Adapting Statistical Language Identification Methods for Short Queries

Alexandru-Lucian Gînscă, Emanuela Boroş, Adrian Iftene

UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University,
General Berthelot, 16, 700483, Iasi, Romania
{lucian.ginsca, emanuela.boros, adiftene}@infoiasi.ro

Abstract. This paper describes the participation of UAIC team at the LogCLEF 2011 initiative, language identification task. Our approach is an aggregation of known methods for recognizing languages. Short texts are a real challenge in applying a language identification tool; so, our methods had to comply with it by resisting to noisy data as only one letter, only numbers, links, different symbols. We applied n-grams extraction with distance measurement computing and a learning algorithm. The results were satisfying on specific languages, considering that our system supports only a limited number of languages.

Keywords: Language Identification, Multilingual Context, Statistical Methods

1 Introduction

The LogCLEF multilingual log analysis evaluation initiative has created the first long-term standard collection for evaluation purposes in the area of log analysis. The LogCLEF 2011[1] lab it is the continuation of the past two editions: as a pilot task in CLEF 2009, and a workshop in CLEF 2010.

In the last years, Cross-Language Information Retrieval (CLIR) systems and projects like Europeana[2], CACAO[3] or MICHAEL[4] were oriented to support multilingual resources and to perform operations in a multilingual context [1].

The aim of LogCLEF 2011[5] is the analysis and classification of queries in multilingual contexts. Log data constitute an important aspect that allows us to evaluate a search engine and the quality of a multilingual search service.

At LogCLEF 2011, organizers proposed three tasks[6]: (1) *Language identification task*: where participants are required to recognize the actual language of the query submitted; (2) *Query classification*: where participants are required to annotate each query with a label which represents a category of interest (for example, category can be *Person*, *Geographic Location*, *Event*, *Work title*, *Domain Specific*, *Other*); (3)

[1] LogCLEF 2011: http://ims.dei.unipd.it/websites/LogCLEF/Overview.html
[2] Europeana: http://version1.europeana.eu/web/europeana-project/
[3] CACAO: http://www.cacaoproject.eu/
[4] MICHAEL: http://www.michael-culture.eu/
[5] LogCLEF Topic and Goal: http://ims.dei.unipd.it/websites/LogCLEF/Topic_and_Goal.html
[6] LogCLEF Tasks: http://ims.dei.unipd.it/websites/LogCLEF/Tasks.html

Success of a query: where participants are required to study the trend of the success of a search. The success can be defined in terms of time spent on a page, number of clicked items, actions performed during the browsing of the result list.

In the following, we present the approach of our group to build a system for the first task.

2 Language Identification

Our core language identification module is a component of the Sentimatrix[7] system [2]. Language identification and modeling are used in many natural language processing applications such as speech recognition, machine translation, part-of-speech tagging, parsing and information retrieval. These processes represent important steps in creating a viable system.

2.1 General language identification methods

Language detection is a preprocessing step problem of classifying a sample of characters based on its features (language-specific models). Currently, the system supports German, English, Greek, Spanish, French, Hungarian, Italian, Latvian, Dutch, Polish, Portuguese, Romanian, Russian, Slovene, Czech and unknown language. We combined three methods for identifying the language: *N-grams detection*, strictly *trigrams detection* and *unigrams, bigrams and trigrams detection* [3, 4, 5]. We created a corpus for every language. This is constructed by samples of text and n-grams models. The models are created by extracting the n-grams from large data collection from European Parliament Proceedings Parallel Corpus 1996-2009[8]. The trigrams models are approximately 100KB each.

There are three main methods for language detection: the *first one* is based on the trigrams models [3], the *second one* is based on sample texts [4] and the *third one* on unigrams, bigrams and trigrams models [5]. The language detection in the trigrams cases, for comparing the query's trigrams with corpus data, it is performed a distance measurement between languages profiles.

The N-grams classification method implies, along with computing frequencies, a posterior Naive Bayes implementation [6]. The corpus for this method is used from corpus from the Cybozu Labs language detection library[9]. Each N-gram *i* from every language *j* is mapped with a frequency:

$$P(i,j) = \frac{C(i,j)}{\sum_i C(i,j)} \quad , \qquad where$$

P (*i, j*): Frequency of a N-gram *i* in language *j,*

[7] Sentimatrix: www.sentimatrix.eu

[8] European Parliament Proceedings Parallel Corpus 1996-2009: http://www.statmt.org/europarl/

[9] Language Detection library: http://code.google.com/p/language-detection/

$C(i, j)$: Count of the i-th N-gram in the j-th language,
$\Sigma i\ C(i, j)$: Sum of the counts of all the N-grams in language j.

We compute a posterior Naive Bayes:

$$P(L_k|X) \propto P(L_k|X) \times P(F_j|X), \qquad where$$

L_k: Language category,
X: Document whose language needs to be detected (set of features F_j),
F_j: Feature/N-gram j of document.

$P(L_k|X)$ for every language k knowing in order to classify the test document is computed normalizing at every step, in concordance with $P(i, j)$, the probability until it becomes closer to 1.

2.2 Short or ambiguous query specific methods

The methods described in the previous section are applicable to a general language identification task. We will now present the main issues encountered when dealing with language identification for very short queries and several methods to overcome these problems.

A first issue was the significant number of queries for which the language was unknown or undecided. We preserved the notation used in the annotated queries for this situation and we attributed the "zxx" value to the queries we decided that fall in one of the two previous categories. There are several reasons that a query cannot be linked to a certain language. The most obvious ones were the cases in which the query contained mostly *digits*, such as *dates* or ISBN codes and when the query had less than three characters. These can be easily treated by identifying numerical patterns in the query or, in the second case, by checking the length of the query.

A much more difficult task appears when an undecided language tag is associated with the query due to the fact that in the query appears a *named entity*, which can be a person or a geographical entity or the title of a literary work is found in the query. This kind of query is generally treated as language independent, but sometimes language specific diacritics or spelling can suggest the origin of the named entity without any other background knowledge. It can be observed that even the inner annotator agreement is low on this situation. For example, the query "*marquis angelo gabrielli*" marked as having the type *Person* has "*zxx*" in the language tag, but the query "*karvinen marita*", also a *Person* is considered to be in the "*fin*" (Finish) language.

We managed to improve our results by building and using dictionary that maps specific diacritics to a source language. If a special character can be found only in a single language, then the problem is solved. If the character is common in more than one language, then the probabilities of belonging to one of those languages, calculated by the methods described in the 2.1 section, are given a boost.

As a solution for when we weren't able to identify the language strictly from the form of the query, we introduced a threshold for the probabilistic values. If the

language with highest probability for a query has the score under the threshold, then the language will be "*zxx*". We settled on a threshold value of *0.70*. We will discuss how we obtained this value in the next section.

3 Experiments and results

We used the queries provided after the query annotation task of this year's LogCLEF initiative for our system's evaluation. There were 25 languages used, including "*zxx*" for unknown or undecided totaling a number of 510 queries. We show in Table 1 the distribution of queries by language.

Table 1: Number of queries for each language

Lang	Queries	Lang	Queries	Lang	Queries	Lang	Queries	Lang	Queries
zxx	199	**spa**	16	**dut**	6	**cat**	1	**Hrv**	1
eng	159	**ita**	10	**gre**	4	**slvo**	1	**Sv**	1
fre	35	**pol**	7	**cze**	4	**fin**	1	**Lit**	1
ger	28	**por**	6	**srp**	2	**rum**	1	**Tur**	1
rus	17	**lat**	6	**ukr**	1	**slv**	1	**Cs**	1

As it can be seen in Table 1, an important number of languages are poorly represented and we didn't train our system for some of these languages. This translates to lowering the maximum achievable accuracy. If we disregard the languages that have less than 10 queries in the collection, we can expect a maximum 9.01% drop of accuracy.

In our best experiment, we obtained a global accuracy of *62.54* %. In Figure 1, we provide detailed results of the accuracy obtained by our system for every language, including the ones that are represented by only one query. We can observe very encouraging results, *90* % accuracy for queries marked as unknown or undecided. These results are important because this is one of the top priority values that we tried to maximize. One of the focuses of our research for the language identification task was to find ways to improve the number of correctly identified "*zxx*" queries. On the other hand, we obtained a less than expected accuracy for the *English* language.

An interesting result is the high accuracy for languages that use a particular character set, such as *Russian* or *Greek*. This gives us confidence in our diacritics dictionary method.

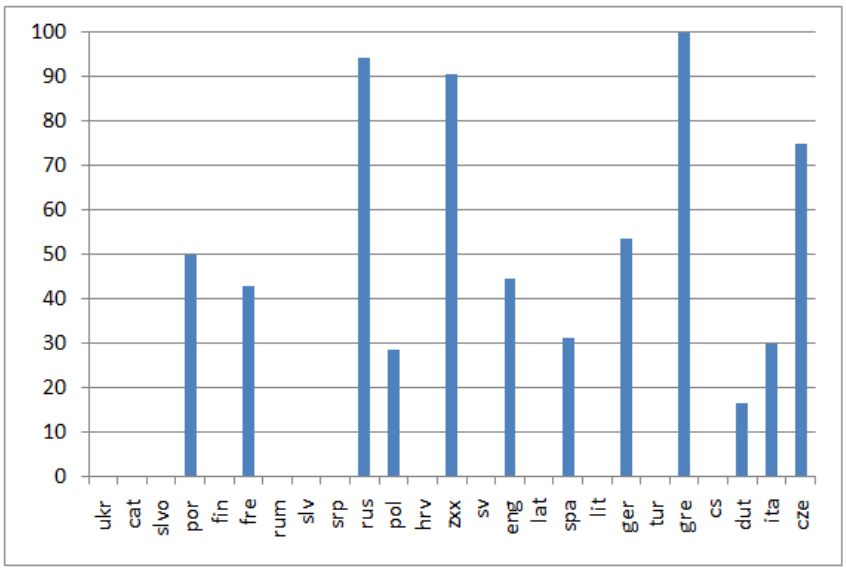

Figure 1: System accuracy for each language

In Figure 2, we can observe the influence of the threshold introduced in the previous section over the global accuracy and the accuracy for the "*zxx*" tagged queries. As expected, from a certain point a tradeoff appears between the global accuracy and the "*zxx*" accuracy. We chose *0.70* as the threshold due to the fact that it gives the best value for the general accuracy.

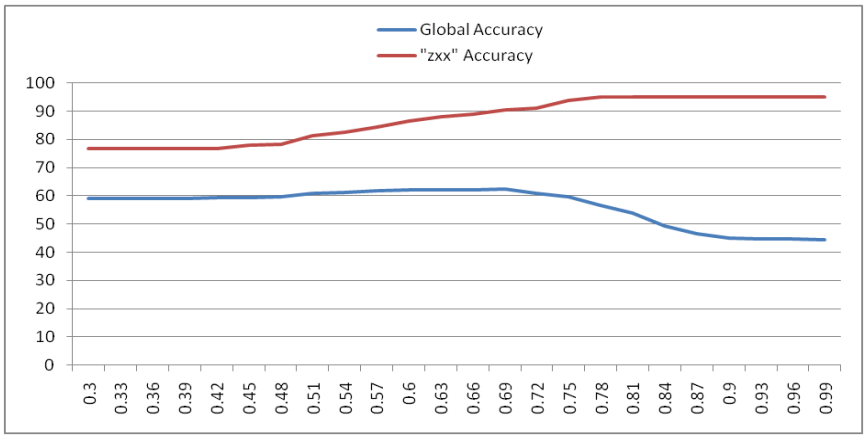

Figure 2: Threshold influence over the results

5 Conclusions

Language identification and modeling deserve necessary involvement from out team and it is important to continue investigating N-grams extracting more accurately. A large corpus would be needed, along with manual help. Noisy data was an important challenge that we think that we managed to cover partially, by using Naïve Bayes Classifier and alphabet diacritics that basically covered more than thirty percent of queries.

In conclusion, we need to better apply more language modeling techniques and to improve the ability of training the system on more languages. In addition, it will be interesting to participate in further LogCLEF initiative tasks, channeling our attention on more tasks.

Acknowledgements. The research presented in this paper was funded by the Sector Operational Program for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs" POSDRU/89/1.5/S/49944.

References

1. Bosca, A., Dini, L.: Language Identification Strategies for Cross Language Information Retrieval. In Proceedings of CLEF (Notebook Papers/LABs/Workshops)'2010. (2010)
2. Gînscă, A. L., Boroş, E., Iftene, A., Trandabăţ, D., Toader, M., Corîci, M., Perez, C. A., Cristea, D.: Sentimatrix - Multilingual Sentiment Analysis Service. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-WASSA2011). ISBN-13 9781937284060, Portland, Oregon, USA, June 19-24. (2011)
3. Cavnar, W. B., Trenkle, J. M.: N-Gram-Based Text Categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (1994)
4. Ahmed, B., Cha, S.: Language Identification from Text Using N-gram Based Cumulative Frequency Addition. Proceedings of CSIS 2004, Pace University, May 7th, (2004)
5. Ceylan, H., Kim, Y.: Language Identification of Search Engine Queries. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 1066–1074, Suntec, Singapore, 2-7 August (2009)
6. McCallum, A., Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. Proc. AAAI-98 Workshop Learning for Text Categorization. (1998)