Natural languages and the World Wide Web

Stefan Langer
slanger@elexir.de

ELEXIR GmbH
(Subsidiary of Fast Search and Transfer ASA)

Abstract
In this article I will try to demonstrate the importance of language specific processing on the world wide web. The starting point is the description of language identification carried out on the search engine AllTheWeb and a quantitative importance 44 languages on the Internet. The main possibilities for improving Internet search services through language specific processing are listed in the last section.

1 Introduction

In order to determine the needs for language specific processing it is helpful to know the distribution and relative importance of languages on the web. Thus, the starting point of this paper on languages on the web is the exemplary presentation of the functionality and the performance of the Elexir language identifier used in the index of the AllTheWeb search engine (www.AllTheWeb.com), followed by a presentation of a quantitative analysis of the AllTheWeb index. With the help of the identifier I determined the percentages of 44 languages on the Internet. The basis is an evaluation of the index in summer 1999 and spring 2001. The tables also present these figures and their relation to the total number of speakers for the different languages - which gives an impression of the Internet pervasion for the language communities.
In the last section I describe necessary steps to make search services on the web multilingual.

2 Language identification

2.1 Definition

A language identifier in the sense this notion is used in this paper is a piece of software for the automatic recognition of the language of an electronic text document. In particular, I do not refer to identification of spoken language.

2.2 Application areas

Language identifiers for electronic documents are mainly used in Internet search engines, Intranet search and multilingual text archives.
All big international Internet search engines have language support - users are able to restrict their search to documents written in one or several preselected language(s). The number of languages supported vary between the different search engines. Here the numbers for some more popular interfaces:

Google	25
Excite	11
Altavista	19
AllTheWeb	44

(Numbers from early march 2001)

Html-documents can provide language and character encoding information, and the use of tags to indicate the language is recommended by the W3C-consortium. However, classification of documents for Internet search engines is not (or at least not primarily) carried out by using meta information in html-documents (either language or character encoding tags), because this information is often missing or unreliable. The language is determined by means of an automatic language identifier that processes the document text.
Other application areas for language identification tools are multilingual text archives and the intranet of large companies. In linguistic research language identifiers have been used for the automatic building of corpora from the WWW or other multilingual text collections (Cowie et al 1998).

2.2.1 Methods and algorithms

There are in principle two different techniques for the automatic identification of the language of a text document (for a more detailed performance comparison between the two approaches s. Grefenstette 1995). The two identification types are the word based language identification on the one hand and the N-gram based identification on the other.

Word based methods

- Words found in the document are compared with frequency list of words for the supported languages, and matches are counted. The language with the highest match score, exceeding a predefined threshold, is identified.
- Word lists are compiled from a corpus or constructed from available electronic dictionary resources. Word list have to be cleaned - esp. they should not contain too many internationalisms or, in the case of the application to web search, international Internet related words (such as "html" etc.).
- The word based technique is only applicable, if words are marked in the text. It cannot be used for languages that do not systematically mark word boundaries by

blanks or punctuation marks, such as Japanese, Chinese and Korean. To tokenize a text chunk in these languages, the language has to be known beforehand.

- The size of the word list depends on the performance requirements and on the document type. Longer documents can be identified even with lists of the most frequent word forms (Grefenstette 1995). Very short documents are difficult to classify. The appropriate length of the word list also depends on the morphological system of a language. Languages with many word forms need longer word lists in order to achieve similar recall values. The word lists for the different languages should cover roughly the same percentage of the training corpora.
- Very short words (esp. words consisting of one character) should have a decreased weight or should not be counted at all.
- An advantage of the word based technique is that sources of errors can be easily detected as long as they are related to the dictionaries - word lists can be easily controlled manually.

N-gram based methods (e.g. Cavnar/Trenkle 1994),

- N-grams (sequences of N units) of characters or bytes of text are compared with statistics build on the training corpora. In most cases statistics on trigrams (sequences of three bytes) are used.
- Contrary to word based methods, this technique can be used for any written language, also for such that do not mark word boundaries (Chinese, Japanese, Thai, Korean).
- Like the word based technique, this method is problematic for short documents.
- Sources of errors in the statistics can not be detected easily.

Both techniques for language identification require a parser - an extractor for natural language sequences for natural language sequences. In the case of web applications, an html-parser is required.

2.3 The Elexir language identifier

The language identifier of AllTheWeb, which was used for this study, is a hybrid identifier - it uses both techniques named in the previous section. First, the word based method is used to identify languages that mark word boundaries. For this purpose, word form lists were built for 40 languages. The average length of this word lists is 5000 words. The identifier parses the text of a document and compare the words in the document with the word form lists. The score of a language for a documents depends on the following parameters:

- overall number of words in the document
- number of identified words for the language
- frequency of the identified words in the language
- frequency of word forms in the document (upper threshold)
- length of identified words

If the identification based on these word lists does not lead to a reliable result, a bigram based algorithm is run to try to identify one of the languages that do not mark word boundaries (Japanese, Chinese, Korean and Thai).

Language identification is closely related to the identification of the character encoding (often referred to as character set). If the encoding of a document is not known beforehand, the language identifier needs to work for all possible encodings. This becomes evident when considering cases like Russian in Cyrillic script, where at least 5 encodings commonly appear on the web (ISO-8859-5, Windows Cyrillic code page, KOI-8-R, Dos Cyrillic code page and Mac Cyrillic). To identify a document as Russian, all possible words in the documents have to be compared to word lists in any of these encodings. Apparently, this means that the encoding of the words in the document is detected simultaneously with the language. Word lists for the Elexir language identifier are constructed an maintained in one encoding and then converted to any encoding common for the specific language.
In the case of bigram statistics for Asian languages, no conversion is carried out. These statistics are not constructed on languages only but rather on corpora in one single language and encoding.

In general, language identifiers work very reliably for long an medium sized documents. For documents with more than 50 bytes of text, the presented identifier works with a recall of ca 96% and precision of near 100%. This performance decreases dramatically for very short documents. This does, however, not affect the performance for a web search engine very much, because queries are normally carried out with search words, and very short documents are very unlikely to contain the search word. The problematic cases can be detected in the AllTheWeb index when searching for a specific language without any query terms.
Misclassified pages on AllTheWeb are mostly framesets and other pages containing almost no text. It seems difficult to raise precision for these cases. An immediate solution would be not to classify short pages at all.
Recall for long pages could be easily raised above 96%, but in the case of Internet search engines, precision is much more important, as languages are not equally distributed on the WWW. Even a very low percentage of documents of the whole index that erroneously get a tag for a very rare language (such as Faeroese) will lead to a high percentage of misclassified documents for that language.

3 Language statistics: evaluation set-up

The page counts for the different languages that are presented in this paper were carried out with the identifier presented in the previous section. Both counts (1999 and 2001) took the index of AllTheWeb as a basis.

Web indexes can be taken as representative for the whole of the WWW, because they preferably index linked pages. In this way they represent the web in the form it is also perceived by Internet users - who often use search engines to access web pages, and who almost exclusively find linked sites. The aleatory access method presented in

O'Neill/McClain/Lavoie (1997), who generate random IP-numbers of web servers, is more objective, but also covers web sites that are not really accessible for users.

The evaluation for the 1999 index, containing ca 150 Mio. pages, was carried out off line. The primary language of three million web sites was identified for 3 Mio. documents with more than 10 words. Figures for the entire index were then estimated based on these counts.
Later in 1999, the language identifier was integrated into the AllTheWeb search engine. Languages can be selected on the advanced search interface. Based on this index, I could determine the figures for the march 2001 directly.

3.1 Related work

Lavoie and O'Neill (1999) present an evaluation of 1257 (1998) respectively 2229 (1999) web sites, which were selected by a random generation of possible IP-numbers. Countries were determined manually, and languages semi-automatic. The language statistics for 1999 show differences compared with our figures, but they are hardly comparable because of the different experimental settings (much smaller sample, web sites and not web pages, manual classification).

Grefenstette/Nioche (2000) use the frequency of common words in 32 languages in the AltaVista search index to determine the overall word count for these languages. They come to results that are very similar to the figures presented in this paper for the languages that appear in both statistics.

4 Quantitative distribution of languages on the web

The tables present the results of the study. Table 1 presents the percentages of the supported languages in the off-line index from summer 1999 and the figures for march 2001. In both studies, the language of web pages could be identified in 96% of all cases. The rate of non-classified pages was ca 6%. A sample testing of non classified pages gave the following reasons:

- the document was not classifiable - no language could be assigned manually (directory listing, data garbage, name lists)
- The page was multilingual
- The document only contained few, infrequent words in a language
- The document was written in an unsupported language

For most documents, a single language could be identified. As known beforehand, English showed out to be predominant in both counts with ca 2/3 of pages. The other top positions are not surprising either. Mainly languages from highly industrialized countries (German, Japanese) or countries with a very high population (China) occupy these positions.

The last positions are held by languages with very few speakers (Faeroese), minority languages (Galician) or languages from countries with a very low degree of industrialization (Byelorussian).

Table 2 shows the ration between the number of web pages in the index and the speaker counts for the languages supported. It gives an impression of the importance of the internet for the different speaker communities. The analysis is based on a number of 300 web pages and the speaker counts in Grimes (1997) - necessarily these figures are only estimates for most of the languages.
In this statistics, the Scandinavian countries occupy the leading positions after English (the leading position of English is very much dependent on not counting second language speakers in India and some other countries, where English is one of the official languages).

In both statistics, there is a striking low relevance of Arabic - apart from obvious socio-cultural reasons, the fact that many web browsers did not support Arabic until recently, might be a reason for that. The 2001 counts show that the percentage of Arabic web pages as grown by a factor of 8 - but still this percentage is very low compared with the number of speaker. This disproportions are even more dramatic for languages that do not appear in the statistics such as Hindi or Bengali - there was not enough training material available in 1999 to build up satisfying word lists.

4.1 Table 1: languages in % of all evaluated web pages

(base: index AllTheWeb, summer 1999; index AllTheWeb, 5. march 2001
Ranking based on 1999 counts; some dramatic changes marked bold)

	Language	1999	2001
1	English	64,55	60,75
2	German	4,92	**6,09**
3	Japanese	5,94	**5,19**
4	Chinese	2,28	**3,94**
5	French	3,08	3,04
6	Spanish	2,25	2,68
7	Russian	1,58	1,89
8	Italian	1,66	1,59
9	Korean	0,97	1,45
10	Portuguese	1,11	1,39
11	Dutch	0,93	1,03
12	Swedish	1,03	0,82
13	Polish	0,41	0,54
14	Czech	0,58	0,50
15	Danish	0,35	0,46
16	Finnish	0,54	0,41
17	Norwegian	0,41	0,39
18	Hungarian	0,23	0,23
19	Turkish	0,10	**0,17**
20	Malay	0,088	0,14
21	Catalan	0,16	0,12
22	Slovak	0,11	0,12
23	Thai	0,11	0,12
24	Greek	0,082	0,10
25	Estonian	0,044	0,093
26	Arabic	0,0117	**0,089**
27	Croatian	0,085	0,071
28	Slovenian	0,057	0,062
29	Ukrainian	0,052	0,048
30	Hebrew	0,052	0,046
31	Romanian	0,042	0,046
32	Icelandic	0,050	0,038
33	Vietnamese	0,02	0,028
34	Lithuanian	0,019	0,028
35	Bulgarian	0,022	0,024
36	Latvian	0,0178	0,0158
37	Afrikaans	0,0073	0,0108
38	Basque	0,0141	0,0075
39	Galician	0,0070	0,0073
40	Welsh	0,0048	0,0065
41	Latin	0,0075	0,0053
42	Byeloruss.	0,0018	0,0037
43	Faeroese	0,0015	0,0024
44	W-Frisian	0,0004	0,0006
	unclass.	5,9273	6,09

4.2 Table 2: web pages/speaker[1]

(p/S: pages per speaker; assumed index size: 1999: 150 Mio, 2001: 560 Mio)

	Language	p/S 1999.	p/S 2001
1	Icelandic	0,3372	0,9391
2	English	0,2478	0,8475
3	Swedish	0,1759	0,5111
4	Danish	0,1037	0,4906
5	Norwegian	0,1284	0,44
6	Estonian	0,057	0,433
7	Finnish	0,1391	0,3833
8	German	0,0772	0,3469
9	Dutch	0,072	0,29
10	Faeroese	0,0489	0,2894
11	Italian	0,0691	0,240
12	French	0,0658	0,236
13	Czech	0,075	0,2333
14	Japanese	0,073	0,232
15	Catalan	0,0627	0,175
16	Slovenian	0,0442	0,175
17	Slovak	0,033	0,134
18	Korean	0,0199	0,108
19	Hungarian	0,025	0,0897
20	Croatian	0,0264	0,08
21	Basque	0,0362	0,07
22	Polish	0,0144	0,0682
23	Latvian	0,0196	0,0629
24	Russian	0,0143	0,0624
25	Welsh	0,0123	0,06
26	Greek	0,0126	0,057
27	Hebrew	0,0161	0,052
28	Lithuanian	0,0097	0,052
29	Portuguese	0,0101	0,0459
30	Spanish	0,0104	0,0452
31	Malay	0,0068	0,0399
32	Thai	0,0065	0,0288
33	Chinese	0,0036	0,0223
34	Bulgarian	0,0043	0,0173
35	Turkish	0,0027	0,0164
36	Galician	0,0034	0,0128
37	Romanian	0,0026	0,0104
38	Afrikaans	0,0018	0,0098
39	W-Frisian	0,0016	0,009
40	Ukrainian	0,0025	0,0084
41	Byeloruss.	0,0003	0,0026
42	Vietnamese	0,0005	0,0023
43	Arabic	0,0001	**0,0023**

[1] Speaker populations are based on Grimes (1997); estimates only. Latin is missing here for obvious reasons.

5 Internationalisation of search services

5.1 Requirements

Search engines are based on words and multiword lexemes: Documents on the WWW are fetched, parsed, and words in the document are put into an index. This index is queried by the users of the search engines.
From the tables presented in the last section it can be clearly seen that other languages than English become more and more important in quantitative measures.
Whereas the task of parsing and indexation is - from a linguistic point of view - relatively simple for English and some other languages (e.g. Danish, Norwegian or Swedish) with very reduced inflectional systems, the handling of documents in other languages definitely requires specific solutions. But even the simple tasks of language specific tokenisation (separation into words) and lemmatisation (reduction to canonical base forms) are currently not carried out most of the large international search engines.

5.2 Some possible enhancements

Tokenization: Some Asian languages, such as Japanese, Chinese, Korean and Thai do not overtly mark word boundaries by introducing blanks. In order to index words in these languages, word boundaries have to be detected by language specific software. This requires large dictionaries and morphological tools. Although software for this purpose does exist on the market, it is not widely used in for Internet search engines.
Lemmatisation - base form reduction: Morphological base form reduction, not very important for English with a very reduced morphology, is indespensible for many languages with rich morphology, especially if nouns are affected, as nouns are most often used as search words. Languages with a very rich noun morphology are e.g. Russian, Finnish - for these languages meaningful search cannot be offered without lemmatisation. In spite of the obvious benefits of lemmatisation, none of large international search engines apply lemmatisation systematically for morphological rich languages, such as the ones mentioned above.
Decompounding: For languages like German and the Scandinavian language, with a very rich compounding system that leads to new words that do not contain intermediate blanks, a decompounding algorithms can increase recall dramatically.
High-level NLP: While tokenization and lemmatization are pre-requisites to carry out a meaningful search at all for many languages, other applications of NLP can be used to improve and refine search results.
Spell checking and correction is only partially dependent on processing for specific languages - many algorithms can be used language independently.
High level NLP: The big international search engines currently have indexes that give access to nearly more than half a Billion documents (according to Notess (2001)).

Processing these documents for indexing needs considerable resources. Simple tasks, such as language identification can be carried out, but additional linguistic analysis of the whole index is not feasible for the whole collection of documents, at least not, if the processing is computionally intensive. Complex algorithms can thus be applied only for smaller subsets of the WWW. Possible developments include: Text classification and text filtering and automated abstracting.

6 Conclusions

Most international search engines provide a means for restricting the language(s) of searched documents. In principle language identification tools produce very reliable results. However, there seems to be a lower threshold for the size of documents that can be classified correctly. When inspecting such documents, it is clear that many of them could be classified manually. It seems feasible to develop methods to achieve a better performance on short documents. Another challenge is the development of tools that can detect and classify multilingual documents. This is easy as long as whole paragraphs can be classified, and gets more difficult for documents where the text chunks for different languages get very short.

Another interesting task is to follow the development of the different languages on the web. From a linguistic point of view, it would be desirable to include some more of the smaller languages, including minority languages, in order to have a full picture of the Internet usage of language communities world wide.

Currently, the language specific processing in most search engines is very reduced. In the last part of this paper I have shown that there is a high potential for the improvement of search engines through language specific processing.

References

Cavnar, W. B., Trenkle, J. M. (1994): N-Gram-Based Text Categorization. In *Symposion On Document Analysis and Information Retrieval* pp. 161-176. Las Vegas: University of Nevada.

Cowie, J., Ludovig E., Zacharski, Ron. (1998): An Autonomous, Web-based, Multilingual Corpus Collection Tool. *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*, pp 142-148.

Grefenstette, G. (1995): Comparing two language identification schemes, in *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT?95)*.

Grefenstette, G., Nioche, J. (2000): Estimation of English and non-English Language Use on the WWW. *Proceedings of RIAO'2000, "Content-Based Multimedia Information Access", Paris, April 12-14,2000*, pp. 237-246.

Grimes, B. F. (1997): Ethnologue. Languages of the World. Dallas, Texas : Summer Inst. of Linguistics. Web: http://www.sil.org/ethnologue/

Lavoie, B. F., O'Neill, E. T. (1999): How "World Wide" is the Web? Trends in the Internationalization of Web Sites. Dublin/Ohio: Online Computer Library Center. Web: http://oclc.com/oclc/research/publications/review99/

Notess, G. R. (2001): Search Engine Showdown: The Users' Guide to Web Searching. http://searchengineshowdown.com/

O'Neill, E. T., McClain, P. D., Lavoie, B. F. (1997): A Methodology for Sampling the World Wide Web. Dublin/Ohio: Online Computer Library Center. Web: http://oclc.com/oclc/research/publications/review97/