A Perplexity-Based Method for Similar Languages Discrimination

Pablo Gamallo
CiTIUS
Univ. of Santiago de Compostela
Galiza
pablo.gamallo@usc.es

Jose Ramom Pichel
imaxin|software,
Santiago de Compostela,
Galiza
jramompichel@imaxin.com

Iñaki Alegria
IXA group
Univ. of the Basque Country
UPV/EHU
i.alegria@ehu.eus

Abstract

This article describes the system submitted by the Citius_Ixa_Imaxin team to the VarDial 2017 (DSL and GDI tasks). The strategy underlying our system is based on a language distance computed by means of model perplexity. The best model configuration we have tested is a voting system making use of several n-grams models of both words and characters, even if word unigrams turned out to be a very competitive model with reasonable results in the tasks we have participated. An error analysis has been performed in which we identified many test examples with no linguistic evidences to distinguish among the variants.

1 Introduction

Language detection is not a solved problem if the task is applied to the identification of similar languages and varieties. Closely related languages or language varieties are much more difficult to identify and separate than languages belonging to different linguistic families. In this article, we describe the system submitted by the Citius_Ixa_Imaxin team to the VarDial 2017. We have participated in two task: Discriminating between Similar Languages (DSL) and German Dialect Identification (GDI). The strategy underlying our system is based on comparing language models using perplexity. Perplexity is defined as the inverse probability of the test text given the model. Most of the best systems for language identification use probability-based metrics with n-grams models. This report paper (Zampieri et al., 2017) describes the shared task and compares all the presented systems.

DSL is focused on discriminating between similar languages and national language varieties, including six different groups of related languages or language varieties:

- Bosnian, Croatian, and Serbian
- Malay and Indonesian
- Persian and Dari
- Canadian and Hexagonal French
- Argentine, Peninsular, and Peruvian Spanish
- Brazilian and European Portuguese

The objective of GDI is the identification of German varieties (four Swiss German dialect areas: Basel, Bern, Lucerne, Zurich) based on speech transcripts.

Analysis about previous results on the two scenarios can be found in Goutte et al. (2016) and Malmasi et al. (2015). The latter is focused on Arabic varieties but the scenario is similar to the GDI task.

2 Related Work

2.1 Language Identification and Similar Languages

Two specific tasks for language identification have attracted a lot of research attention in recent years, namely discriminating among closely related languages (Malmasi et al., 2016) and language detection on noisy short texts such as tweets (Zubiaga et al., 2015).

The Discriminating between Similar Languages (DSL) workshop (Zampieri et al., 2014; Zampieri et al., 2015; Goutte et al., 2016) is a shared task where participants are asked to train systems to discriminate between similar languages, language

Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 109–114,
Valencia, Spain, April 3, 2017.

varieties, and dialects. In the three editions organized so far, most of the best systems were based on models built with high-order character n-grams ($>= 5$) using traditional supervised learning methods such as SVMS, logistic regression, or Bayesian classifiers. By contrast, deep learning approaches based on neural algorithms did not perform very well (Bjerva, 2016).

In our previous participation (Gamallo et al., 2016) in the DSL 2016 shared task we presented two very basic systems: classification with ranked dictionaries and Naive Bayes classifiers. The results showed 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. We also observed that basic n-gram models of characters and words work pretty well even if they are used with simple learning systems. In the current participation we decided to use basic n-grams with a very intuitive strategy: to measure the distance between languages on the basis of the perplexity of their models.

2.2 Perplexity

The most widely-used evaluation metric for language models is the perplexity of test data. In language modeling, perplexity is frequently used as a quality measure for language models built with n-grams extracted from text corpora (Chen and Goodman, 1996; Sennrich, 2012). It has also been used in very specific tasks, such as to classify between formal and colloquial tweets (González, 2015).

3 Methodology

Our method is based on perplexity. Perplexity is a measure of how well a model fits the test data. More formally, the perplexity (called PP for short) of a language model on a test set is the inverse probability of the test set. For a test set of sequences of characters $CH = ch_1, ch_2, ..., ch_n$ and a language model LM with n-gram probabilities $P(\cdot)$ estimated on a training set, the perplexity *PP* of *CH* given a character-based n-gram model LM is computed as follows:

$$PP(CH, LM) = \sqrt[n]{\prod_{i}^{n} \frac{1}{P(ch_i|ch_1^{i-1})}} \quad (1)$$

where n-gram probabilities $P(\cdot)$ are defined in this way:

$$P(ch_n|ch_1^{n-1}) = \frac{C(ch_1^{n-1}ch_n)}{C(ch_1^{n-1})} \quad (2)$$

Equation 2 estimates the n-gram probability by dividing the observed frequency (C) of a particular sequence of characters by the observed frequency of the prefix, where the prefix stands for the same sequence without the last character. To take into account unseen n-grams, we use a smoothing technique based on linear interpolation.

A perplexity-based distance between two languages is defined by comparing the n-grams of a text in one language with the n-gram model trained for the other language. Then, the perplexity of the test text *CH* in language $L2$, given the language model *LM* of language $L1$, can be used to define the distance, $Dist_{perp}$, between $L1$ and $L2$:

$$Dist_{perp}(L1, L2) = PP(CH_{L2}, LM_{L1}) \quad (3)$$

The lower the perplexity of CH_{L2} given LM_{L1}, the lower the distance between languages $L1$ and $L2$. The distance $Dist_{perp}$ is an asymmetric measure.

In order to apply this measure to language identification given a test text, we compute the perplexity-based distance for all the language models and the test text, and the closest model is selected.

4 Experiments

4.1 Runs and Data

In the DSL task we have taken part in both tracks: closed and open. The open model was trained with the datasets released in previous DSL tasks (Malmasi et al., 2016; Zampieri et al., 2015; Zampieri et al., 2014).

We prepared three runs for each task. All of them are based on perplexity but using different model configuration:

- Run1 uses perplexity with a voting system over 6 n-gram models: 1-grams, 2-grams and 3-grams of words, and 5-grams, 6-grams and 7-grams of characters. We observed that short n-grams of words clearly outperform longer word n-grams, while long n-grams of

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
run1	0.903	0.903	0.9025	0.9025
run2	0.9016	0.9016	0.9013	0.9013
run3	0.8791	0.8791	0.8787	0.8787

Table 1: Results for the DSL task (closed).

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
run1	0.9028	0.9028	0.9016	0.9016
run2	0.9069	0.9069	0.9065	0.9065
run3	0.8788	0.8788	0.8773	0.8773

Table 2: Results for the DSL task (open).

characters perform better than shorter ones. In previous experiments, this system configuration reached a similar score to the best system in the DSL Task 2016, namely 0.8926 accuracy, very close to 0.8938 reached by the best system in task A (Çöltekin and Rama, 2016).

- Run2 uses perplexity with just 1-grams of words. In the development tests, we observed that this simple model is very stable over different situations and tasks.

- Run3 also uses perplexity but with 7-grams of characters, since long n-grams of characters tend to perform better than short ones.

4.2 Results

In the first task (Discriminating between Similar Languages) we submitted systems generated with both closed and open training.

4.2.1 DSL Closed

The results obtained by our runs in the DSL task are shown in Table 1. The random baseline (14 classes) is 0.071 and the references from the best system in 2016 is 0.8938 accuracy. However, it is worth noticing that 2016 and 2017 DSL tasks are not comparable because the varieties proposed for the two shared tasks are not exactly the same.

The table shows that best results are obtained using the two first configurations: Run1 and Run2. Let us notice that the second one reaches good results even if it is based on a very simple models (just words unigrams). This is also true for the GDI task (see below in the Discussion section).

Our best run in task DSL achieved 0.903 accuracy (9th position out of 11 systems) while the best system in this task reached 0.927.

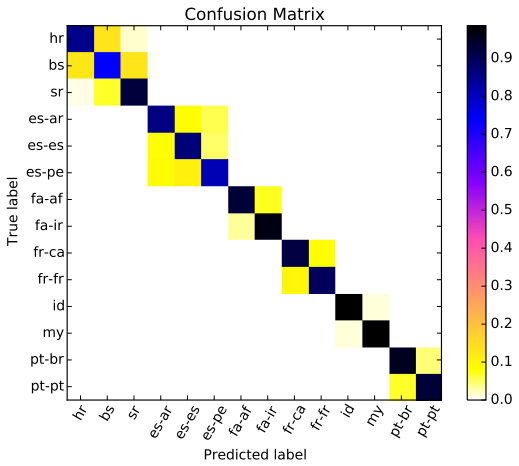

Figure 1: Confusion matrix: DSL run2

The confusion matrix for Run2 is shown in Figure 1. Bosnian and Peruvian Spanish seem to be the most difficult languages/varieties to be distinguished.

Comparing confusion matrices for Spanish variants between Run1 and Run2, we can observe that although the results are similar in both cases, they guess and fail in a different way (Table 3). So, they seem to be quite complementary strategies.

4.3 DSL Open Training

We tried to improve the results by adding more training data from previous shared tasks. Table 2 shows that the simplest configuration (Run2) gets better results than in the closed training task, but only a slight improvement (0.5 %) was obtained. No comparison can be made with other systems because the other participants did not take part in this track.

	run1			run2		
	es-ar	es-es	es-pe	es-ar	es-es	es-pe
es-ar	892	67	36	861	81	56
es-es	88	871	35	78	870	48
es-pe	111	126	763	87	104	809

Table 3: Confusion matrices in run1 and run2 for variants of Spanish

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
run1	0.6262	0.6262	0.6118	0.6108
run2	0.6308	0.6308	0.613	0.612
run3	0.5921	0.5921	0.5785	0.5774

Table 4: Results for the GDI task.

4.4 GDI

The results for the GDI task are shown in Table 4. The majority class baseline is 0.258 and there were no previous results to compare with. However, the best results for Arabic dialects in VarDial 2016 (in similar conditions to GDI) were 0.513 (F-score).

The results are much lower than in DSL task. Several factors which can influence these results are the following:

- the GDI task has unbalanced test sets,
- the data are from speech transcription,
- the task itself is more difficult given the strong similarity of the varieties.

In this task, our best configuration is Run2, which, in spite of its simple model, improves the voting-based system. The confusion matrix for Run2 (see Figure 2) shows that the scores obtained for Lucerne dialect are very poor.

Run2 achieved 0.630 accuracy (8th position out of 10 systems) while the best system in this task reached 0.680. It is worth noticing that only two systems also involved in DSL 2016 task improve our results in GDI.

5 Discussion

The results show that our system, despite its simplicity, performs reasonably well. For the DSL task 2016 we obtained the second best performance even if the results are more discrete in 2017; and for the GDI task the results are better than the best score in 2016 for the Arabic Dialectal Identification task.

It can be underlined that the configuration of our run2 is very simple (just unigrams of words) and

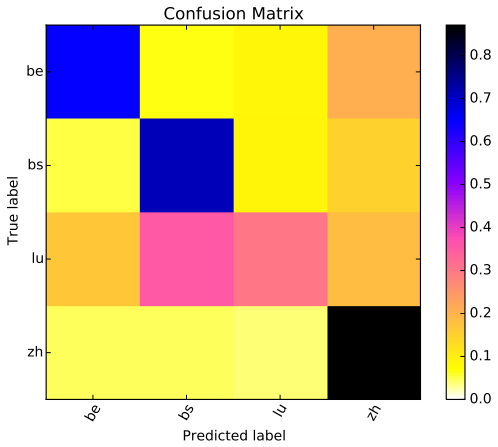

Figure 2: confusion matrix: GDI run2

results using perplexity are very competitive. It could be considered as a baseline for the future.

In order to find key elements for further improvement, we decided to carry out an analysis of errors on variants that we know quite well (variants of Spanish).

5.1 Analysis of errors in Spanish

From the list of errors among Spanish texts extracted from the evaluation carried on the development corpus we selected randomly 50 cases.

We decided to classify these texts on the following categories:

- Not distinguishable: the dialect is impossible or very difficult to classify. There are no specific language features allowing to make a distinction. For instance: *La propuesta de reunir en un mismo lugar a las etiquetas premium de las principales bodegas del país ha*

Cases	number	freq.
not distinguishable	18	0.36
distinguishable by named entities	17	0.34
distinguishable by dialectal uses	7	0.14
others	8	0.16

Table 5: Figures from error analysis on Spanish texts.

logrado cautivar al público amante del buen vino, siendo hoy el evento del sector más esperado del año. is classified by our system as Spanish from Argentina (es-AR) but it was annotated as Spanish from Spain (es-ES). However, the text has no relevant dialectal characteristic.

- Distinguishable by named entities: including geographical names (*Argentina*, *Galicia*, ...), organizations (*PP, PSOE*), localization information (*euro, peso, peruano, Buenos Aires*, etc.). For instance: *Los ingresos tributarios totales de la provincia ascendieron en marzo a 1.305.180.533,54 pesos, un 10,37 por ciento por encima del monto presupuestado para ese mes* is classified by our system as es-ES, but it contains the term *pesos* which refers to the Argentinian currency.

- Distinguishable by dialectal uses. These are cases in which it is possible to find words such as *mamá* or *tercerizar* that are more frequent in some of the variants.

- Others: more complex cases in which it is difficult to make a decision since there are no clear language features from one particular variety. In some of the examples, several hypotheses were possible.

The figures for each case are shown in Table 5. We can observe that the first two cases (i.e not distinguishable and distinguishable by named entities) are the more frequent in the test test.

5.2 Future Work

Based on the error analysis we are planning to test a variant of our system with two new features:

- The system will be provided with the *none* category for those cases where there is no enough evidence to make a decision. This can increase the precision of the system.

- The system will be enriched with lists (gazetteers) of named entities linked to the dialects or geographical locations. These gazetteers could be used to assign weights to n-grams or as new features in the voting system. However, it will be necessary to consider the interferences that this new information might add to the system. For instance, in the following example (*Es indudable que los que utilice en los partidos amistosos que jugaremos contra España, en Huelva el 28 de mayo, y ante México...*), the use of localized named entities could generate a false positive for Spanish from Spain (es-ES).

Additionally we intend to test the perplexity strategy to measure the distance among the language or dialects in a diachronic mode. This would allow us to observe the quantitative transformations of the languages/dialects and the relations among them.

Finally, we will perform further experiments with different voting systems in order to find the most appropriate for our models.

Our perplexity-based system to measure the distance between languages is freely available at `https://github.com/gamallo/Perplexity`.

Acknowledgments

This work has been supported by a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, by TelePares (MINECO, ref:FFI2014-51978-C2-1-R) and TADeep (MINECO, ref: TIN2015-70214-P) projects. It also has received financial support from the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF).

The authors thanks the referees for thoughtful comments and helpful suggestions.

References

Johannes Bjerva. 2016. Byte-based language identification with deep convolutional networks. In *Proceedings of the 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (VarDial)*, Osaka, Japan.

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.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pablo Gamallo, Iñaki Alegria, José Ramom Pichel, and Manex Agirrezabal. 2016. Comparing Two Basic Methods for Discriminating Between Similar Languages and Varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 170–177, Osaka, Japan.

Meritxell González. 2015. An analysis of twitter corpora and the differences between formal and colloquial tweets. In *Proceedings of the Tweet Translation Workshop 2015*, pages 1–7.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating Similar Languages: Evaluations and Explorations.

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.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

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.

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.