Comparing two Basic Methods for Discriminating Between Similar Languages and Varieties*

Pablo Gamallo
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
Univ. of Santiago de Compostela
Galiza
pablo.gamallo@usc.es

José Ramom Pichel
Imaxin—Software,
Galiza
joseramompichel@imaxin.com

Iñaki Alegria, Manex Agirrezabal
IXA Nlp group
Univ. of the Basque Country
UPV/EHU
i.alegria@ehu.eus
manex.aguirrezabal@ehu.eus

Abstract

This article describes the systems submitted by the Citius_Ixa_Imaxin team to the Discriminating Similar Languages Shared Task 2016. The systems are based on two different strategies: classification with ranked dictionaries and Naive Bayes classifiers. The results of the evaluation show that ranking dictionaries are more sound and stable across different domains while basic bayesian models perform reasonably well on in-domain datasets, but their performance drops when they are applied on out-of-domain texts.

1 Introduction

McNamee (2005) argued that language detection is a solved problem since the performance of most systems approaches 100% accuracy. However, this can be true only if we assume that the systems are tested on relatively long and well written texts. In recent experiments, the accuracy of the language detection starts to decrease much faster with respect to relatively longer texts having at least 400 characters (Tromp and Pechenizkiy, 2011). In consequence, language detection is not a solved problem if we consider noisy short texts such as those written in social networks. Apart from the size and the written quality of input texts, it is also necessary to take into account another important factor that can hurt the performance of language detectors, namely language proximity and variety detection. Closely related languages or language varieties are more difficult to identify and separate than languages belonging to different linguistic families.

DSL Shared Task 2016 (Malmasi et al., 2016; Goutte et al., 2016) is aimed to compare language identification systems on the specific task of discriminating between similar languages or varieties. This is the third edition of the shared taks, which is divided into two sub-tasks.

First, the sub-task 1 is focused on discriminating between similar languages and national language varieties, including five different groups of related languages or language varieties:

- Bosnian, Croatian, and Serbian
- Malay and Indonesian
- Portuguese varieties: Brazil and Portugal
- Spanish varieties: Argentina, Mexico, and Spain
- French varieties: France and Canada

The objective of sub-task 2 is the identification of Arabic varieties. As Arabic is mostly written using the modern standard, the sub-task is focused on conversational speech which is divided into many different diatopical varieties. For this purpose, the DSL organizers provided a dataset containing automatic speech recognition transcripts five Arabic varieties: Egyptian, Gulf, Levantine, North-African, and Modern Standard Arabic (Malmasi et al., 2015).

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
pages 170–177, Osaka, Japan, December 12 2016.

Our team, Citius_Ixa_Imaxin, participated in all DSL sub-tasks with the following objective: to compare two very basic methods for language detection and observe how they behave when they are applied on the difficult task of discriminating between similar languages or varieties. On the one hand, we describe and evaluate a ranking approach based on small dictionaries built according to the Zipf's law, i.e. the frequency of any word is inversely proportional to its rank in the frequency table. On the other hand, we also describe and evaluate a Naive Bayes system relying on word unigrams.

2 Related Work

Two types of models have been used for language detection in general: those made of n-grams of characters (Beesley, 1988; Dunning, 1994) and those based on word unigrams or dictionaries (Grefenstette, 1995; Rehurek and Kolkus, 2009). In the latter approaches, models are dictionaries built with words ranked by their frequency in a reference corpus, and their ranking is used to compute their "relevance" in the input text. In Cavnar and Trenkle (1994), they construct a language model by making use of the ranking of the most frequent character n-grams for each language during the training phase (n-gram profiles). So, even if this is an approach based on character n-grams, it also uses the ranking strategy which is characteristic of the dictionary-based approach.

According to Rehurek (2009), very simple dictionary-based methods are better suited to work on close varieties than other more complex methods for language identification. In order to verify such a hypothesis, in the DSL shared task we will compare a dictionary-based approach with a another standard strategy based on bayesian classification.

Two former editions of DSL shared task took place in the two previous years (Zampieri et al., 2015; Zampieri et al., 2014). One of the best systems in the two previous editions makes classification in two steps: it first makes a prediction about the language group and then it selects a specific language from that language group (Goutte et al., 2014; **?**). In 2014 edition, it achieved the best performance in the closed submission task, while in 2015, it was the first system in the open task. At the last edition, the winner system in the closed submission track relies on an ensemble of SVM classifiers using features such as character n-grams from one to six n-grams (Malmasi and Dras, 2015). Notice that the two winner systems rely on complex strategies since the first one requires several steps to perform classification and the second one needs to work with several classifiers. By contrast, we propose very basic classifiers using just word unigrams (tokens) as features. One of our aims is to observe whether baseline strategies are able to be competitive in the DSL tasks.

Another related research direction has been on language identification on Twitter, giving rise to the Tweet-LID shared task (Zubiaga et al., 2014; Zubiaga et al., 2015). This competition aimed at recognizing the language of tweets written in English and in languages spoken on the Iberian peninsula such as Basque, Catalan, Spanish, Galician and Portuguese. Notice that some of these languages, namely Galician and Portuguese, are so close that they could be considered as two varieties of the same language.

3 Methodology and Data

In this section, we describe two basic strategies for language identification: a dictionary-based approach and a bayesian classifier, which also participated at TweetLID 2014 Shared Task (Gamallo et al., 2014).

3.1 Quelingua: A Dictionary-Based Approach

Quelingua[1] has been implemented using a dictionary-based method and a ranking algorithm. It is based on the observation that for each language, there is a set of words that make up a large portion of any text and their presence is to be expected as word distribution follows Zipf's law.

For each word w found in a corpus of a particular language, and for the N most frequent words in that corpus, we define its *inverse ranking* (IR) as follows:

$$IR(w) = N - (rank(w) - 1) \quad (1)$$

[1]Freely available at: `https://github.com/gamallo/QueLingua`

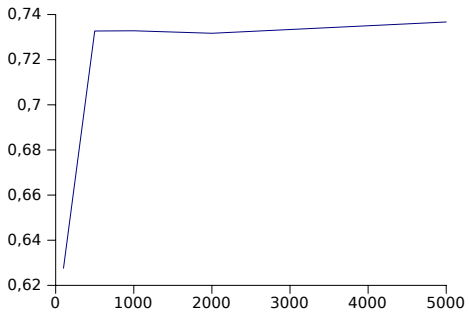

Figure 1: Growth curve of F1-Score (y axis) as a function of the dictionary size (x axis)

where $rank(w)$ is the rank of w in the dictionary of N most frequent words. For instance, if the dictionary contains 1000 words, the IR for the most frequent word (ranking 1) is 1000. Specifying the size N of the dictionary is a critical issue of the method. The final weight of a specific language $lang$ given a text is computed in equation 2, where K is the size of the input text:

$$weight(lang, text) = \sum_{i=1}^{K} IR(word_i) \tag{2}$$

This is computed for all available languages, and that with the highest weight is selected as the detected language for the input text.

In order to give more coverage to the system, we added a suffix module containing the most frequent suffixes of the target languages. For instance, "-ção" is associated to Portuguese, "-ak" to Basque, "-ción" to Spanish and Galician, etc. This information can be automatically extracted or manually added to the module. The IR of any word that is not in the dictionary but has a suffix found in the suffix module is computed as the average IR, i.e.: $N/2$. However, for the DSL task, the suffix module has not been used because we did not find any relevant suffixes allowing to discriminate between similar varieties. This module is useful for distinguish between different languages within very short texts, but it is not suited to deal with similar varieties.

We performed some preliminary experiments focused on determining the best size of the dictionary (i.e. of the language model). Figure 1 depicts the growth curve of F1-Score as a function of the size of the dictionary for one of the varieties (es-ES). It shows that the peak is achieved with a size of 1000 words. We obtained similar results for all language varieties. So, for all tracks of the DSL shared task, Quelingua was trained with a dictionary of this size.

3.2 A Naive Bayes Classifier

To compare our dictionary-based system with a state-of-the-art approach, we implemented a Naive Bayes (NB) classifier based on the system we previously created for a sentiment analysis task, and described in Gamallo (2013). According to Winkelmolen and Mascardi (2011; Vatanen et al. (2010), language detection based on NB algorithms performs well on short texts. In Vatanen (2010), a NB classifier built with character n-gram models clearly outperformed the ranking method by Cavnar and Trenkle (1994) when the tests were performed on noisy short texts.

Our NB classifier was trained with two different models: a model based on character n-grams and another one based on word unigrams (bag of words). The smoothing technique used by our classifiers for unseen features (n-grams or words) is a version of Good-Turing estimation (Gale, 1995).

We made preliminary experiments on similar languages with both character n-grams and word unigrams. Concerning character-based models, the highest scores were reached using short n-grams. This was also predicted by Winkelmolen and Mascardi (2011; Vatanen et al. (2010), who claimed that NB classifiers for language detection perform better using short n-grams, with $n < 4$. However, in our preliminary experiments the best results were achieved using word unigrams, which outperformed the best

character-based models. This is in accordance with Rehurek (2009), who tried to prove that word-based methods are more reliable than character-based models for language discrimination between similar languages/varieties. Therefore, for the tracks of DSL, we will only use word unigrams to train the bayesian classifiers.

4 Experiments

4.1 Training and Test Dataset

For the sub-task 1, the training corpus is a new version of the DSL corpus collection (DSLCC) (Tan et al., 2014). The corpus contains 20,000 instances per country, including excerpt extracted from journalistic texts. In total, the corpus contains 8.6M tokens. For the sub-task 2, the training corpus on Arabic varieties consists on over 7.5K automatic speech recognition transcripts for five varieties (**?**). In total, 331K tokens. These two corpora were used in the closed submission tracks, that is, in those tracks requiring systems to be trained with the corpus collection of the third edition. In order to participate in the open tracks, we also trained our two systems including the corpus released in the second edition of the DSL corpus collection (Zampieri et al., 2015). Given that this collection does not contain any data for Arabic dialects, we have not participated at the open submission track of sub-task 2.

4.2 Preprocessing

Before building the language models, we used a Named Entity Recognition system inspired by that described in Garcia and Gamallo (2015) in order to remove all proper names from the input texts. Even if proper names may help the system find the correct variety in many cases, we reckon that they are useful just because of extra-linguistic or cultural reasons. Proper names can prevent the classifier correctly identifying a specific national variety when the topic of the target text is a person or a location of a country with a different language variety. For this reason, we got rid of named entities before building the language models.

Test Set	Track	Method	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
A	-	Random baseline	-	-	-	0.083
A	closed	Quelingua	0.7756	0.7756	0.771	0.771
A	closed	**NB**	0.8525	0.8525	0.8502	**0.8502**
A	open	Quelingua	0.7759	0.7759	0.7707	0.7707
A	open	**NB**	0.871	0.871	0.8694	**0.8694**
B1/B2	-	Random baseline	-	-	-	0.20
B1	closed	**Quelingua**	0.708	0.708	0.4454	**0.7127**
B1	closed	NB	0.082	0.082	0.049	0.1175
B1	open	Quelingua	0.664	0.664	0.3962	0.6339
B1	open	NB	0.094	0.094	0.054	0.1296
B2	closed	**Quelingua**	0.686	0.686	0.4988	**0.6983**
B2	closed	NB	0.282	0.282	0.1244	0.2987
B2	open	Quelingua	0.692	0.692	0.4345	0.6952
B2	open	NB	0.288	0.288	0.1318	0.3164
C	-	Majority class baseline	-	-	-	0.2279
C	closed	**Quelingua**	0.387	0.387	0.3795	**0.3817**
C	closed	NB	0.3032	0.3032	0.2667	0.2664

Table 1: Results for all runs of Quelingua and NB classifiers.

4.3 Results

Table 1 shows the results obtained by our two classifiers, *Quelingua* (dictionary-based) and *NB* (naive bayes). Four test sets were used for evaluation. Tests A, B1 and B2 belong to the sub-task 1 (5 language groups of similar varieties) while test C is used for sub-task 2 (Arabic varieties). In sub-task 1, test A

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
Quelingua	0.7756	0.7756	0.771	0.771
NB	0.8525	0.8525	0.8502	0.8502

Table 2: Results for test set A (closed training).

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
Quelingua	0.7759	0.7759	0.7707	0.7707
NB	0.871	0.871	0.8694	0.8694

Table 3: Results for test set A (open training).

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
Quelingua	0.708	0.708	0.4454	0.7127
NB	0.082	0.082	0.049	0.1175

Table 4: Results for test set B1 (closed training).

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
Quelingua	0.664	0.664	0.3962	0.6339
NB	0.094	0.094	0.054	0.1296

Table 5: Results for test set B1 (open training).

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
Quelingua	0.686	0.686	0.4988	0.6983
NB	0.282	0.282	0.1244	0.2987

Table 6: Results for test set B2 (closed training).

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
Quelingua	0.692	0.692	0.4345	0.6952
NB	0.288	0.288	0.1318	0.3164

Table 7: Results for test set B2 (open training).

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
Quelingua	0.387	0.387	0.3795	0.3817
NB	0.3032	0.3032	0.2667	0.2664

Table 8: Results for test set C (closed training).

contains newspaper texts. It is thus considered as an in-domain experiment as both training and test datasets belong to the same domain. By contrast tests B (B1 and B2) consist of social media data. As the training corpus is very different from the test dataset, it can be considered as an out-of-domain test. Test set C of sub-task 2 contains automatic speech recognition translations from Arabic varieties. Even if test C belongs to the same genre (spoken language) than the training dataset, it is likely that its content will belong to new and/or different domains than the training corpus.

Concerning the baselines depicted in the table (1), it is important to point out that test set A had 12 classes (i.e. different varieties) while test sets B1 and B2 had only 5 classes. The samples were evenly distributed across the classes and so a random baseline is used. The samples in test set C were slightly unbalanced, so a majority class baseline of 22.79% is used.

As it was stated above, closed submissions use only the training corpus provided by the DSL organizers, while open submissions also use the corpus provided by the previous version of the DSL shared task.

The results of Table 1 show that the best in-domain system is NB, while Quelingua is more stable and performs better accross different domains and genres. Considering the in-domain task (test A), NB scores are not very far from the best systems. For instance, it is less than 2 points far from the top system (0.869 *vs* 0.888) in the open submission. A similar system achieved the best score in the open submission of Tweet-LID shared task 2014 (Gamallo et al., 2014). Nevertheless, the performance of NB drops dramatically in out-of-domain tests (B1 and B2). The dictionary-based approach (Quelingua) achieve similar results in both in-domain and out-domain tests. It is the eighth best system (out of 14) in both B1 and B2 tests (closed submission). Such a result is acceptable if we consider that the system is very basic and simple: its models only make use of the 1k most frequent words per variety. It also outperforms NB in test C, even if the results are quite poor: position 16 out 18 systems and 13 points less than the best one: 0.381 *vs* 0.513. The poor results are likely due to the fact that we used the same preprocessing than that performed for the sub-task 1. However, the transcription of spoken language contains many metacharacters that could have been misinterpreted by our system.

To help readers to understand on which languages or groups of languages the two approaches performed better, we also include seven new tables with the confusion matrix for each test. Tables 2 and 3 for test A, tables 4, 5, 6, and 7 for tests B, and Table 8 for test C.

4.4 Efficiency

In terms of memory use, Quelingua loads a light dictionary of 136Kb (1000 words per language in sub-task 1), while the NB system requires loading much larger language models o(31Mb in sub-task 1, closed submission). Concerning speed, classification based on NB models is much slower than classification with the ranking method of Quelingua. More precisely, Quelingua is about 10 times faster than NB.

5 Discussion

We compared two very basic strategies for similar language/variety detection. We observed that Naive Bayes classifiers perform better on in-domain datasets than dictionary-based strategy, while the latter one is more stable across different domains and performs reasonably well on out-of-domain tests.

Besides the fact of performing reasonably well across different domains and genres, another benefit of the dictionary-based model is its small, transparent, and easy to handle ranked lexicon, which can be easily corrected and updated by human experts.

However, we must clarify that our Naive Bayes classifier is a class of model that can be quite sensitive to specific hyper-parameters (e.g. the kind of smoothing and the type of features - characters *vs* words). So, our work should be seen as just a comparison between a dictionary-based strategy and a particular parameterization of a Naive Bayes classifier.

In future work, we will measure the performance effects of using a manually corrected ranked vocabulary, since the dictionaries used in the described experiments were not corrected by humans. We will also analyze the growth curve of the F1-score obtained by the NB system over the corpus size. Besides, it will be interesting to compare these approaches with contextual-based strategies such as Markov Mod-

els, which were the best systems according to other evaluations (Padrò and Padrò, 2004). Finally, it will be very useful to perform a sound qualitative error analysis of the language varieties we know well: Portuguese, Spanish, and French. We have observed that many of the instances in the training dataset were annotated as belonging to a particular variety even if they did not contain any clear linguistic feature. In many cases, only cultural and extra-linguistic elements (e.g. a localized topic and named entities) could be used to discriminate between the related varieties. Further deeper analyses in this direction are required.

Acknowledgments

This work has been supported by TelePares project, MINECO, ref:FFI2014-51978-C2-1-R.

References

Kenneth R. Beesley. 1988. Language identifier: A computer program for automatic natural-language identication of on-line text. In *29th Annual Conference of the American Translators Association*, pages 47–54.

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

Ted Dunning. 1994. *Statistical Identification of Language*. Technical Report MCCS 94–273. New Mexico State University.

Willian Gale. 1995. Good-turing smoothing without tears. *Journal of Quantitative Linguistics*, 2:217–37.

Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. 2013. Tass: A naive-bayes strategy for sentiment analysis on spanish tweets. In *Workshop on Sentiment Analysis at SEPLN (TASS2013)*, pages 126–132, Madrid, Spain.

Pablo Gamallo, Susana Sotelo, and José Ramom Pichel. 2014. Comparing ranking-based and naive bayes approaches to language detection on tweets. In *Workshop TweetLID: Twitter Language Identification Workshop at SEPLN 2014*, Girona, Spain.

Marcos Garcia and Pablo Gamallo. 2015. Exploring the effectiveness of linguistic knowledge for biographical relation extraction. *Natural Language Engineering*, 21(4):519–551.

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.

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)*.

Gregory Grefenstette. 1995. Comparing two language identification schemes. In *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT 1995)*.

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, Eshrag Refaee, and Mark Dras. 2015. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, pages 209–217, Bali, Indonesia, May.

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.

Paul McNamee. 2005. Language identification: a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 3:94–101.

Muntsa Padrò and Lluís Padrò. 2004. Comparing methods for language identification. *Procesamiento del Lenguage Natural*, 33:151–161.

Radim Rehurek and Milan Kolkus. 2009. Language identification on the web: Extending the dictionary method. *Lecture Notes in Computer Science*, pages 315–345.

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.

Erik Tromp and Mykola Pechenizkiy. 2011. Graph-based n-gram language identification on short texts. In *Proceedings of Benelearn 2011*, pages 27–35, The Hague, Netherlands.

Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. 2010. Slanguage identification of short text segments with n-gram models. In *Proceedings of LREC-2010*.

Fela Winkelmolen and Viviana Mascardi. 2011. Statistical language identification of short texts. In *Proceedings of ICAAR*, pages 498–503.

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, 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 (LT4VarDial)*, pages 1–9, Hissar, Bulgaria.

Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2014. Overview of tweetlid: Tweet language identification at sepln 2014. In *TweetLID - SEPLN 2014*, Girona, Spain.

Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2015. Tweetlid: a benchmark for tweet language identification. *Language Resources and Evaluation*, pages 1–38.