Automatic Identification of European Languages

Anna V. Zhdanova[1,2]

[1] A.P. Ershov Institute of Informatics Systems, Novosibirsk 630090, Russia
[2] Novosibirsk State University, Novosibirsk 630090, Russia
anna@sib3.ru

Abstract. We describe our word-based implementation of a language identifying system for the text messages written in European languages. Specifically, we use and compare linguistic (based on functional words) and statistic (based on the word frequency) approaches to construction of the identifying vocabularies. Our version of the statistic approach copes with the differences in degrees of word overlap among languages and the problem of the small-size messages. In addition, it allows an user to choose the accuracy of language identification. At present, our system identifies 8 languages (Bulgarian, English, French, German, Italian, Russian, Spanish and Swedish) in various encodings. With the identifying vocabularies of limited size (less than 1500 keys per language), the accuracy of identification attains 99% even for the messages containing only one sentence.

1 Introduction

Automatic language identification (LI) is the process by which the language of a textual message or of a digitized speech utterance is recognized by a computer. The understanding and developing of the principles which are or may be used to perform LI is of high current interest due to the growing internationalization of Internet. For European international companies, for example, LI is one of the primary steps in solving the problems of electronic message classification and auto-responding, because no single language is spoken throughout Europe. Considering the situation of real life where incoming messages written in the same language may arrive differently encoded, encoding identification is also important, e.g., for redirecting a message in an appropriate form to the subsequent message classification system or to a human operator. Historically, LI of textual messages has been accomplished by using such techniques employing dictionaries or other language sources as (i) a bigram probability-multiplication method [1] and the trigram-based methods [2, 3], (ii) scoring the words with five characters or less [3], and (iii) scoring the grammatical words and frequent endings [4]. The word-based techniques [(ii) and (iii)] seem to be more convenient, because unlike the letter-based technique (i) they are able to provide LI of multilingual messages.

In addition, the methods for classifying documents by language have been developed. They may use less information about languages than the LI methods and are able to group documents into language similarity clusters. The algorithms for classifying documents by language are based on the vector space model of information retrieval and involve various approaches employing, e.g., linear algebra [5].

At present, there exist online implementations of LI systems such as Euclid [6], Lextek Language Identifier [7] and SILC [8]. All these products seem to employ the letter-based methods or n-gram algorithms, as well as the most mentioned above papers. These methods are good for the typical incoming texts, but they usually fail when a text includes relatively large number of foreign words or transliterated proper and geographical names (e.g., such a message as "Anna Zhdanova will go to Stockholm." may be erroneously identified as not being written in English). Unfortunately, the latter situation is common for e-mail messages and Internet in general. Thus, there is a need of effective methods of word-based vocabularies construction. Although the Lextek Language Identifier operates with up to 260 language and encoding modules and provides rather fast recognition, the processing of messages containing less than 200 characters (about 30 words) is however not allowed there. Thus, the problem of LI of small inputs is still open.

In this paper, we describe and compare two word-based methods of construction of the identifying vocabularies. The first conventional method implies involving experts who know languages [4]. The second new method involves corpora and statistics. In the latter case, we take into account the message size a user expects to have. Thus, the LI system's database is optimized according to the user's needs. This makes it possible to reach the maximum performance efficiency. The resulting set of the identifying vocabularies depends on the languages included in a specific configuration of the LI system. This strategy improves the performance of the LI system. In addition, our system allows a user to choose the needed accuracy of LI.

2 Principles of Word-based Approaches to LI

2.1 Construction of identifying vocabularies

Each LI system usually contains a set of identifying vocabularies. In turn, each identifying vocabulary contains a set of key words corresponding to a given language and encoding. During the parsing of an incoming message, the occurrences of key words are calculated for each vocabulary. When this process is completed, the counts for the vocabularies are compared, and the language and encoding of the vocabulary with the largest count (i.e., the vocabulary containing the key words which occurred most often) are assigned to a message. To prevent misidentification of messages exhibiting only a few key words or written in a language the system does not have, one may establish a threshold value for determining a language. If the ratio of the number of the key words found in the vocabulary with the largest count and the total number of words in a message is lower then the threshold value, no language is identified.

For efficient LI, the identifying vocabularies should be representative. On the other hand, their sizes have to be limited in order to enable the system to work quickly. To satisfy these requirements, we have first constructed the sets of identifying vocabularies based on the linguistic approach, similar to the already known method of grammatical (or functional) words [4]. This method is however not suitable in the case when the processed input is small. For the LI system including four European languages, for example, the grammatical word method becomes inefficient when the number of words in the input is less than eight [4]. With increasing the number of languages, the minimum size of the processed input is expected to be larger.

To provide the possibility to work with relatively small messages, we have used the statistic algorithm. The latter algorithm takes into account the expected number of words in a message and the needed accuracy of LI (these two parameters are set by an user) and compiles the corresponding set of the appropriate identifying vocabularies from the given corpora.

Linguistic approach to construction of identifying vocabularies. Usually text consists of sentences. The main idea of the linguistic approach is that at least one word from a typical sentence written in some language should be included in the corresponding identifying vocabulary. As a rule, typical sentences contain determiners, prepositions, pronouns, auxiliary verbs, numerals and some words specific for a particular subject domain. In the framework of the linguistic approach, the identifying vocabularies are constructed manually by the experts who are acquainted with the languages and know the appropriate words. The vocabularies compiled in our group make it possible to operate with six languages including English, French, German, Russian, Spanish and Swedish. The sizes of the vocabularies are different due to the difference in the structure of languages (inflections, cases, etc.).

Statistic approach to construction of identifying vocabularies. The linguistic approach described above requires involving experts in languages. Such experts are not always available. Furthermore, the results obtained depend on experts' educational background, field of expertise, etc. For this reason, there is no guarantee that the corresponding identifying vocabularies are representative. To avoid these complications, we have developed a statistical algorithm for construction of identifying vocabularies. In the framework of this algorithm, the identifying vocabularies for a set of languages are worked out on a set of corresponding corpora. In particular, we have constructed two sets of identifying vocabularies. The first one is for the same six languages that were handled by using the linguistic approach (Sec. 2.1.1). The second one contains six plus two languages, including English (En), French (Fr), German (Ge), Russian (Ru), Spanish (Sp), Swedish (Sw), Bulgarian (Bu) and Italian (It).

The algorithm used depends on the three parameters: n, the expected number of words in a message or a number of words in a single test entry (for testing); a ($a \in [0, 1]$), the needed accuracy of LI; and $\{c_i\}$, a set of corpora (the subscript i

indicates a language). The numbers n and a are expected to be chosen by an user according to his/her needs. He/she should also list the languages which are to be identified. The set of corpora employed now is "standard", i.e., it corresponds to our own choice. In principle, however, the set may be chosen in order to take into account the specifics of the user's field of interests.

To construct the identifying vocabularies corresponding to given values of n and a, it is convenient to introduce the probability P that a word from a typical message written in one of the included languages belongs to the identifying vocabulary of this language. The probability that a word is not found is accordingly equal to $1-P$. For a typical message with the expected size n, the probability that no words belong to the identifying vocabulary equals $(1-P)^n$, provided that the correlations in the occurrence of words are not significant. The probability that at least one of the words belongs to the identifying vocabulary is equal to $1-(1-P)^n$. The latter probability is identified with the accuracy a, i.e., we use $1-(1-P)^n = a$, or

$$(1-P)^n = 1-a. \tag{1}$$

By solving this equation, we can calculate P for given values of n and a.

With the specification above, the algorithm of the choice of the key words for the identifying vocabularies is as follows:

1. For each language, a list of words used in the corresponding corpus is constructed.
2. The frequency f_i^j of the occurrence of word j for corpus i is calculated. By definition, f_i^j is the ratio of the number of occurrence of word j to the total number of words in c_i.
3. The words found in more than one corpus are removed.
4. The remaining words are sorted in each list according to their frequencies in a decreasing order. The top words (with the highest frequencies) are included into the corresponding identifying vocabularies as long as the sum of their frequencies is lower than P.

For this algorithm, the sizes of the identifying vocabularies for different languages are different. Every time a new language is added into the LI system, all the identifying vocabularies have to be rebuilt. With increasing n, the sizes of the identifying vocabularies are rapidly decreasing.

Theoretically, due to the inter-lingual homonyms removal [step 3] and possible shortcomings in corpora, this algorithm may lead to an undesired situation when at step 4 all the words remaining in a list are already added to the identifying vocabulary, but the sum of the word frequencies is still less than P. In such a case, the LI system includes all the words from a given list to the corresponding identifying vocabulary and simultaneously warns an user that it works at risk of not obtaining the accuracy a. With our identifying vocabularies, we had no this problem provided that $a \leq 0.99$ and $n \geq 10$.

2.2 Inter-lingual homonyms and misrepresentative words

For the traditional linguistic approach and the statistic approach with step 3 omitted, some of the key words usually belong to several identifying vocabularies. This may result in mistakes in LI, e.g., when a word is an inter-lingual homonym, but is not included into all the identifying vocabularies which might contain this word. For example, the word "se" is a verb with the meaning "to see" in Swedish and a reflexive pronoun with the possible meaning "himself"/"herself" in French. The Swedish identifying vocabulary does not contain the key word "se", while the French one does. If the LI system encounters the word "se" in a message written in Swedish, it will add a point to the count of French. As a result, this will head towards the wrong direction.

To tackle the problem above, it makes sense to prevent the appearance of inter-lingual homonyms in the identifying vocabularies. In the framework of the statistical approach, this was attained by introducing step 3. For the linguistic approach, we have first compiled extensive dictionaries by using large corpora. If a key word from an identifying vocabulary belonged to at least two extensive dictionaries, it was removed from the identifying vocabularies.

An opportunity of choosing various corpora for constructing identifying vocabularies is an advantage and a flaw of the statistical method. While an user is able to tune the LI system to a better performance for a specific domain, the identifying vocabularies lose their universality. For example, the identifying vocabularies, compiled by employing the corpora related to programming, usually contain words typical for this field. But these words are misrepresentative for other domains. This problem can partly be resolved by combining corpora corresponding to different fields.

2.3 Multiple Encodings

In reality, incoming messages written in the same language may be differently encoded. For the languages based on the Roman alphabet (e.g., English, German), the encoding Cp1252 (i.e., Windows Western Europe / Latin-1) is usually sufficient to cover all the letters including the "substandard" ones (e.g., Spanish "ñ" or German "ß") and diacritics. But for Russian, Bulgarian, Ukrainian and other languages based on the Cyrillic alphabet, a few encodings such as Cp1251, Cp866 and koi8-r are widespread. For this reason, the key words from the identifying vocabularies for these languages should be written in all the encodings used in real messages.

3 Results

The system performing LI is written in the Java language in accordance with the principles of the object-oriented programming. The language identifier takes a file or a list of files as an input, makes the parsing by its lexical analyzer, matches the picked out words to the identifying vocabularies, calculates the score for every possible language and encoding pair, and after all, returns the language and encoding with the highest score.

Our LI system makes it possible to construct and add new identifying vocabularies for different languages and encodings. To study its performance, we have constructed several sets of identifying vocabularies. In addition, we have formed a database of test entries modeling typical incoming messages. The latter database consists of texts representing the banking, insurance, computer semantic domains and fiction literature. Most texts were extracted from the appropriate web pages and used with no preprocessing except converting to the plain text files by removing tags. Some texts were found to contain foreign words and phrases. While the presence of such texts in the database employed to form test messages is a model of real situation, their presence in the corpora created for construction of the identifying vocabularies resulted in the appearance of the "false" inter-lingual homonyms and eventually slightly deteriorated the identifying vocabularies.

In reality, incoming messages usually contain several sentences. Our LI system processes a whole message as an input. To test the system, we used relatively small entries consisting of a single sentence. A sentence was defined to be the characters between the punctuation signs ".", "?", "!" and the symbols signifying the beginning and end of text files.

3.1 Linguistic approach

In the framework of the linguistic approach, the identifying vocabularies are constructed manually (Sec. 2.1.1). Table 1 shows the sizes of such vocabularies before and after removal of inter-lingual homonyms. The difference between the upper and lower rows is proportional to the degree of overlap among the languages.

Table 1. Sizes (number of words) of the linguistically constructed identifying vocabularies before and after removal of inter-lingual homonyms.

	English	French	German	Russian	Spanish	Swedish
Before removal	581	821	250	1053	181	254
After removal	444	642	210	1053	40	213

To test the identifying vocabularies, the LI system determines a language and encoding for each test entry and compares them to the language and encoding the entry is written in. Table 2 shows the percentage of the correctly identified test entries for the linguistically constructed identifying vocabularies. Spanish and French possessing the highest number of widespread inter- lingual homonyms are seen to have the lowest identification accuracy. This is in agreement with the study [5] of classification of documents by language, where Spanish is mentioned as the most commonly misclassified language.

Table 2. Percentage of correctly identified test entries before and after removal of inter-lingual homonyms for the linguistic approach to construction of the identifying vocabularies.

	English	French	German	Russian	Spanish	Swedish
Before removal	99.0	94.2	98.8	95.7	97.9	98.1
After removal	99.9	90.5	99.7	99.8	70.3	99.9

The results presented in Table 2 indicate that for the linguistic approach the removal of inter-lingual homonyms enlarges the number of unidentified entries. In other words, the removal of homonyms from the manually constructed identifying vocabularies neither improves the performance of the LI system nor solves the problem of small incoming messages.

3.2 Statistic approach

Using the statistic approach, we have constructed and tested four sets of identifying vocabularies. Table 3 demonstrates their sizes [with and without step 3] in the cases of six and six plus two languages. The corpora employed contained 30000 words for each language. The accuracy of LI and the expected size of entries were $a = 0.99$ and $n = 10$, respectively. For these parameters, Eq. (1) yields $P = 0.3691$. Thus, we have filled the identifying vocabularies in such a way that the sum of the word frequencies within every vocabulary is not lower than 0.3691.

Table 3. Sizes (number of words) of the identifying vocabularies (for six and eight languages, $a = 0.99$ and $n = 10$) constructed by employing the statistic algorithm with and without step 3.

Number of languages	En	Fr	Ge	Ru	Sp	Sw	Bu	It
6 (without step 3)	24	34	69	97	33	69	-	-
6 (with step 3)	146	527	270	107	433	173	-	-
8 (without step 3)	24	34	69	97	33	69	91	42
8 (with step 3)	146	995	248	599	714	160	742	1427

The results of the tests, presented in Table 4, indicate that the expected LI accuracy, $a = 0.99$, is obtained when the complete (with step 3) statistical algorithm is used. Adding Italian is seen to affect slightly the identification of the languages, based on the Roman alphabet, mainly due to the imperfect corpora and misrepresentative words. While adding Bulgarian produces an appreciable increase in the number of misidentifications for Russian, based also on the Cyrillic alphabet.

Table 4. Percentage of correctly identified test entries before and after removal of interlingual homonyms for the statistic approach to construction of identifying vocabularies.

Number of languages	En	Fr	Ge	Ru	Sp	Sw	Bu	It
6 (without step 3)	95.3	91.2	99.5	98.9	98.8	98.7	-	-
6 (with step 3)	98.7	99.1	100.0	99.3	99.3	99.8	-	-
8 (without step 3)	94.5	86.6	99.3	49.3	94.4	97.3	96.3	94.5
8 (with step 3)	97.8	98.5	99.2	89.6	98.1	99.9	98.8	98.4

4 Conclusion

Using the word-based approach, we have worked out the LI system for European languages. Specifically, several alternative sets of identifying vocabularies have been constructed by employing the linguistic and statistic methods. The latter method is found to be superior, because it allows one to construct the identifying vocabularies taking into account the chosen expected message size and LI accuracy. The sizes of the identifying vocabularies are minimized according to the special formula [Eq. (1)], and this feature provides fast (and accurate) performance of the LI system. For 12 language-encoding pairs recognized by the LI system, the speed of processing is around 2 seconds per megabyte of text and varies from 0.26 to 0.57 milliseconds per sentence depending on the language (for Intel Celeron, 434.31 MHz). Furthermore, the statistic method makes it possible to add a new language to the LI system without any linguistic knowledge and/or employing additional tools. However, this method needs corpora. The identifying vocabularies are optimal when the corpora represent an user field of interest. Thus, the work of our LI system is highly efficient when tuned to the needs of a particular user.

In comparison with the existing LI systems, our LI system with statistically constructed identifying vocabularies has an advantage of tuning its performance. For this reason, it allows one to operate with messages of smaller size or, alternatively, enlarge the speed of LI when the "small-message" feature is not needed. For the messages containing only one sentence, the accuracy of LI approaching to 99% is achieved independently on the degrees of overlap among the languages and the number of languages included into the LI system.

Acknowledgements

The author thanks F. Dinenberg, A. Kleymenov, I. Kononenko, D. Levin, P. Mankevich, D. Petunin, A. Semenov and D. Shishkin for participation in the project and useful discussions.

References

1. Beesley, K.R.: Language Identifier: A Computer Program for Automatic Natural-Language Identification of On-Line Text. In: Languages at Crossroads; Proceedings

of the 29-th Annual Conference of the American Translators Association (1988) 47-54.

2. Schmitt, J.C.: Trigram-based Method of Language Identification. US Patent 5,062,143 (1991).
3. Grefenstette, G.: Comparing Two Language Identification Schemes. In: Proceedings of 3-rd International Conference on Statistical Analysis of Textual Data (1995).
4. Giguet, E.: Categorization According to Language: A Step Toward Combining Linguistic Knowledge and Statistic Learning. In: Proceedings of the International Workshop on Parsing Technologies (1995).
5. Mather, L.: A Linear Algebra Approach to Language Identification. In: Proceedings of the 4-th International Workshop on Principles of Digital Document Processing (1998) 92-103.
6. Euclid: Encoding and Language Identification.
   http://www.basistech.com/products/text-processing/euclid.html (2002).
7. Lextek Language Identifier. http://www.languageidentifier.com (2001).
8. Système d'Identification de la Langue et du Codage.
   http://www-rali.iro.umontreal.ca/SILC/SILC.en.cgi (2002).