Improving the Character Ngram Model for the DSL Task with BM25 Weighting and Less Frequently Used Feature Sets

Yves Bestgen
Centre for English Corpus Linguistics
Université catholique de Louvain
Place du cardinal Mercier, 10 B-1348 Louvain-la-Neuve Belgium
yves.bestgen@uclouvain.be

Abstract

This paper describes the system developed by the Centre for English Corpus Linguistics (CECL) to discriminating similar languages, language varieties and dialects. Based on a SVM with character and POStag n-grams as features and the BM25 weighting scheme, it achieved 92.7% accuracy in the Discriminating between Similar Languages (DSL) task, ranking first among eleven systems but with a lead over the next three teams of only 0.2%. A simpler version of the system ranked second in the German Dialect Identification (GDI) task thanks to several ad hoc postprocessing steps. Complementary analyses carried out by a cross-validation procedure suggest that the BM25 weighting scheme could be competitive in this type of tasks, at least in comparison with the sublinear TF-IDF. POStag n-grams also improved the system performance.

1 Introduction

This paper presents the participation of the Centre for English Corpus Linguistics (CECL) in the fourth edition of the VarDial Evaluation Campaign, which deals with the automatic identification of similar languages (such as excerpts of journalistic texts in Malay and Indonesian), language varieties (such as excerpts of Canadian and Hexagonal French) and dialects (such as Swiss German dialects) (Zampieri et al., 2017). The VarDial tasks share many similarities with the Native Language Identification (NLI) Task (Tetreault et al., 2013) so that several teams (Gebre et al., 2013; Goutte et al., 2013) relied on their participation in the NLI task to develop a system for VarDial. As we achieved an excellent level of performance in the NLI task (Jarvis et al., 2013), we decided to reuse the approach developed on that occasion, which was based on n-grams of characters, words and part of speech (POS) tags, and on global statistical indices such as the number of tokens per documents or the word mean length.

In the NLI task, n-grams of characters had proved to be as effective as the combination of n-grams of words and POStags. The character n-grams also obtained the best results in the 2016 Discriminating between Similar Languages (DSL) shared task (Malmasi et al., 2016) as well as in previous editions (Goutte et al., 2014; Malmasi and Dras, 2015). These performances led us to privilege this approach especially since we did not have an off-the-shelf POS-tagger for some of the languages to be discriminated. We nevertheless used POStag n-grams in addition to character n-grams for the three languages for which a version of TreeTagger is available (Schmid, 1994).

The CECL system was specifically developed for the DSL task in which it obtained the best performance (0.927) according to the weighted F1 measure, but it should be noted that its lead on the system ranked second is only 0.002. A simplified version, due to the different nature of the material to be processed, was applied to a second task, the German Dialect Identification (GDI) task which was organized for the first time. The task aim was to distinguish manually annotated speech transcripts from four Swiss German dialect areas: Basel (BS), Bern (BE), Lucerne (LU) and Zurich (ZH). This task is particularly difficult because many transcripts are very short and because it is not unusual to find in the learning material identical transcripts (e.g., *aber*) belonging to the four categories. In this task, the CECL system came second, obtaining a weighted F1 of 0.661, 0.001 less than the system ranked first.

The next section presents the main characteris-

Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 115–123,
Valencia, Spain, April 3, 2017. ©2017 Association for Computational Linguistics

tics of the system within the context of previous research. The third section describes the material of each task in which we participated and the technical characteristics of the system. The fourth section reports the results obtained on the test set, but also an evaluation of the benefits/losses brought by the various components of the system by means of a cross-validation procedure. In the conclusion, we discuss the main limits of this work and consider a few avenues for improvement.

2 System Characteristics in Relation to Previous Work

Character n-grams are the main features of the system. In the previous VarDial campaigns, a large number of systems obtained excellent performances using them (Çöltekin and Rama, 2016; Goutte et al., 2014; Zampieri et al., 2015a). Regarding the n-gram length, we choose a span of one to seven characters as in Çöltekin and Rama (2016), but the possibility to use more characters for some languages was left open.

In previous editions, named entities received much attention to such an extent that, in the 2015 edition, the documents of one of the test sets was preprocessed so as to mask them (Zampieri et al., 2015b). Their impact on performance is undoubtedly complex. On the one hand, as the material is composed of excerpts of journalistic texts, the named entities should reflect at least partially the origin of the texts. On the other hand, they could also introduce some noise since some of them can be used in any language. We decided to try to identify them (at a lower cost) so they could be processed in different ways. The solution we ended in is very similar to that used by King et al. (2014) which is based on the fact that the first letter of a named entities is usually capitalised. The goal was to determine whether performance could be improved by eliminating them. As the initial analyses refuted this hypothesis, we evaluated the opposite option, that is adding them as a supplementary feature set. The idea was that when these words are encoded in standard character n-grams, they are merged with n-grams from common words. For example, *bec* (*beak*) is included in *Québec* and in *Québecquois*.

Many previous systems developed for the DSL task also used word n-grams (Purver, 2014; Zampieri et al., 2014). We have not explored this option because there is a partial overlap between them and character n-grams. Such a situation does not occur for POStag n-grams whose usefulness has been advocated by Lui et al. (2014). There were added thus to the feature sets for each language for which we had a POS-tagger at our disposal.

The last set of features used is composed of global statistical indices similar to those employed in previous work (Bestgen, 2012; Jarvis et al., 2013). They are computed on the basis of the number of characters, spaces, uppercase letters and punctuation marks in each document.

An important characteristic of the developed system lies in the weighting function used for scaling every n-gram feature. The best performing systems in the previous VarDial editions often employed TF-IDF (see Zampieri et al. (2015a) for a detailed presentation) whose most classical formula is:

$$\text{TF-IDF} = tf \times \log \frac{N}{df} \quad (1)$$

where tf refers to the frequency of the term in the document, N is the number of documents in the set and df the number of documents that include the term. Zampieri et al. (2015a) and Çöltekin and Rama (2016) took advantage of a variant called Sublinear TF-IDF:

$$\text{(sl)TF-IDF} = (1 + \log(tf)) \times \log \frac{N}{df} \quad (2)$$

Other weighting schemes have been proposed in the literature, some of them are simpler and some more complex (Ács et al., 2015). In the NLI task, we choose the log-entropy weighting scheme often used in latent semantic analysis (Piérard and Bestgen, 2006). In Information Retrieval, the BM25 (for Best Match 25, also called *Okapi BM25*) weighting scheme is considered one of the most efficient (Manning et al., 2008) to the point that it is strongly advocated by Claveau (2012). Our first analyses having shown that BM25 surpassed log-entropy, we opted for this weighting scheme for all the n-gram based features.

BM25 is a kind of TF-IDF with specific choices for each of the two components, but above all it takes into account the length of the document. Its classic formula is (Robertson and Zaragoza,

2009):

$$\text{BM25} = \frac{tf}{tf + k_1 * (1 - b + b * \frac{dl}{dl - avg_{dl}})} \times \log \frac{N - df + 0.5}{df + 0.5} \quad (3)$$

in which

- $\frac{tf}{tf+k_1}$ is the TF component which, contrarily to the usual TF-IDF, has an asymptotic maximum tuned by the k_1 parameter.
- $(1 - b + b * \frac{dl}{dl - avg_{dl}})$, where dl is the length of the document and avg_{dl} the average length of the documents in the set, is the document length normalization factor whose impact is tuned by parameter b (and by k_1).
- The second part of the formula is a variant of the usual IDF, proposed by Robertson and Spärck Jones (Robertson and Zaragoza, 2009).

In our analyses, k_1 was set to 2 and b to 0.75 (Claveau, 2012).

3 Data and System Detailed Description

This section first describes the data provided by the organizers for each of the two tasks in which we participated and then the implementation of the various components of the system. Since the system set up for the GDI task was a simplified version of the one developed for the DSL task, the emphasis is placed on the latter.

3.1 Data

DSL Task: The organizers have made available to participants of the task a multilingual dataset (Tan et al., 2014) containing excerpts of journalistic texts in six groups of languages, each composed of two or three varieties:

- Bosnian (bs), Croatian (hr), and Serbian (sr)
- Malay (my) and Indonesian (id)
- Persian (fa-IR) and Dari (fa-AF)
- Canadian (fr-CA) and Hexagonal French (fr-FR)
- Brazilian (pt-BR) and European Portuguese (pt-PT)
- Argentine (es-AR), Peninsular (es-ES), and Peruvian Spanish (es-PE)

For each of the 14 varieties, the learning set consists of 18000 documents and development set of 2000 documents, for a total of 280000 documents.

GDI Task: The dataset for the German Dialect Identification task, described in Samardzic et al. (2016), consists of manually annotated speech transcripts from four Swiss German dialect areas: 3411 from Basel (BS), 3889 from Bern (BE), 3214 from Lucerne (LU), and 3964 from Zurich (ZH), for a total of 14478 documents. For each area, speeches were collected from several speakers and retranscribed by several annotators using a writing system designed to express the phonetic properties of different Swiss German dialects.

3.2 Detailed System Description

The extraction of all the features described below was performed by means of a series of custom SAS programs running in SAS University (freely available for research at http://www.sas.com/en_us/software/university-edition.html). To construct the predictive models during the development and test phases, we used LibSVM (with a linear kernel) (Chang and Lin, 2011), which is significantly slower than LibLINEAR developed by the same authors. This unfortunate choice prevented further optimization trials.

DSL Task: As Goutte et al. (2014), we used a hierarchical approach with a first model for discriminating language groups and then a specific model for each language group. From our point of view, this approach has two advantages. First, since distinguishing different languages (such as Persian and French) is much simpler than distinguishing language varieties (such as Canadian and Hexagonal French), the first model can be based on a reduced number of features and is thus easier to handle even though it is applied to a much larger dataset. Then, different models can be constructed for each language group in order to try to optimize their effectiveness both in selecting the sets of features and in setting the regularization parameter (C) of the SVM.

1. *Features for the identifying the language groups*: This model is based on the character n-grams of length one to four, which occurs at least 100 times in the whole dataset, weighted by means of BM25. These character n-grams were substrings of the documents

Features	bs-hr-sr	es	fa	fr	id-my	pt
CharNgram	1-7	1-7	1-8	1-7	1-7	1-7
CapCharNgram	1-7	1-7	no	1-7	1-7	1-7
POStagNgram	no	1-5	no	1-5	no	1-5
GlobStat	yes	yes	yes	yes	yes	yes
C	0.30	0.0001	0.00005	0.001	0.00005	0.00005

Table 1: Set of features and C value for the six language groups.

and include whitespace, punctuation, digits and symbols. A special character was used to signal the beginning and the end of the document.

2. *Features for the language specific models*:

   (a) *Character n-grams*: They were extracted exactly as explained above, but they contained from 1 to 7 or 8 characters.

   (b) *Capitalized word character n-grams*: Every word that starts with a capital letter was extracted from each document and the character n-grams it contains were used as supplementary features. Consideration was given to not taking into account the first word of each sentence, but since the material consisted of excerpts from newspaper articles, this criterion would have eliminated many named entities as in *Ottawa demande tout de même à la Cour suprême comment...*. This approach does not work for Persian since it does not use capital letters.

   (c) *POStag n-grams*: We used the TreeTagger (Schmid, 1994) to collect the parts of speech associated with each token in a document for each language for which a parameter file for TreeTagger (http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/) was available, that is French, Spanish and Portuguese. It might be useful to note that each of these parameter files has been built for a language group (i.e., French) and not for a language variety (i.e., Canadian French), and that they are not used to identify the language groups.

   (d) *Global statistics*: We also extracted five global statistics from each document: the proportions of capitalized letters, punctuation marks, spaces, and numerals, and the proportion of characters that are not a space, a numeral or a punctuation mark.

The feature sets and the value of the parameter C used during the test phase are given in Table 1. They were determined by means of (non-systematic) cross-validation analyses using one fifth of the data for learning and the remaining for testing.

GDI Task: Since this task was not our priority, we simply adapted the model designed for the DSL task by removing all the sets of features that were not relevant, that is the capitalized word character n-grams, the POStag n-grams, and the global statistics. Thus, only the character n-grams and the BM25 weighting scheme remain.

4 Analyses and Results

4.1 DSL Task

Performance on the Test Set: A single run was submitted because no attempt to optimize the predictions was possible due to the use of LibSVM. We got an accuracy of 92.74% and a weighted F1 of 0.9271. The system ranked first of the eleven systems that participated in the task but with a lead over the next three teams of only 0.002 (weighted F1).

Table 2 gives the confusion matrix on the test set. As can be seen, the language group of 29 documents was incorrectly identified during the first step, which corresponds to a 99.79% accuracy. It is noteworthy that a document in Persian was incorrectly categorized as in Portuguese whereas these are two totally different writing systems. It is also noted that the varieties within the different language groups did not exhibit the same level of difficulty, the triplet *bs-hr-sr* being much more difficult for the system than the others.

The Good (and the Bad) of the System: After

	bs-hr-sr			es			fa		fr		ma		pt	
	bs	**hr**	**sr**	**ar**	**es**	**pe**	**af**	**ir**	**ca**	**fr**	**id**	**my**	**br**	**pt**
bs	784	112	102	0	0	0	0	0	0	1	0	0	0	1
hr	119	865	15	0	0	0	0	0	0	1	0	0	0	0
sr	71	7	921	0	0	0	0	0	0	0	0	0	1	0
es-ar	0	2	0	855	70	68	0	0	0	0	0	0	3	2
es-es	0	2	0	69	892	34	0	0	0	1	0	0	0	2
es-pe	0	0	0	11	24	965	0	0	0	0	0	0	0	0
fa-af	0	0	0	0	0	0	968	32	0	0	0	0	0	0
fa-ir	0	0	0	0	0	0	32	967	0	0	0	0	1	0
fr-ca	0	0	0	0	0	0	0	0	956	44	0	0	0	0
fr-fr	1	0	0	0	0	0	0	0	53	946	0	0	0	0
id	0	0	0	1	1	0	0	0	0	1	980	17	0	0
my	0	0	0	0	0	1	0	0	0	0	10	989	0	0
pt-br	0	0	0	1	0	1	0	0	0	2	0	0	939	57
pt-pt	0	0	0	0	0	1	0	0	0	2	0	0	41	956

Table 2: Confusion matrix for the DSL task.

the test period (and because we remembered having used LibLINEAR in Jarvis et al. (2013)), we conducted a series of experiments on the learning and development sets to determine the gains/losses made by each component of the system. These experiments were carried out independently for each language group. Since the optimum values for the C parameter had not been determined using a cross-validation procedure, 17 values distributed between 0.000001 and 4 were tested. The results given below corresponds to the average accuracy calculated over all these C values[1]

The first experiment aimed at comparing the effectiveness of the BM25 weighting scheme with that of the sublinear TF-IDF scheme used in the best performing systems of previous years (Çöltekin and Rama, 2016; Zampieri et al., 2015a). In order to be the closest to Çöltekin and Rama (2016) system, only characters n-grams were used. Table 3 show that BM25 produced a superior accuracy in five language groups out of six, the only exception being Malay with an advantage for TF-IDF of 0.08, but it is also the language group for which performance is nearly perfect. The average benefit is 0.47%.

The second experiment used the ablation approach to assess the independent contribution of each set of features to the overall performance of the system. It consists in removing one feature of the system at a time and re-evaluating the model. The results indicated that, in each language group, the most comprehensive model (see Table 1) was always the best. Concerning the different sets of features (see Table 4):

- Deleting the global statistics reduced the accuracy in a minimal way since the difference is at most 0.02% and it is even null in three language groups out of 6.
- Deleting capitalized word character n-grams reduced accuracy by 0.09% to 0.27% depending on the language group, with an average decrease of 0.16%.
- Deleting POStag n-grams had a somewhat greater effect since the decrease is at least 0.23% and can be as high as 0.70%.

Language	**BM25**	**TF-IDF**	**Diff.**
bs-hr-sr	85.06	84.45	0.61
es	90.30	89.70	0.60
fa	96.32	95.98	0.34
fr	94.72	93.96	0.76
id-my	98.27	98.35	-0.08
pt	93.55	92.99	0.56

Table 3: Accuracy for the two weighting schemes (DSL).

4.2 GDI Task

Specificity and Performance of the Three Runs: The system for the GDI task was developed using

[1]The analyses were also performed on the maximum accuracy obtained by each model, assuming that an oracle allowed to know the ideal value of the C parameter, and produced very similar results.

Features	bs-hr-sr	es	fa	fr	id-my	pt
GlobStat	0.02	0.002	0.001	0	0	0
CapCharNgram	0.23	0.27		0.09	0.12	0.10
POStagNgram		0.70		0.23		0.30

Table 4: Benefits in accuracy for the three complementary sets of features (DSL).

Source	BE	BS	LU	ZH
Learning set	26.86	23.56	22.20	27.38
Run 1	26.25	26.36	8.74	38.65
Run 2	24.41	29.49	11.00	35.10
Run 3	23.86	25.12	23.69	27.32

Table 5: Percentage breakdown of the documents into the four categories (GDI).

a 5-fold cross-validation procedure. It led to select a model based on n-grams of 1 to 5 characters and a value of 0.0003 for the C parameter. This model was used to produce the first submitted run. It got the seventh place[2] with a weighted F1 of 0.625, close enough to the system ranked sixth but at 0.012 of the fifth place.

When taking a look at the predictions of this model during the submission period, it appeared that it attributed an unequal breakdown of the documents into the four categories, as shown in the second row of Table 5, and quite different from the breakdown in the learning set (see first row in Table 5). Even if such a distribution were possible, it does not look optimal. A few additional analyses were quickly carried out to try obtaining a more balanced breakdown.

First, we obtained from LibSVM the probability estimates of each document for each class, an option not available in LibLINEAR for SVMs. Since the solution proposed with or without probability estimation is not exactly the same for a given value of C, this solution was submitted as the second run. As shown in the third row of Table 5, the breakdown into the categories is somewhat more homogeneous. This run ranked fifth, with a weighted F1 of 0.638, almost tied with the team ranked fourth since the difference is only 0.0006 but at 0.015 from third place.

These probabilities were then used to try to equalize the headcounts in the four categories. To

[2]To determine this place, we used the ranking provided by the organizers, which only contains the highest score of each team, and inserted our different runs. The rank given therefore includes only the best run of each of the other teams.

BE	BS	LU	ZH	#	%
0	0	0	0	297	8.16
0	0	0	1	877	24.11
0	0	1	0	615	16.90
0	0	1	1	33	0.91
0	1	0	0	794	21.83
0	1	1	0	112	3.08
1	0	0	0	756	20.78
1	0	1	0	150	4.12
1	1	0	0	4	0.11

Table 6: Categorisation of the documents according to the probability estimate ranking (GDI).

do this, the 910 documents[3] with the highest probability estimate of belonging to a category were tentatively assigned to this category. Obviously, this procedure allows the classification of a document into several categories as shown in Table 6. A set of ad hoc rules was then applied to take the final decision. The most obvious one was that documents categorized into only one category were assigned to that one. Other rules apply to documents that were not assigned to any category or to documents that were assigned to two categories by giving priority to the least populated one. The last row of Table 5 confirms that this procedure made it possible to obtain a more homogeneous breakdown into the categories compared to the two other runs.

The resulting submission ranked second in the GDI task with a weighted F1 of 0.661, close to the performance of the team ranked first since the difference is only 0.0013. Thus, these simple changes in the category breakdown, only justified by the fact that one of the objectives of a shared task is to obtain the best performance, made it possible to gain 0.035 in weighted F1 and to climb from the seventh place to the second.

Benefits Brought by BM25: In order to determine whether the use of BM25 instead of sublinear TF-IDF provided a benefit, a 5-fold cross-

[3]That is a quarter of the test set. We could also have relied on the percentages in the learning set given in Table 4.

C	BM25	TF-IDF	Diff.
0.0001	82.62	80.58	2.04
0.0002	84.26	82.52	1.73
0.0003	84.68	83.22	1.46
0.0004	84.57	83.46	1.11
0.0005	84.50	83.46	1.05
0.0006	84.43	83.44	0.99
0.0007	84.26	83.42	0.85
0.0008	83.99	83.29	0.70
0.0009	83.87	83.17	0.70
0.0010	83.69	83.11	0.58

Table 7: Accuracy for the two weighting schemes (GDI).

validation procedure was used to first find the best C value for each weighting scheme and then to compare the levels of accuracy achieved. For both BM25 and sublinear TF-IDF, the optimum value of C was between 0.001 and 0.0001. Table 7 gives, for different C, the average accuracy on the 5 folds for the two weightings. As can be seen, BM25 always performed better than sublinear TF-IDF and the gain in the area where the two weightings got the best results was in the range of 1 to 1.5% accuracy. This gain may seem rather low, but it is obtained at the cost of a minimal modification of the system.

Specific Difficulties with this Task: The preceding analyses and the 2017 VarDial report (Zampieri et al., 2017) show that the performances obtained by a cross-validation procedure on the learning set (accuracy = 84%) were clearly superior to those obtained on the test set by any of the teams (maximum accuracy = 68%). This means that, although no information had been provided on this subject in the task description, the transcripts in the test set were quite different from those in the learning set.

5 Conclusion

This paper describes the system developed by the Centre for English Corpus Linguistics for participating in the fourth edition of the VarDial Evaluation Campaign (Zampieri et al., 2017). It was mainly based on characters n-grams, known for their effectiveness in this kind of task, to which less frequently used sets of features were added. These features were weighted by means of the BM25 scheme. In the two tasks we participated in, the CECL system ranked at least second. The good performance in the GDI task was due to several ad hoc adjustments of the breakdown of the test documents in the categories and cannot therefore be seen as a proof of the intrinsic superiority of the system.

The results obtained and the complementary analyses carried out by means of a cross-validation procedure suggest that the BM25 weighting scheme could be competitive in this type of tasks, at least when compared to the sublinear TF-IDF. However, it should be noted that gains were relatively small. Due to the lack of time, a detailed analysis of BM25 was not carried out to optimize the two parameters, to evaluate alternative formulas (Trotman et al., 2014) or to determine which difference between BM25 and the sub-linear TF-IDF is responsible for the performance gain.

Other options for improving the system include removing the words in English (King et al., 2014) and pre-processing the sentences entirely in capital letters. It would also be interesting to determine whether POStag n-grams could be as effective in the other languages as they were in French, Spanish and Portuguese.

Acknowledgments

This work was supported by the Fonds de la Recherche Scientifique (FRS-FNRS) under Grant J.0025.16. The author is a Research Associate of this institution. Computational ressources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI) funded by the FRS-FNRS.

References

Judit Ács, László Grad-Gyenge, and Thiago Bruno Rodrigues de Rezende Oliveira. 2015. A two-level classifier for discriminating similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 73–77, Hissar, Bulgaria.

Yves Bestgen. 2012. DEFT2009 : essais d'optimisation d'une procédure de base pour la tâche 1. In C. Grouin and D. Forest, editors, *Expérimentations et évaluations en fouille de textes: un panorama des campagnes DEFT*, pages 135–151. Hermes Lavoisier, Paris, France.

Çağri Çöltekin and Taraka Rama. 2016. Discriminating Similar Languages with Linear SVMs and Neural Networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 15–24, Osaka, Japan.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Vincent Claveau. 2012. Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, pages 85–98, Grenoble, France.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with TF-IDF weighting. In *Proceedings of the 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*, pages 216–223.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2013. Feature space selection and combination for native language identification. In *Proceedings of the 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*, pages 96–100.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The NRC system for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 139–145, Dublin, Ireland.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of The 8th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL-HLT)*, pages 111–118.

Ben King, Dragomir Radev, and Steven Abney. 2014. Experiments in sentence language identification with groups of similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 146–154, Dublin, Ireland.

Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook, and Timothy Baldwin. 2014. Exploring methods and resources for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 129–138, Dublin, Ireland.

Shervin Malmasi and Mark Dras. 2015. Language identification using classifier ensembles. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 35–43, Hissar, Bulgaria.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In *Proceedings of the 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (VarDial)*, Osaka, Japan.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *An Introduction to Information Retrieval*. Cambridge University Press.

Sophie Piérard and Yves Bestgen. 2006. Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes. *TAL: Traitement Automatique du Langage*, 47(2):89–110.

Matthew Purver. 2014. A simple baseline for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 155–160, Dublin, Ireland.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, April.

Tanja Samardzic, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob–A corpus of spoken Swiss German. In *Proceedings of the Language Resources and Evaluation (LREC)*, pages 4061–4066, Portoroz, Slovenia).

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

Liling Tan, Marcos Zampieri, Nikola Ljubešic, 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)*, pages 11–15, Reykjavik, Iceland.

Joel R. Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In Joel R. Tetreault, Jill Burstein, and Claudia Leacock, editors, *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57. The Association for Computer Linguistics.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, pages 58–65, New York, NY, USA. ACM.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 58–67, Dublin, Ireland.

Marcos Zampieri, Binyam Gebrekidan Gebre, Hernani Costa, and Josef van Genabith. 2015a. Comparing approaches to the identification of similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 66–72, Hissar, Bulgaria.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015b. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, pages 1–9, Hissar, Bulgaria.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain.