An Unsupervised Morphological Criterion for Discriminating Similar Languages

Adrien Barbaresi
Austrian Academy of Sciences (ÖAW-AC)
Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)
adrien.barbaresi@oeaw.ac.at

Abstract

In this study conducted on the occasion of the Discriminating between Similar Languages shared task, I introduce an additional decision factor focusing on the token and subtoken level. The motivation behind this submission is to test whether a morphologically-informed criterion can add linguistically relevant information to global categorization and thus improve performance. The contributions of this paper are (1) a description of the unsupervised, low-resource method; (2) an evaluation and analysis of its raw performance; and (3) an assessment of its impact within a model comprising common indicators used in language identification. I present and discuss the systems used in the task A, a 12-way language identification task comprising varieties of five main language groups. Additionally I introduce a new off-the-shelf Naive Bayes classifier using a contrastive word and subword n-gram model ("Bayesline") which outperforms the best submissions.

1 Introduction

Language identification is the task of predicting the language(s) that a given document is written in. It can be seen as a text categorization task in which documents are assigned to pre-existing categories. This research field has found renewed interest in the 1990s due to advances in statistical approaches and it has been active ever since, particularly since the methods developed have also been deemed relevant for text categorization, native language identification, authorship attribution, text-based geolocation, and dialectal studies (Lui and Cook, 2013).

As of 2014 and the first Discriminating between Similar Languages (DSL) shared task (Zampieri et al., 2014), a unified dataset (Tan et al., 2014) comprising news texts of closely-related language varieties has been used to test and benchmark systems. A second shared task took place in 2015 (Zampieri et al., 2015), an analysis of recent developments can be found in Goutte el al. (2016). The documents to be classified are quite short and may even be difficult to distinguish for humans, thus adding to the difficulty and the interest of the task.

The present study was conducted on the occasion of the third DSL shared task (Malmasi et al., 2016). It focuses on submissions to task A, a 12-way language identification task comprising varieties of five main language groups: Bosnian (bs), Croatian (hr), and Serbian (sr); Argentine (es-AR), Mexican (es-MX), and Peninsular Spanish (es-ES); Québec French (fr-CA) and Metropolitan French (fr-FR); Malay (*Bahasa Melayu*, my) and Indonesian (*Bahasa Indonesia*, id); Brazilian Portuguese (pt-BR) and European Portuguese (pt-PT). Not all varieties are to be considered equally since differences may stem from extra-linguistic factors. It is for instance assumed that Malay and Indonesian derive from a millenium-old *lingua franca*, so that shorter texts have been considered to be a problem for language identification (Bali, 2006). Besides, the Bosnian/Serbian language pair seems to be difficult to tell apart whereas Croatian distinguishes itself from the two other varieties mostly because of political motives (Ljubešić et al., 2007; Tiedemann and Ljubešić, 2012).

This work is licenced under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
pages 212–220, Osaka, Japan, December 12 2016.

The contributions of this paper are (1) a description of an unsupervised, low-resource method comprising morphological features; (2) an evaluation and analysis of its raw performance; and (3) an assessment of its impact in a model comprising common indicators used in language identification. In addition, I will demonstrate that an off-the-shelf method working on the subtoken level can outperform the best submissions in the shared task. The remainder of this paper is organized as follows: in section 2 the method is presented, a evaluation follows in section 3, the systems used for the shared task is described and a new baseline for task A is proposed in section 4.

2 Method

2.1 General principles

Statistical indicators such as character- and token-based language models have proven to be efficient on short text samples, especially character n-gram frequency profiles from length 1 to 5 (Cavnar and Trenkle, 1994). In the context of the shared task, a simple approach using n-gram features and discriminative classification achieved competitive results (Purver, 2014). Although features relying on the output of instruments may yield useful information such as POS-features used for Spanish (Zampieri et al., 2013), the diversity of the languages to classify as well as the prevalence of statistical methods call for low-resource methods that can be trained and applied easily.

Morphological features are not prominent in the literature, although the indirect word stemming performed by character n-grams is highlighted (Cavnar and Trenkle, 1994), and morphological ending frequency mentioned as future work topic (Bali, 2006). The motivation behind this submission was to test whether a morphologically-informed criterion can add linguistically relevant information to the global decision and thus improve performance. This article protocols an attempt at developing an unsupervised morphological model for each language present in the shared task. In order for this to be used in competition, it has to be learned from the training data ("closed" submission track).

The method is based on segmentation and affix analysis. The original idea behind this simple yet efficient principle seems to go back to Harris' letter successor variety which grounds on transitional probabilities to detect morpheme boundaries (Harris, 1955). The principle has proven valuable to construct stem dictionaries for document classification (Hafer and Weiss, 1974), it has been used in the past by spell-checkers (Peterson, 1980; Jones and Silverman, 1985), as it is linguistically relevant and computationally efficient. Relevant information is stored in a trie (Fredkin, 1960), a data structure allowing for prefix search and its reverse opposite in order to look for sublexicons, which greatly extends lexical coverage. Forward (prefix) and backward (suffix) tries are used in a similar fashion, albeit with different constraints. This approach does not necessarily perform evenly across languages; it has for example led to considerable progress in morphologically-rich languages such as Arabic (Ben Hamadou, 1986) or Basque (Agirre et al., 1992).

Similar approaches have been used successfully to segment words into morphemes in an unsupervised way and for several languages. A more recent implementation has been the RePortS algorithm which gained attention in the context of the PASCAL challenge (Keshava and Pitler, 2006; Dasgupta and Ng, 2007; Demberg, 2007) by outperforming most of the other systems. The present approach makes similar assumptions as the work cited and adapts the base algorithm to the task at hand, that is the identification of in- and out-of-vocabulary words and ultimately language identification. I have used this method in previous work to overcome data sparsity in the case of retro-digitized East German corpora, an under-resourced variety of written German, as I showed that it could trump full-fledged morphological analysis to predict whether a given token is to be considered as part of the language or as an error (Barbaresi, 2016a). The present experiment consists of testing if an unsupervised morphological analysis of surface forms can be useful in the context of similar language discrimination.

2.2 Current implementation

In order to build the corresponding models, a dictionary is built by observing unigrams in the training data for each language, then prefix and suffix trees are constructed using this dictionary. An affix candidate list is constituted by decomposing the tokens in the training data and the residues are added to the list if they are under a fixed length. The 5% most frequent affixes are stored and used in the identification phase, as relative corpus frequency is an efficient model construction principle (Dasgupta and Ng, 2007). Parameter tuning, that is the determination of the best result for the shared task settings, is performed empirically, in a one-against-all way with the concurrent languages. Token and affix length as well as frequency-based thresholds and blacklists have been tested. In the end, only token and affix length constraints have been used, as blacklisting in the higher or lower frequency range did not lead to noticeable improvements.

The identification algorithm aims at the decomposition into possibly known parts. It consists of two main phases: first a prefix/suffix search over respective trees in order to look for the longest possible known subwords, and secondly sanity checks to see if the rest could itself be an affix or a word out of the dictionary. If $\alpha\beta$ is a concatenation absent of the dictionary and if α and β are both identified as longest affix and in-vocabulary words, then $\alpha\beta$ is considered to be part of the target language. If one of the components is a word and if the other one is in the affix dictionary, then the token is also considered valid. The segmentation can be repeated twice if necessary, it can thus identify up to 4 components. It is performed both forward and backward since tests showed small improvements in cross-language efficiency.

For example, the token *cantalapiedra* in the Spanish corpus is not necessarily in the dictionary, but it can be decomposed into *canta+lapiedra* and ultimately into *canta+la+piedra*. The method can be robust: *especialemente* for instance can be considered to be a spelling error, but it can still be decomposed into *especial+e+mente* and qualifies as a word if the remaining *e* is in the affix list of the corresponding variety; in this particular context this is not the case, and the token is not considered to be a valid word. In the Malay/Indonesian corpus, the token *abdul_rahman* is probably a Twitter username, and its parts *abdul* and *rahman* are both in the dictionary. If punctuation signs are added to the affix list, then this token is correctly analyzed as part of the target language. On the opposite side, the token *mempertanyakan* (to put into question, to doubt) is only present in the Indonesian corpus, and the affix *memper-* is more frequent in this corpus. The model for Malay decomposes the word as *mempe+r+tanyakan*, because the word *mempe* is seen once in the training data (which stems from a spelling error: *mempe ngerusikan* should be spelled *mempengerusikan* and analyzed as *mem+pengerusi+kan*). Since *r* is in the affix list it concludes that *mempertanyakan* is a valid word. The right decomposition would have been *memper+tanyakan* or even *memper+tanya+kan*. This composite could easily be a valid Malay word but it is more frequent in Indonesian. Since *memper-* does not occur as a token, it is not decomposed correctly. Additionally, the model does not presently yield information about such frequency effects.

The models are indeed restricted to concatenative morphology, and the fact that a stem has to be in the dictionary is a strong limitation impeding performance (Demberg, 2007), in particular recall. However, it has been kept here as it prevents the models from overgenerating because of differences in the languages examined.

3 Evaluation

After empirical testing, the smallest possible token length for learning and searching is fixed to 5 characters, there is no upper bound on token length, and the maximum affix length is set to 2 to provide a safer setting overall, although affix lengths of 3 or 4 occasionally lead to better results. Despite the possibility of populating a blacklist out of common tokens present in the lower and higher frequency ranges, experiments have not been conclusive, so that no blacklisting has been used for the task.

3.1 Raw performance

Table 1 describes the results of morphological training. The coverage displayed is the total percentage of words considered to be in-dictionary by the model, for the target language and for the concurrent language(s) respectively. For Southeastern-European languages, I find a lower lexical overlap than Tiedemann and Ljubešić (2012). The Spanish varieties have the smallest coverage spread. The assumption that Malay and Indonesian feature more than 90% lexical similarity (Bali, 2006) is only partially confirmed: it seems that Indonesian has more to do with Malay than vice versa and the news samples used

Trad. assumed	**Languages**		**Coverage**		**Benchmark (precision)**		
lang. type	Target	Concurrent	Target	Other	Baseline	Method	Bayesline
Fusional	bs	hr,sr	0.88	0.84	0.70	0.71	0.81
	hr	bs,sr	0.90	0.79	0.87	0.87	0.83
	sr	bs,hr	0.90	0.76	0.92	0.92	0.86
Fusional	es-AR	es-ES,es-MX	0.96	0.89	0.85	0.86	0.79
	es-ES	es-AR,es-MX	0.95	0.92	0.69	0.69	0.58
	es-MX	es-AR,es-ES	0.93	0.92	0.66	0.65	0.78
Fusional	fr-CA	fr-FR	0.97	0.87	0.92	0.92	0.95
	fr-FR	fr-CA	0.94	0.92	0.84	0.85	0.85
Agglutinative	id	my	0.95	0.85	0.98	0.97	0.98
	my	id	0.96	0.78	0.99	0.98	0.99
Fusional	pt-BR	pt-PT	0.95	0.89	0.89	0.91	0.93
	pt-PT	pt-BR	0.95	0.89	0.92	0.93	0.93

Table 1: Results of morphological induction on training set in terms of coverage and precision of classification on the development set. The unigram baseline and unigram Bayesline (Tan et al. 2014) are given for comparison.

for the tests seem to be relatively easy to tell apart, since they feature the largest coverage spread. This distinction within the Bahasa complex and the rest is reflected as being traditionally assumed in language typology. However, finer differences do exist between fusional/inflectional languages (Dryer and Haspelmath, 2013)[1], and the results of the morphological induction phase constitute further evidence of subtle differences, among other things on the morpholexical level.

Concerning the benchmark, the method is compared to a unigram baseline in terms of raw precision: for each instance, potential candidates (alphabetic tokens of 5 characters and more) are analyzed and classified as in- or out-of-vocabulary. The number of in-vocabulary tokens is divided by the number of candidates, and the instance is classified according to the model which yields the highest proportion of recognized tokens. This proportion has to be strictly superior to all others, which means that this indicator (as all unigram models) can be undecided due to coverage problems, especially in short instances. Thus, I used precision as a benchmark in order to judge cases where the indicator actually predicts something, in other words the positive predictive value.

The precision displayed has been calculated accordingly on the development set, by using the highest score per instance and taking language families in isolation, i.e. by reducing the 12-way classification to a 2- or 3-way one. The method mostly achives equal or better results than the unigram baseline, which proves that the concept is working, and that it might lead to better predictions for unseen samples. A “Bayesline” is used as implemented for the previous DSL editions (Tan et al., 2014), it grounds on unigrams for the sake of comparison and integrates a Naive Bayes classifier[2], whereas the baseline and my method yield “raw” results at this point. In line with expectations, the Bayesline generally achieves better results. There are interesting discrepancies though: Argentine Spanish and Serbian seem to stand out in a morpholexical perspective, meaning that the method could add relevant information to a global system.

3.2 Impact in a composite system

The morphological criterion is not meant to be used by itself, but rather as a part of a combination of features which are learned and weighted by statistical learning techniques as usually done in the literature (Goutte et al., 2016). Since the criterion does not systematically lead to an unequivocal result, it will be treated as a sparse feature by the models. The question is now to determine both the impact and the

[1] http://wals.info/chapter/26

[2] `CountVectorizer(analyzer='word', ngram_range=(1,1))`, followed by a multinomial Naive Bayes classifier

Language	Morphology		Char 4-grams		Word bigrams	
	LM	RF	LM	RF	LM	RF
bs	***	**	.	*	***	*
hr	.	*		*	***	*
sr	***	*	***	*	*	.
es-AR	***	*	.	.	**	.
es-ES		**	.	**	***	*
es-MX	***	**	*	*	***	.
fr-CA	*	**	***	**	***	.
fr-FR	***	***		*	***	.
id	***	***	*	**	***	.
my	***	**	***	*	**	.
pt-BR	***	**	***	*	***	.
pt-PT	***	**	***	**	***	.

Table 2: Results of relevance tests on development set
Linear model (LM) significance levels: 0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1
Random Forest (RF) relative feature importances: > 80% “***” > 60% “**” > 40% “*” > 20% “.”

potential for generalization of the morphological criterion presented in this article, all the more since the closed training sets are restricted in size.

To test for variable significance, two distinct classification models are applied. The first one consists of a regression analysis using a linear model, from a family of models commonly used to classify data and select variables (Friedman et al., 2009), and previously used for classification of web documents in web corpus construction (Barbaresi, 2015). The second one resides in random forests (Breiman, 2001). It has been shown that in spite of their variability they can be used for interpretation, with respect to variable importance, and also for feature selection in terms of model efficiency (Genuer et al., 2010). Previous editions of the shared task have highlighted that higher-order character n-grams and lower-order word n-grams allow for an efficient combination of accuracy and efficiency (Goutte et al., 2014; Malmasi and Dras, 2015). Following from these results, character 4-grams and word bigrams are taken as a reference for relevance tests.

Table 2 shows that word bigrams are the most relevant indicator according to the linear model, the morphological criterion is used the most by the random forests. Overall, the most relevant feature is the morphological criterion, although it is not equally important across all languages (especially for 3-way concurrencies) and although the overall model is well-balanced. In fact, nearly all if not all the features tend to be used even after feature selection by both methods, which means that the criterion introduced here qualifies as relevant input from a statistical point of view and may be used as a sparse feature to discriminate similar languages.

4 Shared task systems

The systems described in this section have been submitted as team XAC (Cross-Academies). An additional Bayesline is introduced, used as a system component. It only became apparent after the release of the gold dataset that it actually performs better on it than all other features and, most importantly, better than the other competing systems.

After significance tests conducted as described above, a combination of features has been used to set up a classification system for the DSL shared task. The instances in the data are tokenized using the *SoMaJo* tokenizer (Proisl and Uhrig, 2016), which achieves state-of-the-art accuracies on both web and CMC data for German. As it is rule-based, it is deemed efficient enough for the languages of the shared task. The features used comprise instance statistics such as length or number of capital letters and most importantly the following series of continuous variables yielded by models trained for each language variety: the normalized morphological criterion (feature scaling by standardization); character and word n-grams language models perplexities on lowercase tokenized text, respectively character 5-grams with Kneser-Ney smoothing (Kneser and Ney, 1995) as computed by *OpenGrm* (Roark et al., 2012); and word 2-, 3-, and 4-grams with modified Kneser-Ney smoothing as computed by *KenLM* (Heafield, 2011; Heafield et al., 2013); the online learning toolkit *Vowpal Wabbit* (Langford et al., 2007; Langford et al., 2009), which achieved the best performance when used separately on the development set; and probabilities given by the Bayesline proposed below (as a Naive Bayes classifier yields probabilities for each category). It was not clear in the development data that this new Bayesline would perform better when applied alone on the gold set, the combination appeared to lead to the best overall performance.

Classification is performed using existing implementations by the *scikit-learn* toolkit (Pedregosa et al., 2011). Random forests (Breiman, 2001) were used in the two first runs because of encouraging results on the development set, but they were outperformed by a gradient boosting classifier (Friedman, 2001) on the test set as shown in Table 3 (run 3), probably because of the robustness of this method, which is known to perform well with heterogeneous features. The baseline is calculated according to the DSL Bayesline (Tan et al., 2014) as described above, with an adapted setting to focus on character 4-grams.[3]

The best run was ranked 8th out of 17 teams on the task A in closed training, i.e. without using external resources or past DSL data, with an accuracy of 0.879; the baseline of the first edition was 0.859 and the best ranked submission reached an accuracy of 0.893. The confusion matrix on Figure 1 hints at a lower classification accuracy concerning the three-way concurrencies, Spanish in particular. I hypothesize that statistical models reach their limits here, especially concerning the Mexican Spanish, which is both heavily influenced by other varieties and not homogeneous enough, so that frequency information cannot be used reliably. Finally, the results on the gold set are not in line with the development set, where cross-validated accuracies around 0.92 have been observed. The systems used may have been too complex or not well-weighted.

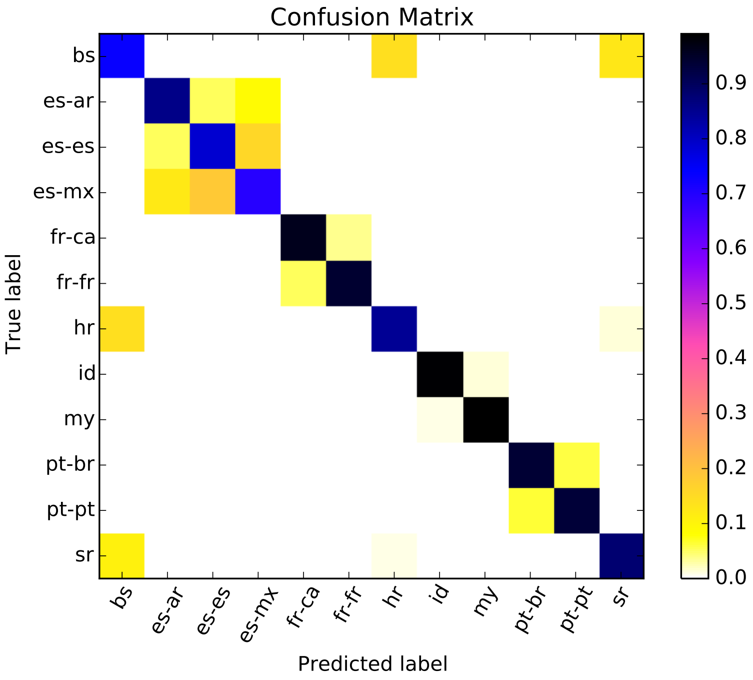

Figure 1: Confusion matrix for test set A (closed training)

In view of this I would like to introduce a refined version of the Bayesline (Tan et al., 2014) in the form of a similar off-the-shelf Naive Bayes classifier using a contrastive subword n-gram model[4], which

[3] `CountVectorizer(analyzer='char', ngram_range=(4,4))`

[4] `TfidfVectorizer(analyzer='char', ngram_range=(2,7), strip_accents=None, lowercase=True)` followed by `MultinomialNB(alpha=0.005)`, adapted from https://web.archive.org/web/2016-

outperforms the best teams for task A (with an accuracy of 0.902), even without taking the development data into consideration (accuracy of 0.898). This shows that meaningful word and subword features can give a boost to existing systems, even if they are based on simple extraction methods and/or used alone.

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
Reference Bayesline	0.859	0.859	0.858	0.858
run 1	0.861	0.861	0.860	0.860
run 2	0.870	0.870	0.869	0.869
run 3	0.879	0.879	0.879	0.879
Proposed Bayesline	0.902	0.902	0.902	0.902

Table 3: Results for test set A (closed training). Bayeslines inspired by Tan et al. (2014)

Finally, I wish to bring to the reader's attention that I tried to gather web texts for an open submission using existing techniques (Barbaresi, 2013; Barbaresi, 2016b) and focusing on top-level domains. Although the quality of corpora did not seem to be a problem apart from the Bosnian domain (.ba), the variation contained in web texts was not a good match for the news texts of the shared task. As observed in previous editions, performance decreased as further texts were included, so that no open submission was made.

5 Conclusion

I have presented a method to build an unsupervised morphological model for all the languages of the shared task. The resulting segmentation analysis is not the most efficient feature in itself, but I have shown that this criterion qualifies as relevant input from a statistical point of view and may be used as a sparse feature to discriminate similar languages. A reasonable hypothesis is that it adds new linguistically motivated information, dealing with the morpho-lexical logic of the languages to be classified, also yielding insights on linguistic typology. Unevenly distributed characteristics across the languages account for noise which is filtered accordingly by the models.

Meaningful subword features could well give a boost to existing systems, even if they are based on simple extraction methods. In fact, an off-the-shelf Naive Bayes classifier using a contrastive word and subword n-gram model outperforms the best submission for classification across 12 languages, which casts the best possible light on this topic. In this respect, future work includes a refinement of feature extraction processes on this level, especially concerning frequency, whose role in linguistically relevant units is more difficult to assess, probably because more training data is needed than for character n-grams.

The efficiency of the proposed Bayesline as well as the difficulty to reach higher scores in open training could be explained by artificial regularities in the test data. The results for the Malay/Indonesian pair are striking, this clear distinction does not reflect the known commonalities between these varieties. This seems to be an artifact of the data which feature standard language of a different nature than the continuum "on the field", that is between both countries and within Indonesia. The conflict between in-vitro and real-world language identification has already been emphasized in the past (Baldwin and Lui, 2010), it calls for the inclusion of web texts into the existing task reference.

Acknowledgements

Thanks to three anonymous reviewers for their comments, to Antonio Ruiz Tinoco (Sophia University) for his input on variation in Spanish, and to Artem Sharandin (Russian Academy of Sciences) for his help with Bahasa Melayu.

References

Eneko Agirre, Inaki Alegria, Xabier Arregi, Xabier Artola, A Díaz de Ilarraza, Montse Maritxalar, Kepa Sarasola, and Miriam Urkia. 1992. XUXEN: A spelling checker/corrector for Basque based on two-level morphology.

0403184050/http://scikit-learn.org/stable/auto_examples/text/document_classification_20newsgroups.html

In *Proceedings of the 3rd conference on Applied Natural Language Processing*, pages 119–125. Association for Computational Linguistics.

Timothy Baldwin and Marco Lui. 2010. 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.

Ranaivo-Malançon Bali. 2006. Automatic Identification of Close Languages–Case Study: Malay and Indonesian. *ECTI Transaction on Computer and Information Technology*, 2(2):126–133.

Adrien Barbaresi. 2013. Challenges in web corpus construction for low-resource languages in a post-BootCaT world. In *6th Language & Technology Conference, Less Resourced Languages special track*, pages 69–73.

Adrien Barbaresi. 2015. *Ad hoc and general-purpose corpus construction from web sources*. Ph.D. thesis, École Normale Supérieure de Lyon.

Adrien Barbaresi. 2016a. Bootstrapped OCR error detection for a less-resourced language variant. In Stefanie Dipper, Friedrich Neubarth, and Heike Zinsmeister, editors, *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 21–26. University of Bochum.

Adrien Barbaresi. 2016b. Efficient construction of metadata-enhanced web corpora. In *Proceedings of the 10th Web as Corpus Workshop*, pages 7–16. Association for Computational Linguistics.

Abdelmajid Ben Hamadou. 1986. A compression technique for Arabic dictionaries: the affix analysis. In *Proceedings of the 11th Conference on Computational Linguistics*, pages 286–288. Association for Computational Linguistics.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

William B. Cavnar and John M. Trenkle. 1994. N-Gram-Based Text Categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Sajib Dasgupta and Vincent Ng. 2007. High-performance, language-independent morphological segmentation. In *HLT-NAACL*, pages 155–163.

Vera Demberg. 2007. A language-independent Unsupervised Model for Morphological Segmentation. In *Annual Meeting of the Association for Computational Linguistics*, volume 45, pages 920–927.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Edward Fredkin. 1960. Trie Memory. *Communications of the ACM*, 3(9):490–499.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2009. *The Elements of Statistical Learning*, volume 1. Springer, 2nd edition.

Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232.

Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. 2010. Variable selection using Random Forests. *Pattern Recognition Letters*, 31(14):2225–2236.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The NRC system for discriminating similar languages. In *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 139–145.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating Similar Languages: Evaluations and Explorations. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1800–1807. European Language Resources Association (ELRA).

Margaret A. Hafer and Stephen F. Weiss. 1974. Word Segmentation by Letter Successor Varieties. *Information Storage and Retrieval*, 10:371–385.

Zellig S. Harris. 1955. From Phoneme to Morphemes. *Language*, 31(2):190–222.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Mark A. Jones and Alex Silverman. 1985. A spelling checker based on affix classes. In Jagdish C. Agrawal and Pranas Zunde, editors, *Empirical Foundations of Information and Software Science*, pages 373–379. Springer US, Boston, MA.

Samarth Keshava and Emily Pitler. 2006. A simpler, intuitive approach to morpheme induction. In *Proceedings of 2nd Pascal Challenges Workshop*, pages 31–35.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 181–184.

John Langford, Lihong Li, and Alexander L. Strehl. 2007. Vowpal wabbit (fast online learning). Technical report. http://hunch.net/∼vw/.

John Langford, Lihong Li, and Tong Zhang. 2009. Sparse Online Learning via Truncated Gradient. *Journal of Machine Learning Research*, 10(Mar):777–801.

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. IEEE.

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

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*, pages 35–43.

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)*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James L. Peterson. 1980. Computer programs for detecting and correcting spelling errors. *Communications of the ACM*, 23(12):676–687.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop*, pages 57–62. Association for Computational Linguistics.

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*, pages 155–160.

Brian Roark, Richard Sproat, Cyril Allauzen, Michael Riley, Jeffrey Sorensen, and Terry Tai. 2012. The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the ACL 2012 System Demonstrations*, pages 61–66. Association for Computational Linguistics.

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*, pages 11–15.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING*, pages 2619–2633.

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.

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*, pages 58–67.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL Shared Task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9.