Language Identification based on a Discriminative Text Categorization Technique

Miguel A. Caraballo, Luis F. D'Haro, Ricardo Cordoba, Rubén San-Segundo, José M. Pardo

Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid - E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040-Madrid, Spain

{macaraballo,lfdharo,cordoba,lapiz,pardo}@die.upm.es

Abstract. In this paper, we describe new results and improvements to a language identification (LID) system based on PPRLM previously introduced in [1] and [2]. In this case, we use as parallel phone recognizers the ones provided by the Brno University of Technology for Czech, Hungarian, and Russian languages, and instead of using traditional n-gram language models we use a language model that is created using a ranking with the most frequent and discriminative n-grams. In this language model approach, the distance between the ranking for the input sentence and the ranking for each language is computed, based on the difference in relative positions for each n-gram. This approach is able to model reliably longer span information than in traditional language models obtaining more reliable estimations. We also describe the modifications that we have being introducing along the time to the original ranking technique, e.g., different discriminative formulas to establish the ranking, variations of the template size, the suppression of repeated consecutive phones, and a new clustering technique for the ranking scores. Results show that this technique provides a 12.9% relative improvement over PPRLM. Finally, we also describe results where the traditional PPRLM and our ranking technique are combined.

Keywords: Language Identification, n-gram frequency ranking, discriminative rankings, text categorization, PPRLM

1 Introduction

Currently, one of the most used technique in Language identification (LID) is the phone-based approach, like Parallel phone recognition followed by language modeling (PPRLM) [3]. In PPRLM, the language is classified based on statistical characteristics extracted from the sequence of recognized allophones.

In spite of the high LID accuracy results obtained by PPRLM, the accuracy is reduced because PPRLM does not model correctly long-span dependencies (i.e. to use high order n-gram language models) probably due to an unreliable estimation of the n-gram probabilities. We propose to use a ranking of occurrences of each n-gram with higher n-grams, in a similar way to [4] and [5] where the ranking is applied to written

© Springer-Verlag Berlin Heidelberg 2011

text. Although the information source is very similar to PPRLM (frequency of occurrence of n-grams), results are much better, as we will see.

This paper is a continuation of the work done in [1] and [2] but tested on a new database with 4 languages and with new improvements to the ranking algorithm. Section 2 describes the system setup and basic techniques. In Section 3, the basic n-gram ranking technique and the new discriminative n-gram ranking are described, together with the results considering all the new alternatives considered. Finally, conclusions and future works are presented in Section 4.

2 System description

2.1 Database

We have used the C-ORAL-ROM database [6], which consists of spontaneous speech for 4 main Romance Languages: Spanish, French, Portuguese, and Italian. This database is made up of 772 spoken texts with more than 120 hours of speech and around 300K words for each language.

The database transcriptions and annotations were validated by both external and internal reviewers. The database includes recordings in two different types: formal and informal. The formal recordings consist of three different contexts: natural (e.g. political speech, teaching, preaching, etc.), media (e.g. talk shows, news, scientific press, etc), and telephone (e.g. private and human-machine). The informal recordings include monologues, dialogues, and conversations in familiar and public contexts.

We needed to do several tasks to adapt the database to our experiments and recognition system: a) Most of the sound files were sampled to 22,050 Hz @ 16 bits and some others to 11 KHz @ 16 bits, all of them were sub-sampled to 8 KHz @ 16 bits in order to use them with the acoustic models of our recognizer. b) Some recordings in the database were too long (i.e. longer than 10 minutes) so they were split into shorter files. We also eliminated sections with noises, c) finally, we generated random recording lists in order to avoid any kind of bias at training. Table 1 shows the number of sentences in the database that we have finally used. The average sentence length is 6.2 seconds.

Table 1. Number of sentences by language

	Spanish	**French**	**Italian**	**Portuguese**
Sentences	17634	16474	19074	17946

2.2 General conditions of the experiments

In our previous work, we used two phoneme recognizers, for Spanish and English, with context-independent continuous HMM models. Now, we present the results using three phone recognizers in Czech, Hungarian and Russian developed by Brno

University of Technology, which are based on using neural network classifiers and were trained on the SpeechDat-E databases.

The phoneme recognizers output contains many relevant errors for several reasons: a) there is a mismatch between the recognizers' languages and the four languages to be identified; b) the recordings still contain different kind of noises, background music, etc., and very spontaneous speech; c) the acoustic models were not adapted to this database. So, there is a clear mismatch in the languages and in the channel conditions. At least, improvements obtained with our techniques will be more evident, as we will see.

In order to increase the reliability of the results presented in the next sections, we performed a cross-fold validation, dividing all the available material in 9 subsets: 5 subsets to estimate the LMs, 2 subsets to estimate the Gaussian classifier, 1 subset for development, and 1 subset for test.

2.3 Description of PPRLM

Nowadays, PPRLM is one of the two typical approaches to language identification. The main objective of PPRLM is to model the frequency of occurrence of different allophone sequences in each language. The technique can be divided into two stages. First, several parallel phone recognizers take the speech utterance and outputs a sequence of allophones corresponding to the phone sets used for each recognizer. Second, the sequence of allophones is used as input to a bank of n-gram language models (LM) in order to capture phonotactics information. The LM module provides the probability that the sequence of allophones corresponds to a given language.

The main advantages of PPRLM are: a) since it uses many recognizers, it is possible to cover most of the phonetic realizations of every language. b) It is possible to have phone recognizers modeled for languages different to the languages that we have to identify, which is especially useful in situations where the training data is not enough to obtain reliable models. On the other hand, PPRLM presents a major weakness: the data sparsity limits the LMs ability to model long span information.

For score normalization, given the good results obtained in [7], we decided to continue using a Gaussian Classifier as a backend. These classifiers also benefit from applying normalization of the scores (e.g., the T-norm normalization). In our system, we use what we call "differential scores", which applies a similar normalization.

Regarding solutions for the problem of including long span information to the language models, in [8] they describe slight improvements on the LID rate when using the skip-gram technique. In [5] they present LID experiments on written text for six languages using three different kinds of LM: Markov models, trigram frequency vectors, and n-gram text categorization, with good results for the last technique. Finally, in [9] an interesting algorithm for selecting high order n-grams based on dynamically keeping the most frequent ones is presented but the selection is not based on discriminative information among languages. In our case, we have used and extended the n-gram text categorization technique proposed in [4].

2.4 PPRLM Results for LID

One problem with the PPRLM approach is that it is affected with a bias that appears in the log-likelihood score for the languages considered. To tackle this issue, we proposed in [7] to use a Gaussian classifier instead of the usual decision formula applied in PPRLM. With all the scores provided by every LM in the PPRLM module we prepare a score vector. With all the sentences in the training database, we estimate a multi-Gaussian distribution for each language. In recognition, the distance between the input vector of LM scores and the Gaussian distributions for every language is computed, using a diagonal covariance matrix, and the distribution which is closer to the input vector is the one selected as identified language.

One important conclusion of our work in [7] is that, instead of absolute values, we need to use differential scores: the difference between the score obtained by one LM and the average score obtained by the other 'competing' languages: (SCi' = SCi – Aver(SCj, j≠i)). We applied it to unigram, bigram and trigram separately, with 4 languages x 3 phone decoders x 3 n-grams = 36 features in total in the feature vector.

The average result in LID for PPRLM is 29.89% error rate. It seems a high rate, but, as we mentioned in Sections 2.1 and 2.2, the performance of the acoustic models is really poor and the sentences average length is short.

3 N-Gram Frequency Ranking

In this section, we will describe the original text categorization technique and the modifications that we have made to improve it, as well as the algorithm for selecting the most discriminative n-grams and to rank them.

3.1 Description of the Basic Technique

In [4], an interesting technique that combines local information (n-grams) and long-span information (collected counts from the whole utterance) is described. In summary, during training, the original technique proposes the creation of a ranked template with the N (typically 400) most frequent n-grams (up to n-grams of order five) of the character sequences in the train corpus for each language sorted by occurrence.

During the evaluation, a dynamic ranked template is created for the phoneme sequence of the recognized sentence following the same procedure. Then a distance measure (OOP, Out-Of-Place) is applied between the input sentence template and each language dependent template previously trained. The distance for a given ranking T is calculated using Equation 1.

$$d^T = \frac{1}{L}\sum_{i=1}^{L} abs(pos\ w_i - pos\ w_i^T) \tag{1}$$

Where L is the number of n-grams generated for a given input sentence. If an n-gram does not appear in the global ranking (meaning that it is not in the top n-grams selected) it is assigned a maximum distance, i.e., the size of the ranking. The selected language is the one that presents the lower distance between templates. Fig. 1 shows an example of one of the templates created in our system using the English phoneme set and the template created for the unknown sentence.

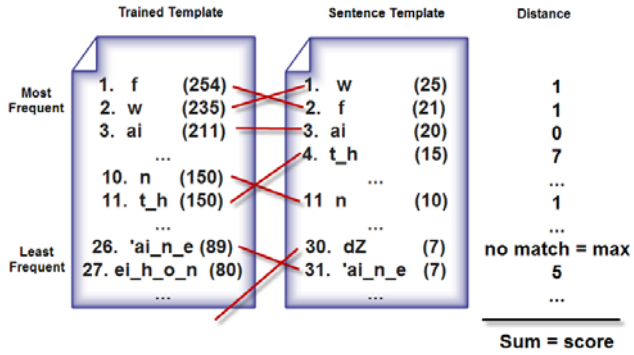

Fig. 1. Example and calculation of distance score using a ranking of n-grams as proposed by [4]

3.2 Our baseline for N-Gram Ranking

In [2] we described several modifications that we made on the basic technique proposed in [4]. We will mention here just the most relevant one. We applied what we called the "golf score". As the number of occurrences of the n-grams in the input sentence is very low, most n-grams have the same number of occurrences and should have the same position in the ranking. It is the same as a ranking in golf (the sport): all players with the same number of strokes share the same position. Fig. 2 shows an example of the modification applied to the original template, which provided a 2.4% relative improvement.

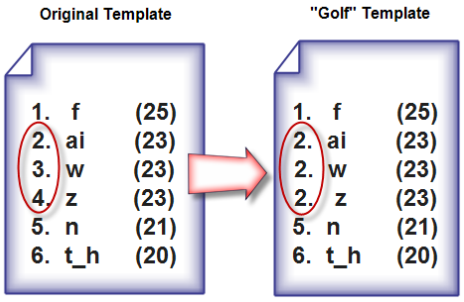

Fig. 2. Ranking template modification with "golf score"

3.3 N-Gram Discriminative Ranking

Inspired in the work presented in [10], where better LID results could be obtained using the most discriminative units, we thought that we should introduce the same concept in the ranking creation process; therefore, we decided to give more relevance (higher positions) in the ranking to the items that are actually more specific to the language that is being identified, i.e. n-grams with a high frequency in one language but with zero or low frequency in the competing languages.

In our work we propose a variation of tf-idf. After the original global rankings are created, we have the number of occurrences of each n-gram: $n_1(w)$ = occurrences of n-gram w in the current language, and $n_2(w)$ = the occurrences of w in the competing language, where T is the whole set of ranking templates created for each language.

$$N_1 = \sum_{\forall w \in T_i} n_1(w) \qquad N_2 = \frac{1}{|T-1|} \sum_{\forall w \in T_j, j \neq i} n_2(w) \tag{2}$$

As the number of total occurrences will be different for each language and n-gram order, before the subtraction a normalization is needed to have comparable amounts. Being N_1 the sum of all occurrences for the current language and N_2 the average for the competing languages (see Equation 2):

$$n_1'(w) = \frac{n_1(w) \times N_2}{N_1 + N_2} \tag{3}$$

$$n_2'(w) = \frac{n_2(w) \times N_1}{N_1 + N_2} \tag{4}$$

Another important issue is to use a threshold value for these normalized values (Equations 3 and 4), i.e., to discard the n-grams that were below a threshold as non-representative. In our case, we obtained an optimum using 6-6-2-2-2 (from 1-gram to 5-gram from left to right). Then, we considered several alternative formulas (shown in Table 2) with the same philosophy as tf-idf for the final number of occurrences used to assign the final position in the ranking (which we will call n_1'').

Table 2. Different formulas for discriminative selection used in the experiments

1	$n_1'' = n_1' * (n_1' - n_2') / (n_1' + n_2')$
2	$n_1'' = \log(n_1') * (n_1' - n_2') / (n_1' + n_2')$
3	$n_1'' = (n_1' - n_2') / (n_1' + n_2')$
4	If $n_1' > n_2' \rightarrow n_1'' = n_1' * (n_1' - n_2') / (n_1' + n_2')^2$ Else $n_1'' = n_2' * (n_1' - n_2') / (n_1' + n_2')^2$
5	If $n_1' > n_2' \rightarrow n_1'' = (n_1' - n_2') / (n_1' + n_2') * [1+\log(n_1' / (n_1' + n_2'))]$ Else $n_1'' = (n_1' - n_2') / (n_1' + n_2') * [1+\log(n_2' / (n_1' + n_2'))]$
6	If $n_1' > n_2' \rightarrow n_1'' = (n_1' - n_2') / (n_1' + n_2') * \text{sqrt}(n_1' / (n_1' + n_2'))$ Else $n_1'' = (n_1' - n_2') / (n_1' + n_2') * \text{sqrt}(n_2' / (n_1' + n_2'))$

In our experiments, we obtained similar results from formulas 4, 5, and 6. In all cases, they improved the baseline with a 5% reduction in LID error rate. We decided to use formula 4 as it meant a more consistent improvement considering several experiments, also because it was slightly better for 3-grams and 4-grams which are the most discriminative ones.

All the formulas that we propose (3-6) normalize the values between 1 and -1: where 1 means that the n-gram appears in the current language but not in the other competing ones (n_2'=0), therefore indicating that the n-gram is especially relevant for that language; -1 means just the opposite (n_1'=0).

3.4 Suppression of repeated consecutive phonemes

In the phone recognizer outputs, several consecutive phonemes are the same, especially for silences. These repeated silences affect the n-gram calculation, especially for 4-gram and 5-gram. Therefore, we decided to suppress them, leaving one instance of each phoneme.

Table 3. LID error rate and improvement obtained after removing repeated consecutive phonemes.

	Original output	Filtered output	Relative improvement
1-gram	46.89	45.35	3.30%
2-gram	35.71	32.07	10.18%
3-gram	30.65	27.56	10.11%
4-gram	33.12	30.03	9.31%
5-gram	46.83	44.07	5.89%
All	30.31	26.91	11.19%

We can see that this decision provides a significant improvement, so we will use it on all experiments.

3.5 Influence of the template size

In our first experiments we worked with template sizes up to 4000. One thing that we observed with this database is that sizes could be increased more drastically with success. Another obvious point is that the template size should be different depending on the n-gram order, as the number of units is clearly different. Therefore, we run a series of experiments varying the template size for the different n-gram orders, which we can see in Fig. 3.

We can draw several conclusions from the figure: 1-gram and 2-gram are not affected by these template sizes, which could be expected as the number of items is usually below 3,000. The saturation points are: 14,000 units (3-gram), 34,000 (4-gram), 66,000 (5-gram). This could also be expected as there are more input n-grams

as the order increases. The best result now is 26.50% LID Error rate, slightly better than the 26.91% that we obtained with the previous non-optimized template sizes.

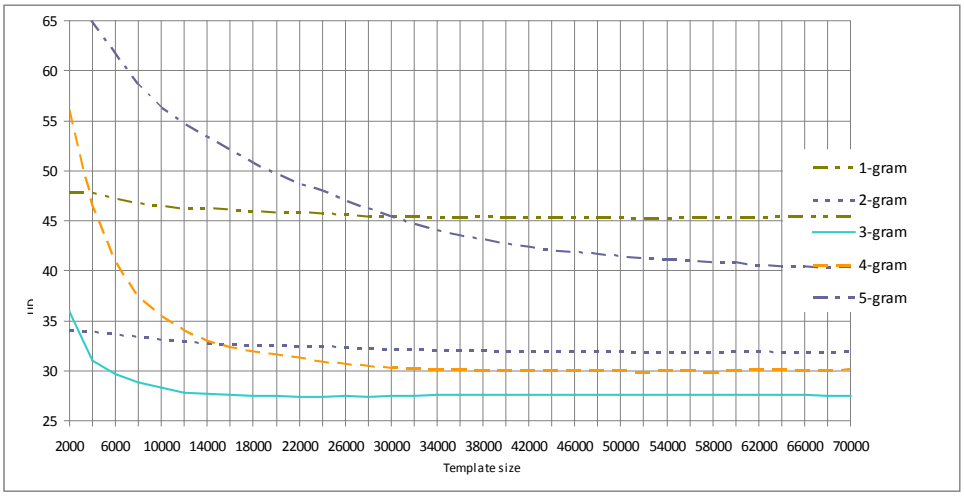

Fig. 3. Template size results for each n-gram

3.6 Clustering of ranking scores

In the previous section, we observed that results tend to saturate as the template size increases. We think this is due to the following: if the size is too big, the out-of-rank units (the ones that do not appear in training or have low values in the scores used for creating the ranking, the n_1'' from Section 3.3) receive a great penalty in the distance formula: they are assigned the last position in the template (the template size).

So, we propose to use another approach. With the n_1'' scores (normalized between 1 and -1) we made a clustering so that units with similar scores share the same position in the ranking. The clusters are created using the classical k-means algorithm. This way we can handle a larger number of units, but still apply a reasonable penalty to unseen units: now, they are assigned the total number of clusters, which is quite a smaller value.

In Table 4 we can see the results, including the total number of clusters obtained with k-means, the total number of units in those clusters and the LID error rates. The results values correspond to the average of all k-fold experiments (hence, the decimal values). We can see that the approach is worse for 1-gram and 2-gram, which is normal as we reduce the number of units, and hence, the precision. But it is extremely nice that the performance increased for 4-gram and slightly for 5-gram. We also observe that the final number of clusters is small for 4-gram and especially 5-gram, which is the result of having many units with very similar scores.

Table 4. Clustering of ranking scores.

	No. units	No. clusters	LID error rate with clustering	LID error rate without clustering	Improvement
1-gram	51.2	27.8	63.41	45.48	-38.71%
2-gram	1,510.2	733.7	31.81	31.96	0.98%
3-gram	20,015.6	1,985.4	27.63	27.34	-0.60%
4-gram	55,397.4	1,225.9	**28.96**	29.85	**2.08%**
5-gram	40,875.5	392.8	37.60	40.38	0.10%
All			26.23	26.50	-0.33%

The conclusion is that we should use our regular templates for 1-gram and 2-gram, and the clustering approach for the rest.

3.7 Combination of PPRLM and N-Gram Discriminative Ranking

We checked whether the two techniques were complementary, so we fused them. The summary of results is as follows:

- PPRLM: 29.89%
- N-gram Discriminative Ranking: 26.23% (12.2% relative improvement)
- PPRLM + N-gram: 26.04% (12.9% relative improvement)

So, the fusion of both of them does not provide significant improvements in these experiments.

4 Conclusions and Future Work

We have shown that the n-gram Frequency Ranking approach overcomes PPRLM due to the longer span that can be modeled, especially for the effect of the 4-gram, and partially of the 5-gram information. The following issues have been crucial:

- n-gram discriminative rankings with the normalized value for the number of occurrences are able to overcome PPRLM (12.9% relative improvement).
- Using a ranking score normalized between 1 and -1 provides significant improvements.
- The ranking size should be different for n-gram orders.
- The clustering of ranking scores provides further improvements as it lets to consider all units appearing in training that are above a threshold.
- The suppression of repeated consecutive phonemes provides a significant improvement, 11.19%.

As future work, we will work with NIST LRE databases, and we will fuse the results with systems based on acoustic information.

5 Acknowledgements

The work leading to these results has received funding from the European Union under grant agreement 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02) and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects.

6 References

1. Cordoba, R., D'Haro, L.F., et al. "n-gram Frequency Ranking with additional sources of information in a multiple-Gaussian classifier for Language Identification". V Jornadas de Tecnología del Habla, pp. 49-52, 2008. Bilbao, Spain.
2. Cordoba, R., D'Haro, L.F., et al. "Language Identification based on n-gram Frequency Ranking". Interspeech 2007, pp. 354- 357.
3. Zissman, M.A., "Comparison of four approaches to automatic language identification of telephone speech," IEEE Trans. Speech&Audio Proc., v. 4, pp. 31-44, 1996.
4. Cavnar, W. B. and Trenkle, J. M., "N-Gram-Based Text Categorization", Proc. 3rd Symposium on Document Analysis & Information Retrieval, pp. 161-175, 1994.
5. Vatanen, t., Väyrynen, J. and Virpioja, S. "Language Identification of Short Text Segments with N-gram Models". Int. Conf. on. Language Resources and Evaluation (LREC'10), 2010.
6. Cresti, E. et al. "The C-ORAL-ROM CORPUS. A Multilingual Resource of Spontaneous Speech for Romance Languages". IV Int. Conf. on Language Resources and Evaluation, 2004.
7. Córdoba, R., et al. "Integration of acoustic information and PPRLM scores in a multiple-Gaussian classifier for Language Identification". IEEE Odyssey 2006.
8. Navratil, and J. Zühlke, W. "Double bigram-decoding in phonotactic language identification". ICASSP, Vol. 2, pp. 1115–1118. 1997.
9. Penagarikano, M. et al. "A dynamic approach to the selection of high order n-grams in phonotactic language recognition".Acoustics, Speech and Signal Processing (ICASSP), 2011.
10. Nagarajan, T., and Murthy, H. A. "Language Identification Using Parallel Syllable-Like Unit Recognition". ICASSP, pp. I-401-404. 2004.