Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection

Liling Tan[1], Marcos Zampieri[1], Nikola Ljubešić[2], Jörg Tiedemann[3]

Saarland University, Germany[1]
University of Zagreb, Croatia[2]
University of Uppsala, Sweden[3]
liling.tan@uni-saarland.de, marcos.zampieri@uni-saarland.de
nikola.ljubesic@ffzg.hr, jorg.tiedemann@lingfil.uu.se

Abstract

This paper presents the compilation of the DSL corpus collection created for the DSL (Discriminating Similar Languages) shared task to be held at the VarDial workshop at COLING 2014. The DSL corpus collection were merged from three comparable corpora to provide a suitable dataset for automatic classification to discriminate similar languages and language varieties. Along with the description of the DSL corpus collection we also present results of baseline discrimination experiments reporting performance of up to 87.4% accuracy.

Keywords: language identification, language discrimination, comparable corpus, similar languages, language varieties

1. Introduction

The interest in building language resources for similar languages, dialects and varieties (*SimDiVa*) has been growing significantly in the past few years. Along with these resources, we have recently seen a substantial growth in studies creating NLP tools to process and analyse *SimDiVa*; for instance, adapting character and word-level models for machine translation between similar languages (Nakov and Tiedemann, 2012), lexicon extraction from comparable corpora for closely related languages (Fišer and Ljubešić, 2011), identification of lexical variation between language varieties (Piersman et al., 2010) and automatically extracting comparable lexical and syntactic differences between language varieties (Anstein, 2013).

Along with recently published studies, the growth of interest in varieties and dialects within the NLP community is evidenced by recent events held at international NLP conferences such as the DIALECTS workshop[1] at the 2011 edition of EMNLP and '*Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*' held at the latest RANLP2013 in Bulgaria[2].

In like manner, forthcoming workshops such as LT4CloseLang[3] at EMNLP 2014 and the VarDial workshop at COLING 2014 express the same interest in SimDiVa. The VarDial workshop will host the *Discriminating Similar Language* (DSL) shared task which uses the corpus collection that this paper describes.

1.1. DSL Shared Task

Within the scope of the DSL shared task and also the VarDial workshop, we do not make a distinction between similar languages, dialects and language varieties and we aim to discuss them collectively.

From a computational perspective, language processing and tools adaptation for SimDiVa is analogous; the task of adapting tools to process similar languages (e.g. Croatian and Serbian) is not unlike adapting tools for dialects/language varieties (e.g. Dutch and Flemish; Brazilian and European Portuguese).

The DSL shared task aims at discriminating similar languages and language varieties. We treated similar languages and varieties as classes and grouped by similarity (see section 2). Similar shared tasks have dealt with language identification or discrimination for a specific language/variety group and generic language identification evaluation. For instance, the DEFT2010 attempted to discriminate the country of origin of French texts (e.g. Belgium, France, Quebec, Switzland, etc.) (Grouin et al., 2010) and the Multilingual Language Identification (MLI) shared task focusing on general purpose language identification rather than on similar languages or language varieties (Baldwin and Lui, 2010b). The main motivation of the DSL shared task is to provide a non-partisan platform for comparing classification systems using the same dataset.

For the purpose of the shared task we had to collect datasets for training, development and testing. There was no corpus compiled specifically for the purpose of discriminating similar languages or language varieties. However, there were existing corpora that held data for various languages/varieties of interest to the DSL shared task. Short of collecting data to build a new corpus, we collected corpus subsets from various corpora to build the DSL corpus collection.

To ensure that the systems participating in the shared task were actually distinguishing classes (languages or varieties) rather than text types or genres, we opted for comparable journalistic texts as this is the most common text type that has been used for previous studies on similar language discrimination (as evidenced in 1.2.). Beyond the DSL shared task, the DSL corpus collection is a useful resource for fu-

[1] http://www.ofai.at/ dialects2011/

[2] http://c-phil.informatik.uni-hamburg.de/view/Main/RANLPLangVar2013

[3] http://www.c-phil.uni-hamburg.de/view/Main/LTforCloseLang2014

ture experiments in language identification/discrimination.

1.2. Identifying Similar Languages and Varieties

Distinguishing similar languages is an obstacle in language identification. The DSL shared task aims to fill this gap by providing a dataset for researchers to test their systems in different language groups containing closely related languages or varieties. This aspect of language identification received more attention from the NLP community in the last few years.

One of the first studies to explore this issue is the by Ljubešić et al. (2007). This study proposes a computational model for the identification of Croatian texts in comparison to other closely related South Slavic languages. The study reports 99% recall and precision in three processing stages. One of these processing stages, includes a list of forbidden words, a 'black list', that appear only in Croatian texts. Tiedemann and Ljubešić (2012) improve this method and apply it to Bosnian, Serbian and Croatian texts. The study reports significantly higher performance than the accuracy of general-purpose methods, such as *TextCat* (Cavnar and Trenkle, 1994) and *langid.py* (Lui and Baldwin, 2012). Bosnian, Serbian and Croatian datasets provided are included in the DSL corpus collection as group A.

Another study presents a semi-supervised character-based model to distinguish between Indonesian and Malay (Ranaivo-Malancon, 2006), two closely related languages from the Austronesian family also represented in our dataset. The study uses different features such as the frequency and rank of character trigrams extracted from the most frequent words in each language, lists of exclusive words in each of the classes, and the format of numbers (Malay uses decimal point and Indonesian uses comma). The authors compare the performance obtained by their approach with the one obtained by *TextCat*. From the previously mentioned DEFT 2010 shared task, Mohkov (2010) proposes a classification method based on the MARF framework.

One of the methods proposed to identify language varieties is by Huang and Lee (2008). This study presented a bag-of-words approach to distinguish Chinese texts from the mainland and Taiwan. Authors report results of up to 92% accuracy. Another study is the one presented by Zampieri and Gebre (2012) for Portuguese. In this study, the authors proposed a log-likelihood estimation to identify two varieties of Portuguese (Brazilian and European). Their approach was trained and tested using journalistic texts with accuracy results above 99.5% for character n-grams. The algorithm was later adapted to classify Spanish texts using not only the classical word and character n-grams but also POS and morphology information (Zampieri et al., 2013).

The most recent experiments, to our knowledge, aim to distinguish between Australian, Canadian and British English (Lui and Cook, 2013). This study investigates the performance of classifier across different domains and the results obtained suggest that the characteristics of each variety are consistent across them. Portuguese, Spanish and English are also represented in the DSL dataset with two varieties for each language.

2. DSL Corpus Collection

The availability of adequate language resources has been a bottleneck for most language technology applications. Reusing and merging existing resources is not altogether unknown (Pustejovsky et al., 2005; Silvia et al., 2011; Eckle-Kohler and Gurevych, 2012). Since there was no existing resources specifically designed for discriminating similar languages or language varieties, we merged different corpora subsets for the purpose of the DSL shared task. The DSL corpus collection comprises news data from various corpora to emulate the diverse news content across different languages, viz. SETimes Corpus[4] (Ljubešić, 2011; Tyers and Alperen, 2010), HC Corpora (Christensen, 2014) and Leipzig Corpora Collection (Biemann et al., 2007).

2.1. Corpora Cleaning

Although the source corpora for the DSL corpora used a standardized Unicode encoding (UTF-8), the web-crawled nature of news texts from Leipzig Corpora Collection and HC Corpora contains various (X)HTML markups (e.g. `&mdash;` and `&rsquo;`) and control-characters (e.g. `U+0091` to `U+009F`), that requires cleaning prior to data usage for the DSL task. The HTMLParser[5] was used to resolve the (X)HTML markups and a python code snippet[6] was used to replace control characters with a null string.

Group	Language/Variety	Code
A	Bosnian	*bs*
	Croatian	*hr*
	Serbian	*sr*
B	Indonesian	*id*
	Malay	*my*
C	Czech	*cz*
	Slovak	*sk*
D	Brazilian Portuguese	*pt-BR*
	European Portuguese	*pt-PT*
E	Argentine Spanish	*es-AR*
	Castilian Spanish	*es-ES*
F	British English	*en-GB*
	American English	*en-US*

Table 1: Closely Related Language/Language Variety Groups

2.2. Size, Format and Representation

For each language/variety, the DSL corpus collection contains 18,000 randomly sampled training sentences, 2,000 development sentences and 1,000 test sentences; each sentence contains at least 20 tokens. We note that our naive notion of "tokens" here refer to orthographic units delimited by white spaces and this is not necessarily scalable to disambiguate language/variety groups that do not overtly mark word boundaries such as Chinese *vs* Cantonese. But for the purpose of the shared task, tokenization at codepoint

[4] published in OPUS (Tiedemann, 2012)

[5] www.docs.python.org/2/library/htmlparser.html

[6] www.pastebin.com/1aR1ivaR

is sufficient because (i) the datasets are of a single encoding and (ii) all languages involved use white spaces in their orthography.

These sentences were randomly selected from the corpora collections for each language/variety, the dataset compiled can be treated as a balanced comparable corpora sample of the news domain. To distinguish between the languages we refer to them by the language code using ISO 639-1 convention[7] and for language varieties, we use a common convention in localization, where the country code is appended to the ISO code, e.g. *en-GB* refers to the British variety of English.

The DSL Corpus Collection are in tab delimited format; the first column presents a sentence in the language/variety, the second column states its group and the last column refers to its language code. Table 1 summarizes the language/variety groups and their respective sources.

3. Baseline Discrimination Experiment

Using all 234,000 sentences of the training dataset, we trained the Naive Bayes classification models with character and word ngrams features to discriminate between the datasets. And we report the accuracy of the baseline system on the 13,000 test sentences (1000 from each language/variety).

3.1. Models

We used a lightweight Naive Bayes classification model that was previously described in language identification studies (Baldwin and Lui, 2010a; Zampieri and Gebre, 2012; Tiedemann and Ljubešić, 2012). Naive Bayes is a popular classification model due to its robustness and speed. The language of test document *D* is predicted by maximizing the sum of the logarithmic probability of a feature (i.e. word/character ngrams frequency) *w* given a language *l*:

$$\hat{l}(D) = \underset{l_i \epsilon L}{argmax} \sum_{j=1}^{|V|} \log P(w_j|l_i) \quad (1)$$

where L is the set of languages/varieties in each language group, *N* is the frequency of the *j*th word/character ngram in *D* and *V* is the set of all word/character ngrams in the training data. We use the `sklearn` implementation of multinomial Naive Bayes in our experiments[8] (Kibriya et al., 2004), which calculates:

$$P(w|l_i) = \frac{\sum_{k=1}^{|\delta|} N_{k,w} + \alpha}{|V| + \sum_{j=1}^{|V|} \sum_{k=1}^{|\delta|} N_{k,w_j}} \quad (2)$$

where δ is the set of features from the test document D and α is the smoothing factor; setting α=1 results in Laplace smoothing and α<1 for Lidstone smoothing. We used Laplace smoothing for our experiments.

[7] http://www.loc.gov/standards/iso639-2/php/English_list.php

[8] www.scikit-learn.org

3.2. Preliminary Results

The best results obtained in our baseline experiments reported 87.4% accuracy when training on character 5grams features (Table 2) and 87.1% when training on word unigrams features (Table 3).

Character Ngrams	Accuracy
2grams	0.763
3grams	0.837
4grams	0.867
5grams	**0.874**
6grams	0.873

Table 2: Discrimination Results with Character Ngrams Features

Word Ngrams	Accuracy
unigrams	**0.871**
bigrams	0.841
trigrams	0.736
uni+bigrams	0.857

Table 3: Discrimination Results with Word Ngrams features

As a sanity check, we selected a subset of the training data (108,000 sentences) and the testing data (6000 sentences) from the first language of each language/variety group (i.e. *bs, id, cz, pt-BR, es-AR, en-GB*) and ran the same Naive Bayes classification training on the subset and we achieved **99.97%** accuracy (5998 out of 6000 instances) with only character 5grams feature. The contrasting increase in accuracy without the need for discrimination of similar languages reiterates the need for language identification tools to incorporate devices to discriminate similar languages.

	Precision	Recall	F-score
bs	0.908	0.915	0.911
hr	0.957	0.944	0.950
sr	0.947	0.954	0.950
id	0.993	0.994	0.993
my	0.995	0.993	0.994
cz	1.000	1.000	1.000
sk	1.000	1.000	1.000
pt-BR	0.934	0.944	0.939
pt-PT	0.943	0.934	0.938
es-AR	0.927	0.744	0.825
es-ES	0.787	0.941	0.857
en-GB	0.600	0.602	0.601
en-US	0.600	0.598	0.598
Overall	0.889	0.889	0.889

Table 4: Precision, Recall and F-score of best performing system

Table 4 reports the precision, recall and f-score of the the 5-gram classifier for the individual languages/varieties. In the following section, we provide a brief error analysis on the preliminary results from the best performing baseline system.

	bs	*hr*	*sr*	*id*	*my*	*cz*	*sk*	*pt-BR*	*pt-PT*	*es-AR*	*es-ES*	*en-GB*	*en-US*
bs	**915**	35	50										
hr	53	**944**	3										
sr	39	7	**954**										
id				**994**	5							1	
my				7	**993**								
cz						**1000**	-						
sk						-	**1000**						
pt-BR								**944**	56				
pt-PT								66	**934**				
es-AR										**744**	255		1
es-ES										59	**941**		
en-GB												**602**	398
en-US												402	**598**

Table 5: Confusion Matrix for Character 5grams Naive Bayes Discrimination Classifier on Language varieties

3.3. Error Analysis

Table 5 presents the confusion matrix of the error analysis for the character 5gram classifier performance. The table is to be understood as such, when classifying 1000 Bosnian (*bs*) test sentences, the classifier correctly tagged 915 instances (i.e. *true positives*), wrongly tagged 35 and 50 Bosnian sentences as Croatian (*hr*) and Serbian respectively (i.e. *false negatives*) and wrongly tagged 53 Croatian sentences and 39 Serbian as Bosnian (i.e. *false positives*).

We provide a brief error analysis to emphasize the need for language discrimination among similar languages and language varieties. From the confusion matrix, the Naive Bayes classifier overfits in discriminating languages from group A and cast Bosnian features on the other two similar languages, thus resulting in high false negatives and false positives.

For group B and C, Table 4 and 5 suggest that languages that are thought to be similar are not so similar after all; Czech and Slovak (group C) though sharing the same alphabet and Slavic roots can be easily classified using the baseline system. Also, for Indonesian and Malaysian (group B), the common orthography and Austronesian origin did not hinder the performance of the baseline system[9].

Looking at the Portuguese varieties (group D), the baseline classifier performed reasonably well but it still falls behind the state-of-art accuracy (>95%) as reported in classical language identification literature (Cavnar and Trenkle, 1994; Baldwin and Lui, 2010a; Lui and Baldwin, 2012).

From Group E, the Castilian Spanish features overfits and when the classifier tagged Argentine Spanish instances, ~25% of the time, it wrongly tagged them as Castilian Spanish. Group F consisting of British (*en-GB*) and American (*en-US*) English also suffers from classification performance; ~40% of the time the classifier makes mistakes and tags an American test sentence as British and vice versa.

Prior to the DSL shared task, we might consider adding more similar languages to group B and C so as to increase the complexity of DSL task or replace the groups with other groups of similar languages (e.g. Danish and Norwegian (Bokmål) or Dutch and Flemish).

[9]Note that one Indonesian test sentence was wrongly identified as British English *en-GB* and one Argentine Spanish test sentence was wrongly identified as American English *en-US*

4. Conclusion

In this paper, we described the compilation of the DSL corpus collection for the DSL shared task. This was done through merging subsets of existing comparable corpora. Using the DSL corpus collection, we run a simple Naive Bayes discrimination system at the character and word levels to serve as baseline for the shared task. This method achieved an overall accuracy of 87.4% on the whole dataset. The task of distinguishing similar languages and varieties is by no means trivial and with this preliminary baseline results, we would like to encourage the participation of researchers and developers in the DSL shared task. The DSL corpus collection and shared task are aimed at improving the state-of-art language identification systems by tackling a known bottleneck of this task: discriminating similar languages and varieties.

The compilation of the DSL collection fills an important gap as no equivalent resource focusing on similar languages and varieties was available prior to the compilation of this collection. The resource and baseline system presented in this paper can used beyond the context of the shared task to improve/evaluate language identification systems as well as for related NLP tasks.

Acknowledgements

The authors would like to thank the original data source providers for the free and open access to their datasets and also the anonymous reviewers who provided important feedback to increase the quality of this paper.

5. References

Stefanie Anstein. 2013. *Computational approaches to the comparison of regional variety corpora : prototyping a semi-automatic system for German*. Ph.D. thesis, University of Stuttgart.

Timothy Baldwin and Marco Lui. 2010a. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237. Association for Computational Linguistics.

Timothy Baldwin and Marco Lui. 2010b. Multilingual language identification: Altw 2010 shared task data. In *Proceedings of Australasian Language Technology Association Workshop*, pages 4–7.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*.

William Cavnar and John Trenkle. 1994. N-gram-based text catogorization. *3rd Symposium on Document Analysis and Information Retrieval (SDAIR-94)*.

Hans Christensen. 2014. Hc corpora. `http://www.corpora.heliohost.org/`.

Judith Eckle-Kohler and Iryna Gurevych. 2012. Subcatlmf: Fleshing out a standardized format for subcategorization frame interoperability. In *EACL*, pages 550–560.

Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 125–131.

Cyril Grouin, Dominic Forest, Lyne Da Sylva, Patrick Paroubek, and Pierre Zweigenbaum. 2010. Présentation et résultats du défi fouille de texte deft2010 où et quand un article de presse a-t-il été écrit? *Actes du sixième DÉfi Fouille de Textes*.

Chu-ren Huang and Lung-hao Lee. 2008. Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of PACLIC 2008*, pages 404–410.

Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. 2004. Multinomial naive Bayes for text categorization revisited. In *Proc 17th Australian Joint Conference on Artificial Intelligence*, Cairns, Australia, pages 488–499. Springer.

Nikola Ljubešić, Nives Mikelic, and Damir Boras. 2007. Language identification: How to distinguish similar languages? In *Proceedings of the 29th International Conference on Information Technology Interfaces*.

Nikola Ljubešić. 2011. Setimes corpus. `http://nlp.ffzg.hr/resources/corpora/setimes/`.

Marco Lui and Tymothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Meeting of the ACL*.

Marco Lui and Paul Cook. 2013. Classifying english documents by national dialect. In *Proceedings of Australasian Language Tchnology Workshop*, pages 5–15.

Sergei Mokhov. 2010. A marf approach to deft2010. In *Proceedings of TALN2010*, Montreal, Canada.

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 301–305. Association for Computational Linguistics.

Yves Piersman, Dirk Geeraerts, and Dirk Spelman. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engieenir-ing*, 16:469–491.

James Pustejovsky, Adam Meyers, Martha Palmer, and Massimo Poesio. 2005. Merging propbank, nombank, timebank, penn discourse treebank and coreference. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 5–12. Association for Computational Linguistics.

Bali Ranaivo-Malancon. 2006. Automatic identification of close languages - case study: Malay and indonesian. *ECTI Transactions on Computer and Information Technology*, 2:126–134.

Necsulescu Silvia, Núria Bel, Muntsa Padró, Montserrat Marimón, and Eva Revilla. 2011. Towards the automatic merging of language resources.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Francis M. Tyers and Murat Alperen. 2010. Setimes: a parallel corpus of balkan languages. In *Proceedings of the multiLR workshop at LREC*.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS2012*, 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 TALN2013*, Sable d'Olonne, France.