Language Identification Using Global Statistics of Natural Languages

Gergely Windisch, László Csink

Budapest Tech, John von Neumann Faculty of Informatics
Institute of Software Technology
Nagyszombat u. 19, H-1034 Budapest, Hungary
Phone: +36 1 3689840, Fax: +36 1 3689632
winger@freemail.hu, csink.laszlo@nik.bmf.hu

Abstract: This article is about a new method which makes it possible to identify the language of a written document. The method is based on the analysis of simple descriptive statistics of the given text. These simple statistical features include things like average word length or consonant congestion.
In order to measure the effectiveness of the method an application has been developed which can classify English, Hungarian, German, Spanish, Croatian, French and Norwegian documents by analysing the average word length, the ratio of certain characters, word endings and consonant congestion.

Keywords: Natural languages, language identification

1 Possible Use of Language Identification

Identifying the language of a text is useful in a wide range of applications. Due to the nature of its task a language identifier (LI) is usually a module within a larger application.

In an office program suite, and especially in a word processor the LI can be used to automatically set the language of the whole document or even the languages of the individual paragraphs.

An LI is also useful for translators. When they receive a large number of documents an LI-based program can sort them out and send them to the person who does the translations for that language. It is even more important in automated translator applications where setting the input language is the only manual intervention necessary before translating the documents, so an LI is essential for an automatic system. Of course, this is hypothetical because currently the automated translator applications are not powerful enough to do real work, especially for languages beside English, but it is a rapidly improving field and in

the future it will probably be available.

An LI can also be used in web browsers to pre-filter the search results so that the users do not have to browse through thousands of pages that they do not understand.

2 Known Solutions

2.1 N-gram Analysis [2]

An n-gram is an n character long substring of a string. The characters must be consecutive within the given word (a slightly different interpretation of an n-gram can also be found in the literature which states that the letters from a given word can be taken out in an arbitrary order to build an n-gram). In this method the words in the document are decomposed into overlapping substrings and then it is possible to calculate the probability of the languages using the number of occurrences of these substrings. It is possible to create many n-grams from one word due to the overlapping and the fact that usually a number of spaces are appended before and after the word in order to increase the number of these sub strings.

For example the decomposition of the word 'apple' gives the following result:

bigrams: _a, ap, pp, pl, le, e_

trigrams: __a, _ap, app, ppl, ple, le_, e__

quadgrams: ___a, __ap, _app, appl, pple, ple_, le__, e___

Generally speaking it can be said that from an n character long word n+1 bigrams, n+2 trigrams and n+3 quadgrams can be created. The method is based on the assumption that within a language some words are used more frequently than others, and that this holds for the substrings created from those words. The probabilities of each letter groups are directly proportional to the probability of the word those groups were decomposed from [5].

2.2 Short Word Method [3]

The short word method (also known as small word method) is similar to the n-gram analysis with the exception that here whole words are being used for identification. As the name suggests, these words are short, usually less than four letters – adjectives, prepositions – which appear in every document of a given language.

2.3 Bayes Method [4]

In this method conditional probability is used to identify the language. In the teaching phase long documents are used to create a database with as many words as possible and the probabilities of each of them appearing in different languages. Later on, when a text of unknown language has to be analysed a probability is calculated for each language based on the words appearing in the document using the Bayes theorem (hence the name).

2.4 Other Possible Methods

There are other ways to identify the language of a given text apart from these three. Research is being undertaken to use, for example, hidden Markov models for this task [4]. Nowadays the n-gram analysis is the most frequently used method due to its high accuracy.

2.5 Analysing Statistical Features

In this project a different approach has been selected. This new method uses simple statistical features of the document instead of the document itself. As no mention had been found in the literature about any system using this idea, finding those features and setting their parameters by thorough experimenting was undertaken by the authors.

Simple descriptive statistical features like word endings and consonant congestion are analysed and, to keep the method simple and fast, complex mathematical statistical calculations are avoided. In this research seven languages had been analysed, namely English, Hungarian, German, Spanish, Croatian, French and Norwegian.

3 Features

The most important part of the work was finding features that can be used to identify the language. An extensive examination of different documents resulted in five features that can be used for identifying the language of a text. These features are:

- Character occurrence frequency,
- Word ending,
- Ratio of the number of appearances of any two characters,
- Consonant congestion,
- Average word length.

3.1 Character Occurrence Frequency

It shows how often is a certain letter used in a text. For example:

"**Th**is was one of t**h**e many t**h**ings t**h**at Art**h**ur **h**ad told t**h**e ent**h**usiastic little man"[1].

There are 66 letters in this sentence, eight of which is 'h' so the relative frequency of this letter is 0.12 that is 12%.

Not all of the letters can be used for language identification. Looking for special characters like 'é' or 'á' that are not part of every language is not advised as it might give false results with quotes, foreign expressions or jargon words, so only letters of the standard English alphabet are used. And even from those 26 letters some are better then the others. A letter is called "good" – meaning that it can be used to classify languages – if its occurrences in multiple sample texts follow a uniform distribution. In an ideal case, the average of the relative frequency of a letter is different for each language and the variance is zero. In this case this letter would be able to separate every language from the others. Plotting such a letter on a diagram similar to that of *Figure 1* would result in a straight graph.

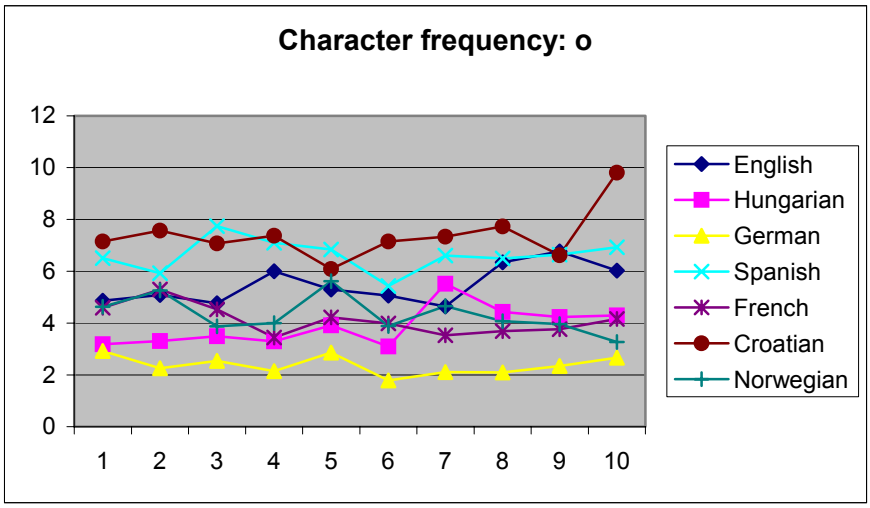

Figure 1
Character frequency: o[2]

1 Douglas Adams: Life, The Universe and Everything, ch. 22, p. 157. ISBN: 0-345-39182-9

2 A diagram like this shows the values of a given feature in ten sample documents for each languages. One graph shows one language. The sample documents are shown on the X axis. The Y axis is the percentage of the occurrence of the given feature

Unfortunately, this never happens in real life. The letter “o” shown in *Figure 1* is among the best but it can be seen that the variance is different from zero and the graphs do overlap a little. On *Figure 2* the letter p is shown. That is an example for a letter that cannot be used for language identification. The reason for this is obvious from the graphs – their variances are big, the averages are close to each other and there are no visible layers as can be seen on *Figure 1*. The letters can be divided into two categories based on their goodness. Characters that can separate languages go to the first group, and the rest go the second group. There are two types of good letters. For a letter of the first type each language has a distinct value of occurrence in the samples. A good example is the letter “o” in *Figure 1* where the layers are apparent. These letters usually divide the languages into groups of two to four. For example, if in an unknown document the frequency of ‘o’ was 4%, then the document had probably been written either in English, Hungarian or Norwegian, but it is definitely not Croatian. Of course, there is no way to tell which one, but that is why more of these letters have to be analysed as they may create different groups with different members, and if the teaching of the system is good than one of the languages should appear in all of these groups.

The other letters among the good ones are those that give a result for one language only but if that comes out the language of the text is bound to be that one. An example of this is the letter ‘j’. In each of the languages examined during this project the frequency of this letter is between one to two percent except in Croatian documents where it is above 10%. If more than five percent of the letters in an unknown text is ‘j’ it means that there is a really high chance it was written in Croatian.

Letters that cannot be used for identification also have two groups and the categorization is based on the reason why they cannot be used. One of the possibilities is shown on *Figure 2* where it can be seen that within the same language the frequency of the letter ‘p’ varies a great deal, and that is true for most languages. The other reason why a character has to be omitted is if it occurs so rarely that its value is impossible to measure. An example for this is the letter ‘q’ with a frequency of only 0.001 (0.1 percent) in any of the seven languages.

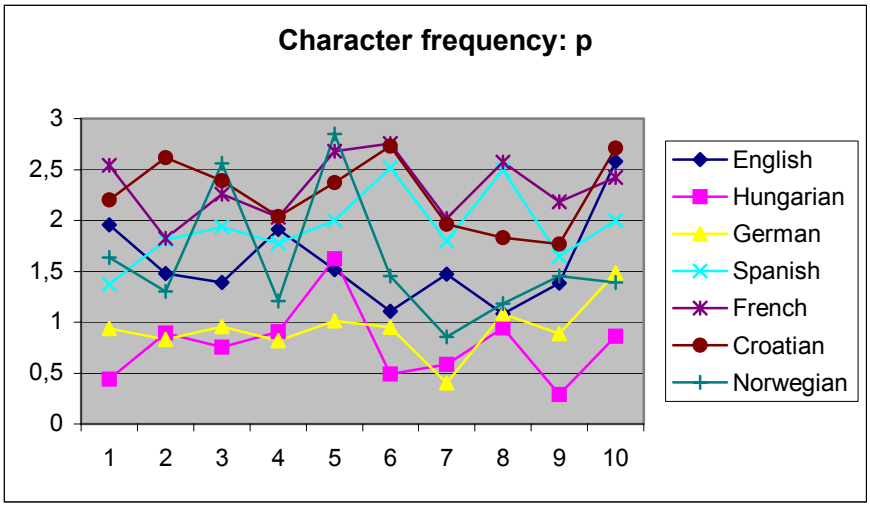

Figure 2
Character frequency: p

3.2 Word Ending

The feature called word ending shows for all words of the document how many of them end in a certain letter. For example:

"Az ott Gandalf, Théoden és az emberei! – Mondta Legolas. – Gyerün**k**, menjün**k** elébü**k**."[3]

Of the 12 words three end in a 'k', which is 25% of the words.

The categorization is the same as that of the character frequency, which has already been discussed in the article. Of these two features word ending was found much more reliable and usable than character frequency. About half of the letters can be used for identification and some of them are close to the previously described ideal case. The reason for this is that certain suffixes are appended to the end of the words – for example the letter 'd' to past tense of English verbs or 's' to plural nouns and these suffixes are typical for the languages.

3.3 Character Ratio

In this feature the number of occurrences of two letters are compared to each other

3 J. R. R. Tolkien: Lord of the Rings – The Two Towers (Hungarian). ch. 10, p. 222. Translated By Árpád Göncz ISBN: 963 07 7050 4

and a ratio is given based on their occurrences.

"A l**a**nt**e**rn th**a**t do**e**sn't shin**e** for **a** m**a**n th**a**t do**e**sn't s**ee**?"[4]

The quoted sentence contains six 'a'-s and six 'e'-s. That gives 6/6=1 for the ratio of a/e.

Comparing the frequencies of two characters gives good results especially for two well-chosen letters. It is not very good for separating languages where the ratios are close to each other (like 0.6 for one language and 0.7 for the other one) but it is really useful for deciding between two languages where one has ratios like 0.6 and and 1.3, for instance. The only language which uses more 'a'-s than 'e'-s is Croatian so the a/e ratio can identify Croatian texts with a high certainty whereas neither the frequency of 'a' nor of 'e' is especially useful.

3.4 Consonant Congregation

This feature shows how often one, two or three consonants stand between two vowels in a word.

"I**ch** bin Geo**rg**e Mi**lt**on. Dies is Le**nn**ie **Kl**ein."[5]

There are 35 letters in this quotation; 20 of which are consonants. 10 of them are standing in pairs, which gives five pairs. These five pairs are 50% of the total number of consonants.

The original idea behind the consonant congregation was that it would be able to identify Slavic languages by searching for congregations of 3 letters or more. This assumption turned out to be wrong, those are not nearly as common as would be needed for classification, however; the single and double consonants do separate the languages very well. A single consonant is frequent in Spanish and French words (about 60% of the total population stand by themselves), whereas in German documents this is as low as 38 percent. Having a pair of consonants is less frequent, it varies from 10 to 20% depending on the language but the values are near constant for a given language, so this feature is useful.

3.5 Additional Features

It would be interesting to experiment with other features as well. One such feature could be the ratio of the frequency of a letter and the percentage it appears at the end of words. It could lead to good results when a certain letter does not occur often enough to be tested, but all of these occurrences are last characters of a

4 Terry Pratchett: Small Gods p. 183. ISBN: 0 552 13890 8

5 John Steinbeck: Of Mice and Man (German) – ch. 2. p. 74. Translated by: Elisabeth Rotten ISBN:N/A

word; then this ratio could actually be used.

Another possible feature is sentence length, although research has shown that it is not nearly as good as the ones mentioned before, as the length of a sentence varies within the language and also separating the sentences is not a trivial task, so it should be used with caution.

3.6 Final Features

Experimenting with these features yielded the following results. It turned out that it is important to find the optimum number of features used to identify the language. Using too few or too many would result in a degradation of the effectiveness of the identification. This is the reason why character frequency has been dropped altogether as it had proved much less effective than word endings. Analysing the last characters of words is so powerful that it alone could be used to identify the language of a given document. The character ratio feature has also proven effective. The system that had been created to test the method uses s/t and a/e ratios. The a/e ratio is especially useful for Croatian texts. Of the consonant congestion, single and double letters are used. Average word length is also a good feature. Results show that of the languages experimented with French words are the shortest (an average of 4.74 characters), and Hungarians use the longest words with an average of 6.13 characters.

According to the above, the system uses the following features:

- Word endings: a, d, e, i, j, n, o, s, t, u
- Average word length
- Character ratios: A/E, S/T
- Consonant congregation: 1, 2, 3

4 Testing

With the help of the language identifier system built for this purpose, the method underwent a series of heavy testing. The texts used for the testing were mostly books downloaded from public libraries on the internet[6]. The books were mainly novels from different authors including Harry Potter by J.K. Rowling, Lord of the Rings by J. R. R. Tolkien and Small Gods by Terry Pratchett. Besides these novels the Bible and a number of Wikipedia[7] articles have also been used. Care has been

6 Hungarian Online Library - http://mek.oszk.hu/
7 http://en.wikipedia.org/wiki/Main_Page

taken to use the same works translated to different languages in order to avoid getting false results due to the differences in style. Unfortunately, Croatian and Norwegian documents had not been available for download in the appropriate amount so these languages have been left out of the testing.

To determine how much the method relies on the length of the text, the sample files containing the books had been cut into smaller pieces.

Altogether the tests were run on about 5000 separate files for each language. Each of these files contained a part of a document. These files consisted of

- short texts – about a 100 characters (a line of text)
- middle length texts – about 500 characters (a paragraph)
- long texts – about 1500 characters (a page)

The results of the tests can be seen on *Figures 3-5*.

4.1 Short Texts

Looking at the table describing the results of the identification of short texts, it is apparent that the method is not effective here. The best identification results came with the Spanish texts, where correct identification was achieved in 81% of the documents, which is not too bad. The results of analysing the French documents, however; are much worse with only 16% of the documents correctly identified.

This result is not at all surprising because the method uses statistical analysis which relies greatly on large samples and a 100 characters long sample is not large enough. Considering the fact that the system mainly uses the last letters of the words to identify the language makes the size of the sample even smaller.

It can be stated that this method cannot be used for such short documents.

Short Texts (<100 characters)[8]

Original Language	**English**	**German**	**French**	**Hungarian**	**Spanish**
Points	English:	German:	Spanish:	Hungarian:	Spanish:
	1887	2249	1315	2787	8061
	Spanish:	English:	English:	French:	French:
	392	1113	891	365	597
	Hungarian:	Hungarian:	French:	Spanish:	Hungarian:
	311	1106	645	321	561
	German:	Spanish:	Hungarian:	English:	English:
	304	796	700	318	445
	French:	French:	German:	German:	German:
	198	207	262	208	295
Summary	3094	5473	3815	4001	9961
Percentage:	60.989%	41.093%	16.907%	69.658%	80.926%

Figure 3
Test results for short texts

4.2 Mid-sized Texts

The accuracy of the identification improves greatly with the increase of the document size. When ran on 500 characters long texts, the system was able to identify about 80% of the documents (average value, for details see *Figure 4*). The identification of French language documents improved the most, from 16% it went up to 62%.

According to these results the method can be used for identifying texts that are at least 500 characters.

8 The title of the table tells the category. The Original Languge row means to which language the document belongs. In the Points section the actual results of the identifications are shown. The Summary row shows how many files there were for a language and below that the values of the Percentage row shows how effective the identification was (e. g. the percentage of the correctly identified documents). For example, from the first column it is apparent that there had been 3094 English language texts, 1887 of which were identified correctly as English, 392 were misidentified as Spanish and so on.

Mid-sized texts (~500 characters)

Original Language	**English**	**German**	**French**	**Hungarian**	**Spanish**
Points	English:	German:	French:	Hungarian:	Spanish:
	3239	4402	2115	3068	3953
	German:	English:	Spanish:	French:	French:
	300	822	751	69	57
	Hungarian:	Hungarian:	Hungarian:	English:	Hungarian:
	205	137	229	34	41
	French:	Spanish:	English:	Spanish:	English:
	133	89	206	28	22
	Spanish:	French:	German:	German:	German:
	57	54	72	6	22
Summary	3936	5506	3383	3207	4097
Percentage:	82.292%	79.949%	62.518%	95.666%	96.485%

Figure 4
Test results for mid-sized texts

With a further increase of the document size, the accuracy improves even more. For documents that are about a page long the worst accuracy is more than 92% while the best is 99.5%.

4.3 Long Texts

For documents with 1500 characters or more this method can identify the language with appropriate certainty.

What is interesting when looking at the tables globally is the lack of symmetry of the results. For example, 35% of the 500 characters long French texts have been misidentified as Spanish whereas only 1% of the Spanish have been misidentified as French. The reason for this is yet unknown, further testing is needed to decide whether there is a linguistic explanation behind this phenomenon or it is due to an error made during the teaching phase of the system.

Long texts (>1500 characters)

Original Language	**English**	**German**	**French**	**Hungarian**	**Spanish**
Points:	English:	German:	French:	Hungarian:	Spanish:
	543	992	557	591	1359
	German:	English:	Spanish:	English:	French:
	16	78	23	1	2
	Hungarian:	Hungarian:	English:	German:	German:
	3	2	4	1	2
	French:	Spanish:	Hungarian:	Spanish:	Hungarian:
	2	2	1	1	1
	Spanish:	Croatian:	Croatian:	Croatian:	English:
	1	1	1	1	1
Summary	567	1077	588	597	1367
Percentage:	95.767%	92.108%	94.728%	98.995%	99.415%

Figure 5
Test results for long texts

5 Comparison with Similar Methods

A comparison had been made between the system created for the statistical language identification project and the language identifier of Peter Bauer [1]. Bauer's application uses n-gram analysis. The test results show that the n-gram method is in fact more accurate on shorter texts; it reaches 99% of accuracy with only a hundred characters. That makes it more suitable for applications where long texts are not available.

As the length grows, the accuracy of the statistical analysis method catches up with the n-gram method. Another important parameter of such a system is their speed. When ran on the same computer on the same documents the n-gram analysis takes about 0.6 seconds whereas the statistical analysis takes only 0.05 seconds, which means that it is about ten times faster making it ideal for applications where speed is a critical issue and the documents are not shorter than a page (typically a server side language identifier is like that). This speed difference comes from the different workflows of the systems. For an n-gram analysis, every word in the document has to be decomposed into 3*(n+2) sub strings and their number of occurrences have to be compared to a few thousand of known sub strings whereas for the statistical analysis only 5 features have to be

measured and then their values compared to about 15 parameters.

Conclusions

This project involves the development of a language identifier system that can classify English, Hungarian, German, Spanish, French, Norwegian and Croatian documents that are at least 500 characters long. The system analyses statistical features like word ending, average word length, character ratio and consonant congestion.

Test results indicate that the system can identify the language of a page long text correctly with about 96% likelyhood. Recognition is even better with longer documents.

The test results prove that the idea behind this method is correct; that it can be used for language identification.

All the features and their parameters are the results of the authors own research work. With a wide array of experimenting and testing, the parameters have been fine tuned to do the language identification correctly. As of this moment the system can only identify the language of a document but we plan further research for author or style identification based on statistics.

References

[1] Bauer, P.: Természetes nyelvek statisztikai elemzése (Statistical analysis of natural languages) 2003, Thesis – Budapest Tech

[2] Cavnar, W. B. és Trenkle, J. M. (1994). N-Gram-Based Text Categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US: 161-175

[3] Grefenstette, G. (1995). Comparing two Language Identification Schemes. In: Proceedings of the 3rd International Conference on the the Statistical Analysis of Textual Data (JADT'95), Rome, Italy

[4] Murray, I. A. (2002). Probabilistic Language Modelling. http://www.inference.phy.cam.ac.uk/is/papers/ [last checked 20 April, 2005]

[5] Zipf, G. K., Human Behavior and the Principle of Least Effort, an Introduction to Human Ecology, Addison-Wesley, Reading, Mass., 1949