Categorization according to Language: A step toward combining Linguistic Knowledge and Statistic Learning[*]

Emmanuel Giguet
GREYC — CNRS URA 1526 — Université de Caen
Esplanade de la Paix
14032 Caen cedex — France
e-mail: Emmanuel.Giguet@info.unicaen.fr

Abstract

In this article, we address the problem of categorization according to language by presenting a method based on natural properties of language which allow us to categorize any kind of sentence with a very high success rate.

The major difficulties in categorization are convergence and textual errors. Convergence since dealing with short entries involve discarding languages from few clues. Textual errors since documents coming from different electronic ways may contain spelling and grammatical errors as well as character recognition errors generated by OCR.

We describe an approach to sentence categorization based on natural properties of languages. We combine several linguistic knowledge sources and we study their role. While our first system achieved a high success only on long sentences, we studied how overcoming the weaknesses and we present an interesting use of statistical profiles. Tested for french, english, spanish and german discrimination, the improvements are appreciable, achieving in one test a perfect discrimination for texts, a perfect discrimination for sentences of more than 7 words and a very high discrimination rate for very short sentences.

The resolution power is based on the comprehension that frequent phenomena are different depending on whether processing long inputs or processing short inputs. Thus, several kinds of linguistic knowledges are necessary and a way to combine them is presented. Having linguistic features of each language at its disposal, the system computes for each of them its likelihood to be selected. The name of the language having the optimum likelihood tag the input — but non resolved ambiguities are maintained.

The implementation is fast, small, robust, evolutive and textual errors tolerant.

1 Introduction

Issues in categorization according to language is fundamental for NLP, especially in document processing. In fact, with the growing amount of texts becoming available, this categorization by revealing multilingual text structure opens a wide range of applications in multilingual text analysis such as information retrieval, preprocessing of multilingual syntactic parser, and any study on one particular language in multilingual texts without parasitic noise (due to other languages).

In our first system (Giguet, 1995), we showed that it was possible to categorize long sentences and texts using only linguistic knowledge. We also remarked that one kind of knowledge could not alone solve the whole problem: long and short inputs can not be categorized using the same method. We analyse the reasons why the system did not succeed in categorizing short sentences and decide to combine linguistic knowledge and statistical knowledge to act on different levels of the input.

In the following sections, we are going to describe the different knowledges we need for categorization and the problems that arise with their combination. The evaluation of the system lead to the fact that this combination is efficient.

[*] This Paper is published in the Proceedings of the International Workshop of Parsing Technologies, Prague - Karlovy Vary, Czech Republic, September 20-24.

Note:

A text is composed by different segments : sentence, included blocks (via quotes, parenthesis, dashes or colons), dialogues, titles... All these segments are possible inputs of the system and we categorize these different entities as if there was no relation of contiguity or insertion between them.

2 Principles of Multilingual Categorization

Studying quantities of texts, we try to understand as well as possible ways to discriminate languages. We present in this section the results of our research.

2.1 Grammatical Words as Discriminant

In this section, we are going to motivate the reasons which lead us to choose grammatical words as discriminant.

What we thought: Grammatical words are proper to each language and are in a whole different from one language to another. Moreover, they are short, not numerous and we can easily build an exhaustive list. So, these words can be use as discriminant of language. But can we use them as discriminant of sentences?

Grammatical words in sentences represent on average about 50% of words. They can't be omitted because they structure sentences and make them understandable. Furthermore, relying on grammatical words allows textual errors tolerance and foreign words import from other languages (usual in scientific texts). It's also important to note that words borrowings concerns nouns, verbs, adjectives but never grammatical words.

What we found: Grammatical words allow us to categorize inputs containing more than 9 words. In fact, in very short inputs (from 1 to 3 words) we find every syntactic categories but grammatical words. So, this method does not give very good results with such inputs.

The success rate is proportionnal to the number of words of the input. In fact, the more the input is large, the more we find grammatical words. The error rate is close to 0% since ambiguities are maintained and grammatical word is a good discriminant.

2.2 Using the Alphabet

In this section, we expose an attempt to classify short sentences, using the alphabet.

What we thought: To improve categorization of short sentences, a simple way is the use of the alphabet. Alphabets are proper to each language and even if they have a great common part, some signs such as accents allows discrimination between them.

What we found: As we did not expect, the alphabet is not very useful. In fact, we saw that accented characters were not as frequent as we thought. Moreover, accented letters does not belong to only one alphabet. So, this clue allows to discard languages but not really to discriminate the right language.

The success rate is not proportionnal to the number of words of the input. It behaves quiet like a boolean choice: wether there are accented letters, or there are not. The error rate is close to 0% since there are very few (accented) words borrowed from one language to an other.

2.3 Using word knowledge

In this section, we expose a successful attempt to improve the classification of short sentences. The idea is to try to characterize non grammatical words. The conclusion on the use of statistics is interesting.

What we thought: To do this job, the traditionnal way is to exploit the difference between letter combinations in different languages (Cavnar and Trenkle, 1994). For each language, the system computes from a training set a profile based on frequency (or probability) of letter sequences (called n-gram). Then, for a given text, it computes a profile and select the language which has the closer profile.

An other way is to use knowledge upon word morphology via *syllabation*. The idea is to check the good syllabation of words in a language, distinguishing the first, middles and last syllabs. In the same way, using words endings and using sequences of voyells or consonants seems interesting.

What we found: To realize the first approach, we trained a program to learn digrams and trigrams on texts. Quadgrams did not converged fast enough to be used. The interesting point is the following: *if we want to improve the categorization of short sentences, it is a mistake to train a program on texts.*

In fact, we saw in §2.1 that a text is mainly made of grammatical words. So most frequent digrams and trigrams are grammatical word ones. A simple test showed that these ngrams did not characterize non grammatical words.

We decided to make an other training on texts with no grammatical words. The profile we get was closer to what we thought, including frequent endings and other frequent ngrams. We see that frequent endings are well-known but the other digrams depends on the training texts. A simple test showed that only endings could really be used.

The major problem is that the quality is entirely based on the training set. Profiles require a lot of data to converge and building a large representative training set is a real problem. We showed that getting statistics profiles can not be done in a fully automatic way. It always requires a manual checking.

The success rate of this method is not proportional to the number of words of the input. It behaves quiet like a boolean choice: wether there are frequent endings, or there are not. The error rate is higher than the other methods since the knowledge is not exhaustive and the quality is based on the training set.

2.4 Using Text Structure

When dealing with texts, we can use heuristical knowledge about text structure:

- In a same paragraph, contiguous sentences are written in the same language
- Titles of a paragraph are written in the same language as their body
- Included blocks in a sentence (via parenthesis, ...) are written in the same language as the sentence.

We did not implement these ideas.

Theoritical issues in this field are in progress (Lucas et al., 1993; Lucas, 1992) but as far as we know no implementation has been done yet.

2.5 Building the lexicons

The build of the lexicons sets the problem of their intersections since there are common words to different languages. For instance, *de* is a french and a spanish preposition, *é* is a french and a spanish letter. A solution consists in discarding them to get empty intersections between all the lexicons. An other one is to keep the common words. The second solution is the only one to be coherent with the principles of evolutivity, flexibility and quality.

2.6 Putting them together

In this section, we are going to describe the way we combine the previous methods and the principles of the score function.

All the methods are independent of the others. They are activated by a general control module which process words step by step. This control module attempt to guess the language of each word to infer the language of the sentence. There is no propagation of guess from one word to another since we consider borrowings. The only propagations that are consistent are those described in §2.4.

The scoring function gives different weights to the methods depending on their nature. A statistic method will have a smaller weight than an exhaustive linguistic knowledge based method. This is to compensate the error rate introduced. In our system, the weight of statistical method has arbitrarily been set to the half of the weight of an exhaustive linguistic knowledge based method.

The control of the global module is the following for each word of the sentence :

For each language:

- Call the *Alphabet method* to check if the word belongs to the language.
- If so,
  - Call the *Grammatical Word method* to ckeck if the word is a grammatical word.
  - If not, Call the *Frequent Ending method* to ckeck if the word morphology lets think it belongs to the language.

The sentence is tagged with the names of the languages which have the same and highest likelihood.

Note

It is interesting that, using these knowledges, this system will be coherent with multilingual syntactic parsers which only rely on grammatical words and endings. So, the categorization system can constitute a switch for these parsers (Vergne, 1993; Vergne, 1994).

3 Evaluation

3.1 The Test-Bed

The test-bed set has been prepared to process French, English, Spanish and German. We use dictionnaries to get the grammatical words of each language and their alphabet. We use a semi-automatic The more frequent endings

Language of Corpus	Number of inputs
French	4502
English	6735
Spanish	94
German	393

Table 1: Size of Corpus

We decided to use different kinds of documents to test robustness, speed, precision and textual errors tolerance. So, we collected scientific texts, emails and novels (see table 1).

3.2 Results

The results of the combination of these methods is very good. The grammatical word method allows a perfect discrimination of inputs of more than 7 words. The *Frequent Ending method* improve categorization of inputs of less than 9 words by a bit increasing the error rate but without loosing the benefit of *Grammatical Words method.* The *Alphabet method* does not really increase the success rate but decrease the error rate introduced by *Frequent Ending method* in short inputs thanks to the different weights assigned to the methods (see §2.6)

The results are quiet identical from one language to another so we are going to explain the results that we get with the french corpus.

From 1 to 2 words, there are mainly total indeterminations. In fact, the corpus shows that we are processing included segments (via quotes and parenthesis) or short answers in dialogues. There are no grammatical words and few clues to rely on. Deductions really start between 3 and 5 words. Here, the cooperation of the three methods is fundamental since every clue is important. From 6 to 8 words, the grammatical word method begins to be really efficient and it is helped by both other methods. From 9 words, the grammatical word method alone is able to do the job.

In table 2, with the french corpus, there were 182 inputs of 3 words. Among these inputs, 5.49% were tagged as indeterminable (i.e no deduction), 23.63% got more than one tag (i.e some languages have been discarded), 70.88% got one tag (i.e one language has been isolated).

3.3 Errors

The assignment of one tag does not mean exactly isolating the right language. The error rate is close to 0% for inputs of more than 3 words and is due to a change of language without punctuation signs in a sentence or due to an unexpected language. For inputs from 1 to 3 words, the error rate is higher and is mainly introduced by the frequent endings method.

Words	Inputs	%Indet	%hesit	%OK
1	467	87.58	3.43	8.99
2	217	24.88	34.10	41.02
3	182	5.49	23.63	70.88
4	157	0.63	10.83	88.54
5	121	0.00	6.61	93.39
6	147	0.00	3.40	96.60
7	135	0.00	1.48	98.52
8	134	0.00	0.00	100.00
9	127	0.00	0.00	100.00
10	125	0.00	0.00	100.00
...	...	0.00	0.00	100.00

Table 2: French Inputs Categorization with Frequent Endings method

4 Conclusion

The aim of this paper is also to point that we must not have an implicit trust to statistical tools. They have to be used carefully, enlightened by linguistic concepts.

We also showed that linguistic knowledge and statistical tools act on different levels of a sentence. Their combination allows to classify every kind of inputs (short sentences, long sentences and texts).

The architecture is evolutive which implies that the system can easily extended with other languages.

This tool is already a switch of Jacques Vergne's multilingual syntactic parser (for french, english and spanish).

References

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Symposium On Document Analysis and Information Retrieval*, pages 161–176, University of Nevada, Las Vegas.

Emmanuel Giguet. 1995. Multilingual sentence categorization according to language. In *Proceedings of the European Chapter of the Association for Computational Linguistics SIGDAT Workshop "From text to tags : Issues in Multilingual Language Analysis"*, pages 73–76, Dublin, Ireland, march.

Nadine Lucas, Nishina Kikuko, Akiba Tomoyoshi, and Surech K.G. 1993. Discourse analysis of scientific textbooks in japanese : a tool for producing automatic summaries. Technical Report 93TR-0004, Department of Computer Science, Tokyo Institute of Technology, Meguro-ku Ookayama 2-12-1, Tokyo 152, Japan, March.

Nadine Lucas. 1992. Syntaxe du paragraphe dans les textes scientifiques en japonais et en français. In *Colloque international : Parcours linguistiques de discours spécialisé*, Université Paris III, Septembre.

Jacques Vergne. 1993. Syntactic properties of natural languages and application to automatic parsing. In *SEPLN 93 congress*, Granada, Spain, August. Sociedad Española para el Procesamiento del Lenguaje Natural.

Jacques Vergne. 1994. A non recursive sentence segmentation, applied to parsing of linear complexity in time. In *New Methods in Language Processing*, pages 234–241, June.