A Comparative Study on Language Identification Methods

Lena Grothe, Ernesto William De Luca and Andreas Nürnberger

University of Magdeburg, Faculty of Computer Science
39106, Magdeburg, Germany
{lena.grothe@st., ernesto.deluca@, andreas.nuernberger@}ovgu.de

Abstract

In this paper we present two experiments conducted for comparison of different language identification algorithms. Short words-, frequent words- and n-gram-based approaches are considered and combined with the Ad-Hoc Ranking classification method. The language identification process can be subdivided into two main steps: First a document model is generated for the document and a language model for the language; second the language of the document is determined on the basis of the language model and is added to the document as additional information. In this work we present our evaluation results and discuss the importance of a dynamic value for the *out-of-place measure*.

1. Introduction

Language identification can be used as a filtering technique to support users of information retrieval systems interested only in documents written in a certain language. Furthermore, language identification is an essential pre-processing step for other language processing techniques like stemming or machine translation that can only be applied to a document when the document's language is known. The goal of this paper is to analyze and compare different language identification methods with respect to their applicability in information retrieval settings.

The language identification process can be subdivided into two main steps: First a document model is generated for the document and a language model for the language; second the language of the document is determined on the basis of the language model and added to the document.

1.1. Language Identification Models

Language identification models contain entities (words or n-grams) encoding language specific features. Depending on their particular frequency of occurrence, these features are listed as entities in a language or document model. In the following, we present three different approaches that are relevant for our study. Different parameter values like word length or frequency are considered.

Short Word-Based Approach The first type of word-based approaches is called *short word*-based approach. It uses only words up to a specific length to construct the language model, independent from the particular word frequency.

Grefenstette (1995) uses one million characters of text for each language, tokenizing them and extracting all words with a length of up to five characters that occurred at least three times. The idea behind this approach is the language specific significance of common words like conjunctions having mostly only marginal lengths. Depending on the language, Grefenstette's language models contain between 980 and 2750 words. In contrast to Grefenstette, Prager (1999) uses only words up to four letters. Prager (1999) notes that short words will perform as good as a set of function words.

Frequent Word-Based Approach The second type of word-based approaches generate the language model using a specific amount of the *most frequent words*. These words describe a set of words having the highest frequency of all words occurring in a text. Different work has been presented in (Martino and Paulsen, 2001; Souter et al., 1994) who use the most frequent one hundred words, while Cowie et al. (1999) uses the most frequent one thousand words to generated a language model.

Souter et al. (1994) takes into account one hundred high frequent words per language extracted from training data for nine languages and 91% of all documents were correctly identified. The utilized frequency - relative or absolute - is not stated in the work.

Martino and Paulsen (2001) use one hundred most frequent words in *word frequency tables* where every word gets a *normalized frequency value*. This value is calculated by dividing the relative frequency of every word of the table with the relative frequency of the word at the first rank of the table.

In contrast to these two works, Cowie et al. (1999) use the one thousand most frequent words and their absolute frequencies of every of the 34 used languages.

N-Gram-Based Approach The third type of language models is generated by the *n-gram*-based approach and uses n-grams of different (Cavnar and Trenkle, 1994) or fixed (Grefenstette, 1995; Prager, 1999) lengths from tokenized words. In contrast, Dunning (1994) generates n-grams of sequences of Bytes.

An n-gram is a sequence of n characters. Before creating n-grams of a word, its beginning and the end are marked, for instance with an underscore. Doing this, start and end n-grams can be discovered.

Cavnar and Trenkle (1994) evaluate their algorithm on a corpus of 3713 documents in 14 languages and they notice that more comprehensive language models of the training files have better results for the language identification. For language models of more than 300 n-grams very good results of 99,8% were achieved.

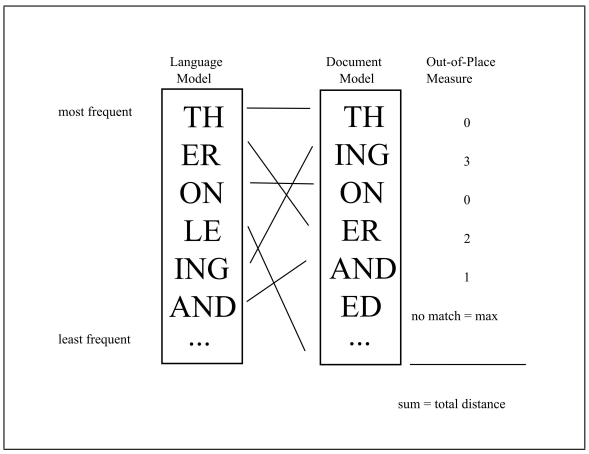

Figure 1: Out-of-Place-Measure Computation. (adapted from (Cavnar and Trenkle, 1994))

The approach of Dunning (1994) is quite similar to the one from Cavnar and Trenkle (1994). However, Dunning's approach does not use the tokenization of the text to build the n-grams, the language models are here generated assuming that the data are sequences of Bytes. For 50 KB on training data this approach identified 92% for test documents of length 20 Byte and 99,9% for longer texts with 500 Byte.

Grefenstette (1995) generates trigrams of whitespace tokenized text adding an underscore at the beginning and at the end of every extracted token. Every trigram has to occur at least 100 times. This results in minimum 2550 and maximum 2560 trigrams for each language.

Prager (1999) uses just as Grefenstette (1995), language models with n-grams of a fixed length and this system reached a performance of 99% of correct identified texts for documents of 200 Byte.

1.2. Language Classification Methods

Classification is the second step of the language identification process. The language of a document is identified using the generated document model as input for the classification method. Different classification approaches can be used for language identification like Vector Space Model (Prager, 1999), Monte Carlo sampling (Poutsma, 2001), Markov Chains in combination with Bayesian Decision Rules (Dunning, 1994), Relative Entropy (Sibun and Reynar, 1996), and Ad-Hoc Ranking (Cavnar and Trenkle, 1994).

Poutsma (2001) involved Monte Carlo sampling on language identification to compare it to the Ad-Hoc Ranking from Cavnar and Trenkle (1994) and the Mutual Information Statistics from Sibun and Reynar (1996). Instead of the necessary generation of language models in order to classify the documents language, the Monte Carlo technique used by Poutsma utilizes dynamic models. These are built by randomly selected features. The selection iteration is executed until the amount of features is adequat enough to determine the entire documents language. The necessary amount of features is investigated calculating the standard error σ of the feature samples.

Prager (1999) applied the Vector Space Model based on the similarity computation via the cosine distance between the training and test language model. The numerical values within one vector are defined by a token's occurrence in the training set times its inverse document frequency. Prager (1999) used three different values for the *idf-weight*: $1/n_i$, $1/\sqrt{n_i}$ and $1/\log(1+n_i)$. The best performing of these was the first, $1/n_i$.

The basic idea behind Dunning (1994) approach is the computation of a token's occurrence in every system supported language. The language of every language model, constructed using Markov Chains, is computed using the Bayesian Decision Rule. All supported languages are possible events given one considered test language model. The most likely language for a given language model is computed by the probability for one language l_i within the language pool L times the probability of the language model given language l_i computed with Markov Chains.

As stated above, Sibun and Reynar (1996) applied Relative Entropy also called Kullback-Leibler distance. The language models describe probability distributions and the Relative Entropy computes their similarity based on the amount necessary encoding information for a second language model given a first one. The respective language model probability distributions are computed by determining the amount of a group of token like trigrams, for instance. The language a language model might be written in most likely will minimize the Relative Entropy the language's training model.

The Ad-Hoc Ranking method of Cavnar and Trenkle (1994) relies on the comparison of two models which are ranked in descending frequency (see Figure 1). For every unclassified document, text features are extracted into a document model.In the same way, text features are extracted into a language model from training data. All features are sorted by their descending frequency (rank). The features contained in the document model are successively searched in the language models. First, the single *out-of-place measure* (Cavnar and Trenkle, 1994) is computed by comparing the entities of the two models and their rank. Then, a value is assigned. The resulting total distance of the out-of-place measures is used for assigning the language to the document. Otherwise, if an entity of the document model is not included in the language model, a maximum value is added to the total distance. This value used for distinguishing the no-matching language from the correct one. Finally, all total distances between the document model and the language models are computed. The language model with the smallest total distance is chosen as describing the language of the document. An example of such computation is given in figure 1, where the entities of the document and language model are compared. The first five entities of the document model are also contained in the

	language	size	source
cat	Catalan	10 million	WWW
dan	Danish	3 million	WWW
dut	Dutch	1 million	Newspaper
eng	English	10 million	Newspaper
est	Estonian	1 million	various
fin	Finnish	3 million	WWW
fre	French	3 million	Newspaper
ger	German	30 million	Newspaper
ice	Icelandic	1 million	Newspaper
ita	Italian	3 million	Newspaper
jap	Japanese	0.3 million	WWW
kor	Korean	1 million	Newspaper
nor	Norwegian	3 million	WWW
ser	Serbian	1 million	various
sor	Sorbian	0.3 million	various
spa	Spanish	1 million	Newspaper
swe	Swedish	3 million	WWW
tur	Turkish	1 million	WWW

Figure 2: Overview of Document Coverage in Leipzig Corpora Collection (Quasthoff et al., 2006)

language model. The out-of-place measures shown on the left side define the spread of their ranks in both models. In contrast, the entity *ED* is not included in the language model. Thus, the out-of-place measure gets the maximum no-matching value, increasing the difference of both models.

Since the Ad-Hoc Ranking from Cavnar and Trenkle (1994) is quite simple and experiments show very good results, we used this classification method as a basis to compare the three language model types - the short words-, frequent words- and n-gram-based approaches with each other and with varying model parameters.

2. Language Model Comparison

Language identification can be very helpful for information retrieval because users can be supported in filtering documents *language-oriented*. This means that every document is first classified into the belonging written language and resorted according to the language described above.

In this work, we study and compare the performance of existing language model approaches utilized as input for the classification method. Depending on the respective model type, every language model has been tested with different parameter values for word length, n-gram length or word frequency.

2.1. Data Collection

The data for our experiments are derived from the *Leipzig Corpora Collection (LCC)* (Quasthoff et al., 2006) and from randomly selected Wikipedia articles.

Leipzig Corpora Collection The Leipzig Corpora Collection contains corpora in different languages intended for natural language processing use. These corpora contain randomly selected sentences and are available in sizes of 100,000 sentences, 300,000 sentences, 1 million sentences (see Table 2). The sources are either newspaper articles or randomly collected web documents. The content is split into sentences. Preprocessing steps have been applied. For each word, the most significant words appearing as immediate left and right neighbors are included as well as co-occurrence information.

In our first experiment (see Table 4), we used the Leipzig Corpora Collection to create our training data for the following languages: Catalan, Danish, Dutch, English, French, German, Italian, Norwegian and Swedish.

Wikipedia Wikipedia[1] is a multilingual, Web-based, free content encyclopedia project. It is written collaboratively by volunteers and has grown rapidly into one of the largest reference Web sites. It contains about 9,000,000 articles in more than 250 languages[2]. The test data set, retrieved randomly from Wikipedia, is composed of 135 documents in the above mentioned respective languages.

In a second evaluation (see Table 5), we used the same training set explained above, extending it with Estonian, Finnish, Sorbian and Turkish. Furthermore, we decided to use another subset of the Leipzig Corpora Collection as test benchmark for verifying the results obtained in the previous experiments.

2.2. Language Models

We evaluated the different language models in order to find out the best performing one in conjunction with different model typical parameter values and in conjunction with experimentally optimized values for the out-of-place measure.

While in frequent words-based related works a fixed number of frequent words is used within a language model, in our work this number is determined dynamically. Thus, these parameter values $(10\%, 25\%, 50\%)$ indicate a percentage of most frequent words to be used. This percentage is computed depending on different lengths of documents (different numbers of words contained in the documents). An example of the 20 most frequent words is given in Table 1 for German, French and Swedish.

For the n-gram-based language model, we selected the parameter values $(3, 4, 5)$ that specify the maximum length of n-grams. As a pre-processing step of the n-gram creation, the beginning and end of a word are marked $n-1$ times on each end with a selected sign like the underscore. This indicates the start and the end of n-grams. A word of length k has then length of $k+(2n-1)$ characters. Afterwards, $k+(n-1)$ n-grams can be composed of a word of length k. In Table 2, we present an extract of the generated n-grams for German, French and Swedish with length values $(3, 4, 5)$.

[1] www.wikipedia.org

[2] http://stats.wikimedia.org/EN/Sitemap.htm

Ger.	freq.	Fre.	freq.	Swe.	freq.
die	2799	de	3690	och	2471
der	2656	la	2154	i	1972
und	1595	le	1779	att	1701
in	1381	l	1502	som	1272
den	938	les	1419	en	1201
das	834	à	1388	för	1197
von	711	et	1266	av	1151
zu	700	des	1123	på	1114
mit	690	d	1065	är	1096
im	647	en	981	med	967
auf	599	un	836	det	931
sich	570	a	824	til	l 871
ist	538	du	798	den	628
ein	535	une	706	har	623
nicht	530	pour	651	ett	600
des	517	dans	556	om	555
für	515	il	541	de	535
dem	513	au	529	kan	441
eine	457	est	502	vi	404
auch	453	qui	501	inte	371

Table 1: 20 Most Frequent Words

Ger.	3-gr.	Fre.	4-gr.	Swe.	5-gr.
und	2164	_dé	812	_på_	1114
_au	2090	ans_	811	__är	1114
_ge	1973	té_	804	_det	1107
gen	1939	_du_	798	_är_	1096
_be	1935	our_	777	_är_	1096
ine	1927	une_	775	ing_	1092
te_	1901	pour	763	__sk	1064
_da	1861	_con	754	__ko	1058
ung	1838	il_	750	den_	1044
_in	1830	ux_	737	ed__	1043
cht	1757	_su	724	__va	1001
te r	1690	c__	724	ta__	997
es_	1680	ts_	722	ill_	992
nde	1612	emen	719	med_	987
_p	1571	_une	710	_med_	967
ste	1561	us_	710	k___	942
hen	1519	atio	706	_det_	931
_zu	1505	ont_	704	__an	915
ver	1456	_av	686	__be	905
_si	1428	me_	684	till_	884

Table 2: Extract of n-grams contained in the language models

Table 3 shows an extract of the short words-based language model for the same languages covered by the n-gram-based one and with the maximum length $(3, 4, 5)$ of words.

Ger.	sw:3	Fre.	sw:4	Swe.	sw:5
sie	377	plus	291	finns	231
am	334	avec	280	jag	224
hat	312	qu	262	vid	222
aus	312	son	253	sig	218
bei	280	ont	250	under	201
wie	253	pas	250	också	178
um	248	ne	249	även	164
vor	228	sont	236	år	161
nur	203	mais	232	han	161
zum	199	été	212	skal	159
so	196	aux	197	vara	156
war	188	c	181	detta	155
bis	168	deux,	175	ska	154
zur	142	on	165	alla	154
sei	132	sa	146	olika	148
man	126	ses	141	när	143
daß	115	ans	139	här	142
wir	104	ces	138	får	139
ich	104	leur	131	där	139
vom	103	elle	120	efter	128

Table 3: Extract of short words contained in the language models

2.3. The Influence of the Out-of-Place Measure

For comparing the language models, we used the Ad-Hoc Ranking classification method (Cavnar and Trenkle, 1994). But analyzing the value of the out-of-place measure used in Cavnar and Trenkle (1994), we can notice that the authors do not specify the maximal value of the out-of-place measure, while Cowie et al. (1999) and Artemenko et al. (2006) give it a fixed value of 10.000.

In our work, we set different experimentally optimized values for the out-of-place measure. These values are dynamically chosen, instead of assigning them a fixed value for all document models, because of the coverage variance of words in the documents. As we already described in Section 1.2., the out-of-place measure is computed considering the single values of the same ranked entity (contained in the document and in the language model). The sum of these measures determines the total distance between a document and a language model. The language model with the smallest distance identifies the language of the document. The maximum value of this measure for a document represents the no-matching languages. The maximum value for a single document entity plays an important role in excluding the no-matching language models. Thus, this possible maximum single distance is used as a threshold for the out-of-place measure and can change the resulting total distance.

Summarizing, we can say that a dynamic value of the out-of-place measure is essential for achieving the best classification performance, as shown in the following section.

3. Evaluation

The language models (LM) discussed above are compared with another using the Ad-Hoc Ranking classification method. The evaluation results are shown in Table 4 and 5, where different parameters are set.

In Table 4 the parameters for the frequent word model (FW) are set to 10%, 25% and 50%, while the length of word and n-grams are respectively set to 3, 4 and 5 for the short word (SW) and the n-gram (NG) based model. The best parameters of each language model are then combined in order to evaluate if the performance of these methods can be increased. Thus, the frequent word model (FW, 25%) is combined with the short word-based (SW, 4) model and then with the n-gram-based (NG, 3) one.

The results are presented in three different columns; the amount of misclassified documents is given in the first column (incor.), while the second one (cor.) contains the number of correct identified documents. In the third column (unkn.) unclassified documents are shown. The fourth column presents the percentage of documents, where the language was correctly identified.

Due to the results of this first evaluation, we chose the best performing parameter settings for every language model: SW(4), FW(25%) and NG(3). The results of this second evaluation are presented in Table 5.

In order to train the classifier, we decided to use for both experiments only one document per language (selected randomly from the Leipzig Corpora Collection). While in the first experiment 15 documents per language have been randomly retrieved from Wikipedia, we chose for the second evaluation a subset of the Leipzig Corpora Collection containing 250 documents per language.

In all experiments, we could notice that the out-of-place measure influences the classification performance of the language models. The value of this measure has been fixed in the first experiments, while it is dynamically chosen for every language model in the last ones.

LM	incor.	cor.	unkn.	cor. %
FW(10%)	2	133	0	98,5
FW(25%)	1	134	0	99,2
FW(50%)	2	133	0	98,5
SW(3)	9	126	0	93,3
SW(4)	8	127	0	94,1
SW(5)	11	124	0	91,8
NG(3)	3	107	25	79,2
NG(4)	7	41	87	30,3
NG(5)	36	2	97	1,5
FW(25%), SW(4)	8	127	0	94,1
FW(25%), NG(3)	18	116	1	85,9

Table 4: Evaluation of Wikipedia Articles (15 documents per language)

First Evaluation with Wikipedia In a first evaluation (see Table 4) best results are achieved by the frequent word approach (FW, 25%) which identified 99,2% respective 134 of 135 documents correctly. Only one document has not been identified with its language. However, the results of the other two parameter values identified only yet another document incorrectly.

Good results are obtained by the short word approach for all three parameter values. However, the range of the results is slightly broader. The short words model with the parameter value 4 achieved the best performance of these three with 94,1% correct identified documents and only 8 incorrect identified ones while the parameter value 5 performed worst identifying only 91,8% documents correctly.

The performance of the n-gram-based language model was quite surprising and not satisfactory at all. Trigrams identified at least 79,2% of all documents with their language. Yet the other two n-gram-based runs performed significantly worse identifying the most documents with no language.

The combination of the frequent word approach FW(25%) and the short word approach SW(4) has indeed no alteration while the combination of FW(25%) and NG(3) shows a small improvement.

LM	incor.	cor.	unkn.	cor. %
SW(4)	0	3250	0	100
FW(25%)	0	3250	0	100
NG(3)	0	3250	0	100

Table 5: Evaluation of Leipzig Corpora Collection Documents (250 documents per language)

Second Evaluation with the Leipzig Corpora Collection In the second evaluation we changed the out-of-place measure for every run dynamically as described in Section 2.3.. The modification of this parameter leads to a very good performance enhancement for every approach. All approaches reach 100% correct classification. This confirms that the out-of-place measure strongly influences the classification performance and plays a crucial role, more than the choice of the language model.

4. Conclusion

In this paper we presented two evaluations for comparing three language identification algorithms. In the first one the frequent word approach (FW, 25%) achieved the best results with 99%. After choosing the best performing parameter values for every implemented method, we could observe that the results of the second evaluation show the importance of the out-of-place measure (Cavnar and Trenkle, 1994), where the choice of one specific approach is no more important. In future work we want to extend the amount of languages and language families (e.g. asian and arabic). In addition, it may be interesting to discover how

other language families behave applying the algorithms discussed in this work.

5. References

O. Artemenko, T. Mandl, M. Shramko, and C. Womser-Hacker. 2006. Evaluation of a language identification system for mono- and multilingual text documents. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, pages 859–860, Dijon, France.

W.B. Cavnar and J.M. Trenkle. 1994. N-gram based text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 161–169.

J. Cowie, Y. Ludovic, and R. Zacharski. 1999. Language recognition for mono- and multilingual documents. In *Proceedings of the Vextal Conference*, Venice.

T. Dunning. 1994. Statistical identification of language. Technical report mccs 94-273, Computing Research Laboratory, New Mexico State University.

G. Grefenstette. 1995. Comparing two language identification schemes. In *3rd International conference On Statistical Analysis of Textual Data*.

M.J. Martino and R.C. Paulsen. 2001. Natural language determination using partial words. U.S. Patent No. 6216102 B1.

A. Poutsma. 2001. Applying monte carlo techniques to language identification. In *Proceedings of Computational Linguistics in the Netherlands*.

J.M. Prager. 1999. Linguini: Language identification for multilingual documents. In *Proceedings of the 32nd Hawaii International Conference on System Sciences*.

U. Quasthoff, C. Biemann, and M. Richter. 2006. Corpus portal for search in monolingual corpora. *Computational Linguistics*, 19(1):61–74.

P. Sibun and J.C. Reynar. 1996. Language identification: Examining the issues. In *5th Symposium on Document Analysis and Information Retrieval*, pages 125–135, Las Vegas, Nevada, U.S.A.

C. Souter, G. Churcher, J. Hayes, J. Hughes, and S. Johnson. 1994. Natural language identification using corpus-based models. *Hermes Journal of Linguistics*, 13:183–203.