УСТОЙЧИВАЯ ИДЕНТИФИКАЦИЯ МНОГОЯЗЫЧНОГО ДОКУМЕНТА

Н. Стефанович

IHS Markit Ltd
Минск, Беларусь
e-mail: nicolas.stefanovitch@ihs.com

Рассмотрим проблему определения языка документа, когда нет никаких предположений о документе: он может быть любого размера и содержать ноль, один или несколько языков. Определение языка считается решенной задачей, но на самом деле, среди прочих недостатков, не касается случая точного наличия или отсутствия нескольких языков в произвольных документах. Для решения этих проблем мы предлагаем подход, основанный на словарях с использованием байесовских статистик и функций Ad-Hoc. На двух наборах данных мы покажем, что с достаточной статистикой наш подход способен дать очень удовлетворительные результаты для двух нерешенных задач: обнаружение документов без каких-либо языков и идентификация языков в многоязычных документах.

Ключевые слова: обработка естественного языка; определение языка; многоязычные документы.

ROBUST MULTILINGUAL DOCUMENT IDENTIFICATION

N. Stefanovitch

IHS Global
Minsk, Belarus

We consider in this paper the problem of detection the language of document when no assumptions are made about a document: it can be of any size and contain zero, one or several languages. Language identification is considered a solved task, but actually, among others shortcomings, does not deal with the case of accurately the presence or absence of several languages in arbitrary documents. In order to tackle these problems, we propose an approach based on word dictionaries using Bayesian statistics and ad-hoc features. We show on two datasets that with sufficient statistics our approach is able to give very satisfying results in dealing with both unsolved tasks: detection of documents with no languages and identification of languages in multilingual documents.

Keywords: Natural Language Processing Language Identification; Multilingual Documents.

1. INTRODUCTION

The problem of language detection of textual document is often, yet inaccurately, considered a trivial and solved task [1]. It is indeed the case that when a document is monolingual, has a relatively large size and contains mostly word tokens, the problem of language identification can be straightforwardly solved. However, when dealing with arbitrary documents, such assumptions cannot be made. It is not possible to assume that the document contains a single language rather than several, it cannot even be assumed that the document contains a language at all. Correct identification of the languages of a document is important in production systems, as errors at this preliminary step can be costly further down the processing chain.

Multilingual detection has been widely studied for audio document but very scarcely for written documents. Detection of documents without language has received no attention.

We present in this paper a solution for the dealing with these problems. This approach, contrary to the vast majority of the state of the art approaches to language detection, is not based on character *n*-grams, but on word dictionary, Bayesian statistics and engineered ad-hoc features. As such it is a refreshing new take on a current unsolved problem with old methods [2].

2. BACKGROUND

The most widespread approach in literature to deal with the task of language identification is the use of *n*-grams, and such is the case of the only other approach we are aware of that specifically tackles with multilingual detection [3]. Character *n*-grams approaches chunk words in overlapping sequence of n consecutive characters. They have been favored as being simple to use, fast, and having a low memory imprint. However, as we experimented with processing real document with arbitrary content using 4-grams it appeared that these approaches are unable to make the distinction between valid textual document and document containing not a single word but containing sequences of bytes appearing in the *n*-gram dictionary. Such documents are typically produced by erroneous conversions from pdf or doc format. Moreover, *n*-grams based approaches are unable to discriminate multilingual documents from monolinguals. While simple at first glance the use of *n*-grams to be effective requires a quite high value of *n*, 4 or 5, which make their memory imprint not so low and their processing time not so fast. Moreover such high value of *n* require dealing with sparsity issues and sophisticated smoothing techniques have to be applied, and thus are not as simple to be straightforwardly applied. Finally, when processing gigabytes of texts their execution time despite code optimization appeared not to be particularly fast. We thus turned our attention to using word dictionaries, an overlooked yet powerful approach which is actually one of the first approaches to have been considered for this task [2] and which recently regained some attention [4].

3. PROCEDURE

3.1. Language score computation

In order to determine the languages present in a document, we first computing the probabilities $P(l|D)$ of each language l in L, where L is the set of all supported languages, given the document D represented as a bag of words and making the assumption that the presence of each word is independent from the other ones:

$$P(l \mid D) = P(l \mid w1, ..., wn) = \prod_i P(l \mid wi) .$$

Using Bayes formula we have that: $P(l \mid w) = \frac{P(w \mid l)P(l)}{\sum_{l' \in L} P(w \mid l')P(l')}$. A naïve Bayes classifier would compute the maximum over all languages of $P(l \mid D)$, and for this purpose discards the computation of the denominator of Bayes formula for faster processing. We are however interested in the actual values in order to compute a confidence score for each language. An additional advantage of computing the denominator is that while storing and processing $P(w \mid l)$ requires logarithmic representation due to numeric stabilities issues, $P(l \mid w)$ can be used directly used, providing as such a confidence criterion that is straightforwardly interpretable and comparable. This fact is specific to our context, where the low number of languages prevents numerical instability from occurring.

$P(w \mid l)$ is the frequency of the word w inside language l and is estimated using a training set. The priors are considered equal. The value of the denominator is computed offline and included in the dictionary to speed up computations, which allows us to get the actual posterior values at no additional cost at runtime.

When building the dictionary, words with frequency lower than 5 are excluded, for each language words are included in the dictionary in decreasing frequency until the cumulative frequency of the words in the dictionary for that language reaches 95 %. A trie is used to represent the dictionary. The final size over the 47 languages supported by our systems is 41M. For comparison the 4-gram dictionary build using the same training dataset is of 7,4 M.

At runtime, each word is processed only once, the probability that this word belong to each language is computed and the posterior probability is updated if non-null. In order to improve accuracy of our detector additional features are computed given the information of the most likely language of a word: the average and maximal length of words in that language and average, maximal sequence length of consecutive words in that language. All these features are then combined by multiplying them, yielding on overall score for each language. In order to help disambiguation of closely related languages [5], the proportion of unique word is also computed.

3.2. Multilingual document detection

The score of each language is then used to determine whether the document is monolingual or multilingual, and which actual language it contains. The threshold for acceptance as monolingual or multilingual can be set depending on the criticality for a given application to wrongly recognize the language of a document. Scores are then normalized to sum to one. As a matter of fact, appreciating what constitutes a multilingual document is actually a subjective matter that depends on assumption on the input of the system and expectation of the user. Depending on the set of parameters used, detection as precise as the presence of a few words in another language can be achieved. Closely related is the question of what proportion of a document should be in a language to be reported. This again is not an exact science as there can be document containing only data and a few sentences of text, and dictionary with about the same proportion of two language but were it may be desirable to report only one. The approach we took to deal with this, is to consider that a language must have a score of a least 5 % of the potential languages, and when ranking language in decreasing order of the score, assuming the first is the main language, and that the others must have a score at least 10 % of the main language to be included. In order to deal with closely related language, languages are filtered out if the proportion of unique words they have is below a given threshold – if all language are filtered, main language is returned, which is an effective yet imperfect way of dealing with this issue, and is subject of future work.

3.3. Non-language detection

In order to determine whether the document actually contains a language, statistics directly deriving from the number of recognized words are used. If they are bellow a cut off value, the document is classified as containing no language.

4. EXPERIMENTS

We compare two approaches an *n*-gram based approach with $n = 4$ and absolute discounting smoothing, and the approach described in this paper, reported under «score» name. Both have been trained using the same data, over a set of 47 languages. Two datasets have been used for evaluation, the, first Random Web Documents, is an internal dataset of pdfs and doc documents randomly downloaded from the internet and converted to text format. Iy mostly contains medium to large size documents (from a few dozen lines to thousands). The second is the YALI public dataset [6] which contains mostly extremely short documents, with three subsets of documents under 30, 140 and 1000 characters, which respectively approximately represent about 4, 15 and 110 words documents (exact number vary with languages). This dataset has 700 documents for each subset-language pairs, with a total of 31 500 documents for each subset. Note that 45 languages are supported in common by YALI and our system. The threshold used for no-language detection requires document larger than the one of the YALI dataset and as a consequence have been tested only on the other dataset.

4.1. YALI dataset

On the YALI dataset we report the performance in table, where the columns are as following, the fact of agreeing (Right) or not (Wrong) is always with respect to the ground truth given by the dataset: «Agree» is when *n*-gram and score both agree, «Disagree» is when both approaches disagree, «*n*-gram wrong» is when *n*-gram disagree, «score wrong difference» is when score is not equal to ground truth, «score wrong inclusion» is when the ground truth is a subset of the language reports by the score method, finally the last column is when the score method reports a better mono lingual result than *n* grams. All performances are reported as a percent of the documents in the subset of a given size.

Performance on the YALI dataset

Size	Agree (%)	Disagree (%)	*n*-gram Wrong (%)	Score Wrong Difference (%)	Score Wrong Inclusion (%)	Score Right *n*-gram Wrong (%)
30	75,13	2,52	0,31	24,19	23,33	0,66
140	96,43	0,51	0,65	3,42	0,017	0,13
1000	98,80	0,18	0,20	1,17	0,002	0,02

4.2. Random Web Documents dataset

The internal random web documents dataset is comprised of 100 000 documents, half of which have been used to tune the parameters of the algorithm, the other half has been used to evaluate the performance. Total size of the evaluation data set is 222 м. This dataset has been used for two purposes: to evaluate multilingual detection and non-language detection. Because of the high cost involved in manual reviewing of documents, precision has been assessed by randomly sampling 100 documents classed as multilingual and 100 documents classed as without language.

Multilingual detection correctly reported 93 % of the multilingual documents with the exact number of languages, out of which 88 are bilingual, 4 are trilingual, and one has 4 languages. 7 documents were misclassified, two were only lists of names, 3 had a borderline minority language that should either be counted has one more language or one language less – 2 of which were indeed multilingual and 1 should have been monolingual, the other two were misconverted documents that should have been flagged has containing no language. *N*-grams reported only one language for all these documents. Note that only 750 of the 50 000 documents were classified as multilingual.

Non-language detection correctly reported 95 % of the documents do not contain languages. Misclassified documents are a list of names, guitar tablatures with almost no words, a very long document of 1527 lines of random characters and 16 lines of German at the beginning, a datasheet with almost no words and perfectly correct Korean grammar in Japanese. *N*-grams reported wrongly monolingual a language for 100 % of these documents.

Time-wise, computation using 120 threads took 50s for the score method and 4,7 min for the *n*-gram approach, making the *n*-gram approach about 5,6 times slower.

DISCUSION AND CONCLUSION

Our system has not been specifically tuned to handle short documents as in the YALI dataset, all the training has been done using the Random Web Documents dataset. The two most significant problems that impacted the performance of our system over the YALI dataset are: first, the short size which imply that out of vocabulary words are not recognized, and as such no language is reported, second that very related language, e. g. like Danish and Norwegian, are hard to distinguish using limited statistics within our framework. The first problem can be addressed using larger dictionaries, but also by recognizing that for extremely short documents n-grams are more adapted, as building a dictionary that covers all the possible word for would be too large and not even possible given the combinatorics of words in some languages. The second problem can be addressed and needs further investigation.

It is interesting to note that on the subset of documents of size 1000, when considering non-highly similar languages, all the remaining 41 reported multilingual documents are actually correctly labelled: all but one are bilingual and one has 4 languages. This fact and the disagreement column show that the YALI dataset contains some errors in its ground truths. This happens because because these documents are random samples from Wikipedia pages, which contains a relatively large number of foreign words.

Results obtained on the Random Web Documents show that when statistics collected from documents are sufficient our approach can effectively identify multilingual documents and documents without languages. To the best of our knowledge this is the only system able to deal with both problem, and only one of the few that are actually able to deal with any of them independently. The approach of [3] is limited to bilingual documents where one of the languages is English. Benefits of using dictionary based approach over n-gram are that that exact matches enable the detection of non-language documents and allows computing more precise probabilities. Remaining work consists in better handling documents contains many data and few words, notably list of names, and improving the overall performance of the system. A very promising way to achieve this goal is using state of the art supervised machine learning techniques exploiting the developed features.

LITERATURE

1. Baldwin T., Lui M. Language identification: The long and the short of the matter. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.
2. Hull D. A., Grefenstette G. Querying across languages: a dictionary-based approach to multilingual information retrieval. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1996.
3. Lau M., Lau J. H., Baldwin T. Automatic detection and language identification of multilingual documents. Transactions of the Association for Computational Linguistics 2, 2014. P. 27–40.
4. Dong-Phuong N., Seza Dogruoz A. Word level language identification in online multilingual communication. Association for Computational Linguistics, 2013.
5. Tiede J., Ljubešić N. Efficient discrimination between closely related languages. COLING 2012, 2012.
6. Majliš M. Yet another language identifier. Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012.