplication, ignoring whitespace and case, was applied.

3 Results and Discussion

Table 4 summarizes the runs submitted by team UniMelb NLP to the VarDial DSL shared task. We submitted the maximum number of runs allowed, i.e. 3 closed runs and 3 open runs, to both the "general" Groups A–E subtask as well as the English-specific Group F subtask. We applied different methods to Group F, as some of the tools (the ERG) and resources (BNC/OANC) were specific to English. For clarity in discussion, we have labeled each of our runs according to a 3-letter code: the first letter indicates the

[5] A sample of the queries was used because of time and resource limitations.

[6] `https://webarchive.jira.com/wiki/display/Heritrix`

Run	Description	Macro-avg F-Score	
		dev	tst
	Grp A-E closed		
AC1	langid.py 13-way	0.822	0.817
AC2	langid.py per-group	0.923	0.918
AC3	POS features	0.683	0.671
	Grp F closed		
FC1	Lextype features	0.559	0.415
FC2	langid.py per-group	0.548	0.403
FC3	POS features	0.545	0.435
	Grp A-E open		
AO1	Ext Corpora (word-level model)	0.705	0.703
AO2	Web Corpora (word-level model)	0.771	0.767
AO3	5-way voting	0.881	0.878
	Grp F open		
FO1	Lextype features using BNC/OANC training data	0.491	0.572
FO2	Web Corpora (word-level model)	0.490	0.581
FO3	5-way voting	0.574	0.442

Table 4: Summary of the official runs submitted by UniMelbNLP. "dev" indicates scores from our internal testing on the development partition of the dataset.

subtask (A for Groups A–E, F for Group F), the second indicates Closed ("C") or Open ("O"), and the final digit indicates the run number.

AC1 represents a benchmark result based on the LangID system (Lui and Baldwin, 2012). We used the training tools provided with langid.py to generate a new model using the training data provided by the shared task organizers, noting that as only data from a single source is used, we are not able to fully exploit the cross-domain feature selection (Lui and Baldwin, 2011) implemented by langid.py. The macro-averaged F-score across groups is substantially lower than that on standard LangID datasets (Lui and Baldwin, 2012).

AC2 and FC2 are a straightforward implementation of hierarchical LangID (Section 2.2), using mostly-default settings of langid.py. A 6-way group-level classifier is trained, and well as 6 different per-group classifiers. We increase the number of features selected per class (i.e. group or language) to 500 from the default of 300, to compensate for the smaller number of classes (langid.py off-the-shelf supports 97 languages). In our internal testing on the provided development data, the group-level classifier achieved 100% accuracy in classifying sentences at the group level, essentially reducing the problem to within-group disambiguation. Despite being one of the simplest approaches, overall this was our best-performing submission for Groups A–E. It also represents a substantial improvement on AC1, further emphasizing the need to implement hierarchical LangID in order to attain high accuracy in discriminating similar languages.

AC3 and FC3 are based solely on POS-tag sequences generated by UMPOS, and implement a hierarchical LangID approach similar to AC2/FC2. Each sentence in the training data is mapped to a POS-tag sequence in the 12-tag universal tagset, using the per-group POS tagger for the language group. Each tag was represented using a single character, allowing us to make use of langid.py to train 6 per-group classifiers based on n-grams of POS-tags. We used n-grams of order 1–6, and selected 5000 top-ranked sequences per-language. To classify test data, the same group-level classifier used in AC2 was used to map sentences to language groups, and then the per-group POS tagger was applied to derive the corresponding stream of POS tags for each sentence. The corresponding per-group classifier trained on POS tag sequences was then applied to produce the final label for the sentence. For Groups A–E, we find that

	bs	hr	sr		id	my		cz	sk
T	53.0	23.2	60.0	VHN	0.9	1.3	.1.1	1.0	1.2
TV	32.4	13.2	43.3	DHN	0.1	0.1	1.1.	2.0	2.2
NT	31.5	13.9	43.2	N.1.	12.1	3.1	1.1	4.0	4.4
TVN	24.8	9.8	34.3	N.N	63.3	48.0	.N.1	0.5	0.7
VT	19.4	6.1	27.1	.DNV	1.8	1.1	.C	39.0	33.5
TN	29.1	10.9	29.6	DH	1.7	2.1	.1..	0.7	1.0
NTV	18.6	8.4	29.4	N.DN	3.2	2.0	.P	51.2	41.8
TVNN	16.8	6.9	23.7	VH	11.3	14.9	1.	14.0	13.9
NVT	11.2	2.9	15.5	PNV1	0.5	0.4	1..	1.2	1.6
VTV	11.0	3.2	17.0	.1.	13.2	3.8	.R	44.0	30.0

	pt-BR	pt-PT		es-AR	es-ES		en-GB	en-US
X	3.4	2.8	..	22.6	43.3	NNN	48.2	43.2
N.NN	22.2	15.3	N..	16.4	31.7	HV	41.5	46.4
.NN	29.9	22.9	.P	52.2	68.3	NN	86.3	83.0
XN	0.4	0.4	P.	6.6	16.8	H	61.8	65.9
NNNN	6.2	3.2	D.	4.4	12.6	R	61.5	65.5
D	99.2	99.5	..$	0.0	0.0	RR	7.2	9.4
NNN	28.3	18.6	J..	5.0	12.6	NNNN	21.7	18.5
.NNN	6.7	4.0	..VV	0.9	5.2	.C	15.8	18.8
N.D	58.6	47.8	DN..	4.2	11.0	...	0.8	0.3
NX	0.8	0.5	.PD	24.5	36.3	N.C	11.3	13.6

Table 5: Top 10 POS features per-group by Information Gain, along with percentage of sentences in each language in which the feature appears. The notation used is as follows: . = punctuation, J = adjective, P = pronoun, R = adverb, C = conjunction, D = determiner/article, N = noun, 1 = numeral, H = pronoun, T = particle, V = verb, and X = others

the POS-tag sequence features are not as effective as the character *n*-grams used in AC2. Nonetheless, the results attained are above baseline, indicating that there are systematic differences between languages in each group that can be captured by an unsupervised approach to POS-tagging using a coarse-grained tagset. This extends the similar observation made by Lui and Cook (2013) on varieties of English, showing that the same is true for the other language groups in this shared task. Also of interest is the higher accuracy attained by the POS-tag features on Groups A–E (i.e. AC3) than on English (Group F, FC3). The top-10 sequences per-group are presented in Table 5, where it can be seen that the sequences are often slightly more common in one language in the group than the other language(s). One limitation of the Information Gain based feature selection used in `langid.py` is that each feature is scored independently, and each language receives a binarized score. This can be seen in the features selected for Group A, where all the top-10 features selected involve particles (labelled `T`). Overall, this indicates that Croatian (hr) appears to use particles much less frequently than Serbian (sr) or Bosnian (bs), which is an intriguing finding. However, most of the top-10 features are redundant in that they all convey very similar information.

Similar to FC3, a hierarchical LangID approach is used in FC1, in conjunction with per-group classifiers based on a sequence of tags derived from the original sentence. The difference between the taggers used for FC3 and FC1 is that the FC3 tagger utilizes the 12-tag universal tagset, whereas the FC1 tagger uses the English-specific lexical types from the ERG (Section 2.3.2), a set of approximately 1000 tags. There is hence a trade-off to be made between the degree of distinction between tags, and the relative sparsity of the data — having a larger tagset means that any given sequence of tags is proportionally less likely to occur. On the basis of the results of FC1 and FC3 on the `dev` data, the lexical type features marginally outperform the coarse-grained universal tagset. However, this result is made harder to interpret by the mismatch between the `dev` and `tst` partitions of the shared task dataset. We will discuss this issue in more detail below, in the context of examining the results on Group F for the open category.

In the open category, we focused primarily on the effect of using different sources of training data. AO1 and AO2 both implement a hierarchical LangID approach, again using the group-level classifier from AC2. For the per-group classifiers, runs AO1 and AO2 use a naive Bayes model on a word-level representation, with feature selection by Information Gain. The difference between the two is that A01 uses samples from existing text corpora (Section 2.4.1), whereas A02 uses web corpora that we prepared specifically for this shared task (Section 2.4.2). In terms of accuracy, both types of corpora perform

substantially better than baseline, indicating that at the word level, there are differences between the language varieties that are consistent across the different corpus types. This result is complementary to Cook and Hirst (2012), who found that web corpora from specific top-level domains were representative of national varieties of English. AO2 (web corpora) outperforms AO1 (existing corpora), further highlighting the relevance of web corpora as a source of training data for discriminating similar languages. However, our models trained on external data were not able to outperform the models trained on the official training data for Groups A–E. A03 consists of a 5-way majority vote between results AC1, AC2, AC3, AO1 and AO2. Including the predictions from the closed submissions substantially improves the result with respect to AO1/AO2, but overall our best result for Groups A–E was obtained by run AC2.

For Group F, FO1 utilizes ERG lexical type features in the same manner as FC1, the difference being that FC1 uses the shared task `trn` partition, whereas FO1 uses sentences sampled from existing corpora, specifically `BNC` for en-GB and `OANC` for en-US. FO2 implements the same concept as AO2, namely a word-level naive Bayes model trained using web corpora. For the Group F (i.e. English) subtask, this was our best-performing submission overall. FO3 is a 5-way vote between FC1, FC2, FC3, FO1 and FO2, similar to AO3. Notably, our Group F submissions based on the supplied training data all performed substantially better on the `dev` partition of the shared task dataset than on the `tst` partition. The inverse is true for our submissions based on external corpora, where all our entries performed substantially better on the `tst` partition than on the `dev` partition. Furthermore, the differences are fairly large, particularly since Group F is a binary classification task with a 50% baseline. This implies that, at least under our models, the en-GB portion of the `trn` partition is a better model of the en-US portion of the `tst` partition than the en-GB portion thereof. This is likely due to the manual intervention that was only carried out on the test portion of the dataset (Zampieri et al., 2014).

Our Group F results appear to be inferior to previous work on discriminating English varieties (Lui and Cook, 2013). However, there are a number of differences that make it difficult to compare the results: Lui and Cook (2013) studied differences between Australian, British and Canadian English, whereas the shared task focused on differences between British and American English. Lui and Cook (2013) also draw on training data from a variety of domains (national corpora, web corpora and Twitter messages), whereas the shared task used a dataset collected from newspaper texts (Tan et al., 2014). Consistent with Cook and Hirst (2012) and Lui and Cook (2013), we found that web corpora appear to be representative of national varieties, and consistent with Lui and Cook (2013) we found that de-lexicalized representations of text are able to provide better than baseline discrimination between national varieties. Overall, these results highlight the need for further research into discriminating between varieties of English.

4 Conclusion

Discriminating between similar languages is an interesting sub-problem in language identification, and the DSL shared task at VarDial has given us an opportunity to examine possible solutions in greater detail. Our most successful methods implement straightforward hierarchical LangID, firstly identifying the language group that a sentence belongs to, before identifying the specific language. We examined a number of text representations for the per-group language identifiers, including a standard representation for language identification based on language-indicative byte sequences, as well as with de-lexicalized text representations. We found that the performance of de-lexicalized representations was above baseline, however we were not able to fully investigate approaches to integrating predictions from lexicalized and de-lexicalized text representations due to time constraints. We also found that when using external corpora, web corpora constructed by scraping per-country top-level domains performed as well as (if not better than) data collected from existing text corpora, supporting the hypothesis that web corpora are representative of national varieties of respective languages. Overall, our best result was obtained by applying two-level hierarchical LangID, firstly identifying the language group that a sentence belongs to, and then disambiguating within each group. Our best result was achieved by applying an existing LangID method (Lui and Baldwin, 2012) to both the group-level and the per-group classification tasks.

Acknowledgments

The authors wish to thank Li Wang, Rebecca Dridan and Bahar Salehi for their kind assistance with this research. 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

Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 229–237, Los Angeles, USA.

Timothy Baldwin, Jonathan Pool, and Susan M Colowick. 2010. Panlex and lextract: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 37–40, Beijing, China.

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)*.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of the Third Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, USA.

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.

Sascha Diwersy, Stefan Evert, and Stella Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*. De Gruyter, Berlin.

Rebecca Dridan. 2013. Ubertagging. Joint segmentation and supertagging for English. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1201–1212, Seattle, USA.

Ted Dunning. 1994. Statistical identification of language. Technical Report MCCS 940-273, Computing Research Laboratory, New Mexico State University.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. 2013. Simpler unsupervised POS tagging with bilingual projections. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 634–639.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop: Can we beat Google*, pages 47–54, Marrakech, Morocco.

Dan Flickinger. 2002. On building a more efficient grammar by exploiting types. In Stephan Oepen, Dan Flickinger, Jun'ichi Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering*. CSLI Publications, Stanford, USA.

Gregory Grefenstette. 1995. Comparing two language identification schemes. In *Proceedings of Analisi Statistica dei Dati Testuali (JADT)*, pages 263–268, Rome, Italy.

Nancy Ide and Catherine Macleod. 2001. The American National Corpus: A standardized resource of American English. In *Proceedings of Corpus Linguistics 2001*, pages 274–280, Lancaster, UK.

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand.

Nikola Ljubešić, Nives Mikelić, and Damir Boras. 2007. Language identification : how to distinguish similar languages ? In *29th International Conference on Information Technology Interfaces*, pages 541–546.

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.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop 2013*, pages 5–15, Brisbane, Australia.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Carl Pollard and Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago, USA.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis, Masaryk University.

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

Bali Ranaivo-Malancon. 2006. Automatic Identification of Close Languages - Case study : Malay and Indonesian. *ECTI Transaction on Computer and Information Technology*, 2(2):126–134.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the Conference on New Methods in Natural Language Processing*, Manchester, 1994.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU Bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland.

Liling Tan, Marcos Zampieri, Nikola Ljubešić, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The DSL corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, Reykjavik, Iceland.

W. J. Teahan. 2000. Text Classification and Segmentation Using Minimum Cross-Entropy. In *Proceedings the 6th International Conference "Recherche dInformation Assistee par Ordinateur" (RIAO00)*, pages 943–961, Paris, France.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 2619–2634, Mumbai, India.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Istanbul, Turkey.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03)*, pages 173–180, Edmonton, Canada.

Peter Trudgill and Jean Hannah. 2008. *International English: A guide to varieties of Standard English*. Hodder Education, London, UK, 5th edition.

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.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS 2012*, pages 233–237, Vienna, Austria.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN 2013*, pages 580–587, Sable d'Olonne, France.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, Dublin, Ireland.