Stemming Strategies for European Languages

Jacques Savoy

Computer Science Dept.
University of Neuchatel
Rue Emile Argand 11
CH - 2009 Neuchatel (Switzerland)
Jacques.Savoy@unine.ch

Abstract: In this paper, we describe and evaluate different general stemming approaches for the French, Portuguese (Brazilian), German and Hungarian languages. Based on the CLEF test-collections, we demonstrate that light stemming approaches are quite effective for the French, Portuguese and Hungarian languages, and perform reasonably well for the German language. Variations in mean average precision among the different stemmers are also evaluated and are sometimes found to be statistically significant.

1 Introduction

In order to perform various text tasks such as text mining, information retrieval, entity extraction, or Web-based lexical statistics, we usually need to transform the words as they appear into their corresponding root or stem forms. Such a procedure is called stemming. In information retrieval (IR), when indexing documents or requests it is assumed that the application of a stemmer is a good practice. For example, when a query contains the word "plane," it seems reasonable to also retrieve documents containing the related word "planes."

As a first approach to designing a stemmer, we begin by removing only inflectional suffixes so that singular and plural word forms (e.g., "cars" and "car") or feminine and masculine variants (e.g., "actress" and "actor") will conflate to the same root. Stemming schemes that remove only morphological inflections are termed as "light" suffix-stripping algorithms, while more sophisticated approaches have also been proposed to remove derivational suffixes (e.g., '-ition', '-able' in "recognition," "recognizable," and "recognize"). Those suggested by Lovins [Lov68] or by Porter [Por80] are both typical examples for the English language

Stemming schemes are usually designed to work with general text. Certain stemming procedures may also be especially designed for a specific domain (e.g., in medicine) or a given document collection, such as that developed by Xu & Croft [XC98], which uses a corpus-based approach. This more closely reflects the language usage (including word frequencies and other co-occurrence statistics), instead of a set of morphological rules in

which the frequency of each rule (and therefore its underlying importance) is not precisely known.

While the English language has already been the object of various stemmer studies, this is not true of other European languages, for which stemmers and appropriate evaluation studies are not readily available. This paper is divided as follows: Section 2 presents some related works, while Section 3 depicts the main characteristics of our test-collections. Section 4 briefly describes the IR models used during our experiments. Section 5 discloses the various stemming approaches suggested, and in Section 6 they are evaluated from various perspectives. The main findings of this paper are presented in Section 7.

2 Related Work

In the IR domain we usually assume that stemming is an effective method of enhancing retrieval effectiveness through conflating several different word variants into a common form (the n-gram indexing strategy [MM04] is a typical exception to this rule). Most stemming approaches are based on morphological rules for the language involved (see [Lov68] or [Por80] for the English language). In such cases, suffix removal is also controlled through the adjunct of quantitative restrictions (e.g., '-ing' would be removed if the resulting stem had more than 3 letters as in "running," but not in "king") or qualitative restrictions (e.g., '-ize' would be removed if the resulting stem did not end with 'e' as in "seize"). Moreover, certain ad hoc spelling correction rules are also applied to improve conflation accuracy (e.g., "running" - 'ing' gives "run" and not "runn").

Simple stemming procedures can process text quickly but by ignoring word meanings they tend to make errors, usually due to over-stemming (e.g., "general" becomes "gener", and "organization" is reduced to "organ") or to under-stemming (e.g., the words "create" and "creation" do not always conflate to the same root). Thus the use of an on-line dictionary has been suggested in order to produce better conflations [KJ04].

Compared to other languages with more complex morphologies [Spr92], English stemmers are quite simple and to reduce their error rate, we may be consider using a dictionary [Sav93]. For those languages having more complex morphological structures, a deeper analysis may be required (e.g., for the Finnish language [KJ04]). Moreover, for other European languages only a few stemming procedures have been suggested, and those schemes available usually apply only to languages that are most spoken. For the African languages (except for Arabic) no stemming procedures are readily available while for the Asian languages, stemming is not always useful. In Chinese for example, morphological variations are usually not indicated (e.g., by a suffix as in Indo-European languages). In Japanese Hiragana characters are mainly used to write grammatical words (e.g., "do", "and", "of"), and inflectional endings (e.g., possessive, subject or object markers) for verbs, adjectives and nouns. Thus the removal of Hiragana characters is a simple process that may be viewed as a stemming procedure.

When analyzing IR stemming performance, Harman [Har91] demonstrated that no statistically significant improvements could be achieved through applying three different stem-

ming strategies, namely that of Lovins [Lov68], Porter [Por80] and another basic stemmer that conflates English singular and plural word forms (based on three rules). A query-by-query analysis revealed that stemming did affect the performance yet the number of queries depicting improvements was nearly equal to the number of queries showing degradation in performance. Other studies [Hul96] limited to only one language (usually English) showed modest improvement when using a stemmer and came to similar conclusions when using one search strategy: The use of a stemmer resulted in only modest improvement, and when compared to an approach ignoring stemming, the difference was not always statistically significant.

It was also surprising to see that during the last CLEF evaluation campaigns [PM05] (see Web site at `www.clef-campaign.org`), only a few stemmers were suggested by other participants and little effort was made to compare stemmers. For example, when evaluating the two statistical stemmers used for five languages, Di Nunzio *et al.* [DNO04] showed for each of these languages there were variations in relative retrieval performance. This means that any given stemming approach may work well for one language, yet poorly for another. When compared to statistical stemmers, Porter's stemmers seemed to work slightly better. For German, Braschler & Ripplinger [BR04] showed that for short queries stemming may enhance mean average precision by 23%, compared to 11% for longer queries. Finally, Tomlinson [Tom04] evaluated the differences between Porter's stemmer and the lexical stemmer (based on a dictionary of the corresponding language). He found that for the Finnish and German languages, the lexical stemmer tended to produce statistically better results, while for seven other languages the performance differences were small and insignificant.

From these facts, the following questions thus arise: 1) Does stemming affect IR performance for European languages other than English, or is the impact of stemming negligible due to their more complex morphology? 2) For these languages, are light stemming approaches less effective than more complex suffix-stripping algorithms? The rest of this paper provides answers to these questions.

3 Test Collections

In our experiments we used the CLEF 2005 corpora made up of newspaper and news agency articles, namely *Le Monde* (French), *SDA* (French), *Público* (Portuguese), *Folha* (Brazilian), *Magyar Hirlap* (Hungarian). The German collection is part of the GIRT corpora and composed of bibliographic records extracted from various sources in the social sciences. A typical record in this German corpus consists of a title, an abstract and a set of manually assigned descriptors. See Kluck [Klu04] for a more complete description of this corpus.

As shown in Table 1, both the French and Portuguese corpora have roughly the same size (487 MB vs. 564 MB), with the German ranking third and the Hungarian fourth, both in size (105 MB) and in number of documents (49,530). The Portuguese corpus has also a larger mean size (212.9 indexing terms/document) or median number of terms per doc-

	French	Portuguese	Hungarian	German
Size	487 MB	564 MB	105 MB	326 MB
# documents	177,452	210,734	49,530	151,319
mean number of terms	178	212.9	142.1	89.6
median size (# terms)	126	171	88	95
# queries	50	50	50	50
# rel. doc./query	50.7	58.1	18.8	86.9

Table 1: Some statistics from our test collections (CLEF)

ument (171) than does the French collection (mean = 178, median = 126). This mean value is slightly smaller for the Hungarian corpus (mean = 142.1, median = 88) and smallest for the German collection (mean = 89.6, median = 95). During the indexing process, we retained the logical sections allowed by CLEF evaluation campaigns. For the German collection, we applied a decompounding procedure [Sav04], and retained compound words and their components in document or topic representations. Compound words (e.g., handgun, worldwide) are widely used in German and can lead to more difficulties they do for the English language. "Computersicherheit" for example is composed of "Computer" + "Sicherheit" (security) and could also appear separately (e.g., "die Sicherheit mit Computern"). Finally, although accents were removed this process may have accidentally conflated words with different meanings into the same form (e.g., in French the word "tâche" (task) and "tache" (mark, spot)).

Each topic was structured into three logical sections comprising a brief title, a one-sentence description, and a narrative part specifying the relevance assessment criteria. In this study, we used the shortest query formulation in order to reflect a more realistic search context. Based on the topic title only, the query had a mean size of 2.8 search terms for the French collection, 2.6 for the Portuguese, 2.2 for the Hungarian and 1.7 for the German.

The available topics covered various subjects (e.g., "Brain-Drain Impact," "Internet Junkies," or "Creutzfeldt-Jakob Disease") and included both regional ("Deutsche Bank Takeovers") and international coverage ("Microsoft Competitors").

As shown in Table 1, the mean number of relevant items per query for the French and Portuguese collection has a relatively similar value (50.7 and 58.1 respectively), but this value is lower for the Hungarian corpus (18.8), a collection whose size is only one quarter that of the French corpus. The mean number of relevant articles per request for the German test-collection was clearly higher, at 86.9

4 IR Models

In order to obtain a broader view of the relative merit of the various retrieval models and stemming approaches, we used seven vector-space schemes and two probabilistic models. First we adopted the classical $tf\ idf$ model, in which the weight attached to each indexing

ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t}(tf_{ik} \cdot idf_k)^2}}$
atn	$w_{ij} = idf_j \cdot \left[\frac{0.5 + 0.5 \cdot tf_{ij}}{max\ tf_{i.}}\right]$
ltn	$w_{ij} = [ln(tf_{ij}) + 1] \cdot idf_j$
dtn	$w_{ij} = [ln(ln(tf_{ij}) + 1) + 1] \cdot idf_j$
ltc	$w_{ij} = \frac{[ln(tf_{ij})+1] \cdot idf_j}{\sqrt{\sum_{k=1}^{t}([ln(tf_{ik})+1] \cdot idf_k)^2}}$
dtu	$w_{ij} = \frac{[ln(ln(tf_{ij})+1)+1] \cdot idf_j}{(1-slope) \cdot pivot + (slope \cdot nt_i)}$
Lnu	$w_{ij} = \frac{\frac{ln(tf_{ij})+1}{ln\left(\frac{l_i}{nt_i}\right)+1}}{(1-slope) \cdot pivot + (slope \cdot nt_i)}$
Okapi	$w_{ij} = \frac{(k_1+1) \cdot tf_{ij}}{K + tf_{ij}}$ with $K = k_1 \cdot \left[(1-b) + b \cdot \frac{l_i}{avdl}\right]$

Table 2: Various Weighting Schemes

term was the product of its term occurrence frequency (or tf_{ij} for indexing term t_j in document D_i) and its inverse document frequency (or $idf_j = ln(n/df_j)$, where n indicates the number of documents in the corpus, and df_j the number of documents in which the term t_j appears). To measure similarities between documents and requests, we computed the inner product after normalizing indexing weights (model denoted "doc=ntc, query=ntc" or "ntc-ntc").

Other variants might also be created, especially in cases when the occurrence of a particular term in a document was deemed a rare event. Thus, it might be good practice to assign more importance to the first occurrence of this word, as compared to any successive, repeating occurrences. Therefore, the tf component might be computed as the $ln(tf) + 1$ (model "doc=ltc, query=ltc") or as $0.5+0.5 \cdot [tf\ /\ max\ tf\ in\ D_i]$. Of course, other weighting formulae could also be used for documents and requests, leading to different weighting combinations (see Table 2 where the length of document D_i is denoted by nt_i, and $avdl$, b, k_1, $pivot$ and $slope$ are constants.). We might also consider that a term's presence in a shorter document would provide stronger evidence than in a longer document, leading to more complex IR models; for example the IR model denoted by "doc=Lnu" [BS96], "doc=dtu" [SP99].

In addition to these vector-space schemes, we also considered probabilistic models such as the Okapi model [RB00]. As shown in Table 2, this model includes some constants fixed as b=0.7, k_1=1.5 (French), b=0.7, k_1=1.5 (Portuguese), b=0.75, k_1=1.2 (Hungarian), and b=0.5, k_1=1.2 (German), while $avdl$ indicates the mean document length (values are given in Table 1). As a second probabilistic approach, we implemented the GL2 approach taken from the *Divergence from Randomness* (DFR) framework [AvR02], based on combining the two information measures formulated below:

$$w_{ij} = Inf_{ij}^1(tf) \cdot Inf_{ij}^2(tf) = -log_2\left[Prob_{ij}^1(tf)\right] \cdot (1 - Prob_{ij}^2(tf))$$

where w_{ij} indicates the indexing weight attached to term t_j in document D_i, $Prob^1_{ij}(tf)$ is the probability of finding tf occurrences of the indexing unit t_j in the document D_i. On the other hand, $Prob^2_{ij}(tf)$ is the probability of encountering a new occurrence of t_j in the document given that we have already found tf occurrences of this indexing unit. Within this framework, the GL2 model is based on the following formulae:

$$Prob^1_{ij}(tf) = [1/(1+\lambda_j)] \cdot [\lambda_j/(1+\lambda_j)]^{tfn_{ij}} \text{ with } \lambda_j = tc_j/n \quad (1)$$

$$Prob^2_{ij}(tf) = tfn_{ij} \,/\, (tfn_{ij}+1) \text{ with} \quad (2)$$

$$tfn_{ij} = tf_{ij} \cdot \log_2 [1 + ((C \cdot mean\ dl)/l_i)] \quad (3)$$

where l_i the number of indexing terms included in the representation of D_i, tc_j represents the number of occurrences of term t_j in the collection, C is a constant fixed at 1.25, and $mean\ dl$ (mean document length) depends on the corpus (values given in Table 1).

5 Stemming Strategies

In our point of view it is important to develop a simple approach, one that does not require a dictionary or any other sophisticated data structures or processing. We also believe that a good IR system stemming procedure should focus mainly on nouns and adjectives, thus ignoring various verb forms (although past participles could be an exception to this rule). Given this assumption, our stemming approach tried to remove morphological variations associated with number (singular vs. plural), gender (masculine or feminine), and various grammatical cases (nominative, accusative, ablative, etc.). For verbal forms we ignored variations which are usually too numerous, while for adjectives we did not attempt to remove comparative and superlative suffixes (e.g., "larger," "largest"), forms that are less frequently used.

An analysis of the grammar of any given language however usually reveals numerous inflectional rules, some of which are used for only one or a few words (e.g., "child" and "children" or "foot" and "feet" in English). As for those languages having morphologies more complex than English, we could develop an even simpler stemmer, based only on a few but frequently used rules. For French, such a stemming approach (label "S-stemmer") is defined as follows.

```
For words of six or more letters
   if final letters are '-aux' then replace -aux by -al
   if final letter is '-x' then remove '-x',
   if final letter is '-s' then remove '-s',
   if final letter is '-r' then remove '-r',
   if final letter is '-e' then remove '-e',
   if final letter is '-é' then remove '-é',
   if final two letters are the same, remove final letter
```

For example, the word "chevaux" (horses) is reduced to "cheval" (horse) and the words

”baronnes” (baronesses) or ”barons” are all reduced to the same stem ”baron”. As a variant for French, we would suggest removing other inflections and also certain derivational suffixes. Labeled ”UniNE” in our experiments, this stemming method is composed of 27 rules (see `www.unine.ch/info/clef/`).

For Portuguese, our suggested stemmer tries to remove inflections attached to both nouns and adjectives, based on rules for the plural form (10 rules) and feminine form (13 rules). In Portuguese as in English the usual plural form is obtained by adding an ’-s’ (e.g., ”amigo” and ”amigos” (friend)). This suffix is also used for adjectives. There are of course various exceptions to the general rule (e.g., ”mar” and ”mares” (sea), ”fuzil” and ”fuzis” (gun), and for the adjective ”fácil” (easy), its plural form is ”fáceis”). The feminine form is usually obtained by replacing the final ’-o’ by an ’-a’ (e.g., ”americano” and ”americana”), but there are various exceptions to be taken into account (e.g., ”inglês” (British) becomes ”inglesa” in the feminine, ”leão” (lion) becomes ”leoa” and ”professor” gives ”professora”)

For German our suggested stemmer incorporates 11 rules to remove both plural forms and grammatical case endings (e.g., those usually used to indicate the genitive case by employing an ’-s’ or ’-es’ as in ”Staates” (of the state), ”Mannes” (of the man)). In German the plural form is denoted using a variety of endings such as ’-en’ (e.g., ”Motor”, ”Motoren” (engine)), ’-er’, ’-e’ (e.g., ”Jahr”, ”Jahre” (year)) or ’-n’ (e.g., ”Name”, ”Namen”). Plural forms also use diacritic characters (e.g., ”Apfel” (apple) becomes ”Äpfel”) or in conjunction with a suffix (e.g., ”Haus” and ”Häuser” (house)). Also frequently used are the suffixes ’-en’ or ’-n’ to indicate grammatical cases or for adjectives (e.g., ”ï einen guten Mann” (a good man) in the accusative singular form).

As with Finnish, Hungarian makes use of a greater number of grammatical cases (usually 18) than does German (four cases). Each case has its own unambiguous suffix however; e.g. the noun ”house” (”hàz” in nominative) may appear as ”hàzat” (accusative case, as in ”(I see) the house”), ”hàzakat” (accusative plural case, as in ”(I see) the houses”), ”hàzamat” (”ï my house”) or ”hàzamait” (”... my houses”). In this language the general construction used for nouns is as follows: ’stem’ ’possessive marker’ ’plural’ ’case’ as in ’hàz’ + ’ak’ + ’at’ (in which the letter ’a’ is introduced to facilitate better pronunciation because ”hàzkt” could be difficult to pronounce). Our suggested ”UniNE” stemming procedure for the plural in this language is based on two rules, plus there are 17 rules for removing various possessive suffixes and 21 rules for removing case markers. In a lighter stemming procedure, we would ignore the possessive marker (under the assumption that such suffixes are infrequently used and in an effort to reduce the number of conflation errors). Moreover, in order to automatically remove the most frequent cases we would apply only 13 rules.

Compared to the 260 rules used by Lovins or the 60 by Porter in their stemmers proposed for the English language, the stemmers we suggest could be viewed as light stemmers for languages having more complex morphologies than English. These stemmers are available at `www.unine.ch/info/clef/`. As an alternative to our light stemmers, we might also employ a more aggressive stemmer, taken from among those found within Porter’s family (available for the French, Portuguese and German languages at `snowball.tartarus.org/`). In the next section, we will evaluate these various stem-

ming approaches and their resultant retrieval effectiveness.

6 Evaluation

To measure retrieval performance, we adopted mean average precision (MAP) as computed by TREC_EVAL. To determine whether or not any given search strategy might be better than another, we applied a statistical test. More precisely, we stated the null hypothesis (denoted H_0) specifying that both retrieval schemes achieved similar performance levels (MAP), and this hypothesis would be rejected at a significance level fixed at α = 5% (two-tailed test). In this paper we have underlined any statistically significant differences that result from a two-sided non-parametric bootstrap test [Sav97].

6.1 IR Models Evaluation

Based on this evaluation methodology, Table 3 depicts the MAP for the French or Portuguese collections, using different stemming approaches. The same information is given in Table 4 for the Hungarian and German corpora. These experiments show that the Okapi probabilistic model usually produces the best retrieval performance (depicted in bold) across the different languages. The Hungarian corpus without stemming is an exception to this finding, for which the MAP difference between the "dtu-dtn" approach (0.1980) and the Okapi model (0.1957) is not however statistically significant (and thus we did not underline this value). Moreover, when considering the French, Portuguese and German corpora, the differences between the Okapi model and other IR models are statistically significant.

For the Hungarian corpus, the difference between the two probabilistic schemes (GL2 and Okapi) and the two best performing vector-processing models ("Lnu-ltc" and "dtu-dtn") is not statistically significant.

6.2 Nonstemming vs. Stemming

In this section we would like to apply a different point of view in order to verify whether or not a given stemming procedure might statistically improve mean average precision. To verify the effectiveness of this approach we adapted retrieval performance without stemming as the baseline (MAP depicted under the label "None" in Tables 3 and 4). For the French collection, all three stemming approaches performed better statistically than the baseline, for the nine IR models. After averaging the percentage of enhancement across these nine models, we found an average increase of 35% when using the UniNE stemmer, 30.5% with Porter's scheme, and 27.3% for the "S-stemmer".

With the Portuguese and German corpora, we found similar conclusions; with the two

	Mean average precision						
Model	French None	French UniNE	French S-stem.	French Porter	Portug. None	Portug. UniNE	Portug. Porter
Okapi	**0.2260**	**0.3045**	**0.2858**	**0.2978**	**0.2238**	**0.2873**	**0.2610**
GL2	0.2125	0.2918	0.2739	0.2878	0.2182	0.2755	0.2502
Lnu-ltc	0.2112	0.2933	0.2717	0.2808	0.1989	0.2611	0.2296
dtu-dtn	0.2062	0.2780	0.2611	0.2758	0.2096	0.2571	0.2189
atn-ntc	0.2088	0.2755	0.2603	0.2695	0.2049	0.2458	0.2128
ltn-ntc	0.1945	0.2466	0.2402	0.2371	0.1758	0.2149	0.1831
lnc-ltc	0.1545	0.2233	0.2080	0.2131	0.1519	0.1811	0.1607
ltc-ltc	0.1461	0.1975	0.1879	0.1922	0.1433	0.1625	0.1415
ntc-ntc	0.1462	0.1918	0.1807	0.1758	0.1344	0.1553	0.1422

Table 3: MAP of various IR models applying different stemming strategies (French & Portuguese corpus)

	Mean average precision					
Model	Hungarian None	Hungarian Light	Hungarian UniNE	German None	German UniNE	German Porter
Okapi	0.1957	**0.2988**	**0.3076**	**0.3552**	**0.3931**	**0.4058**
GL2	0.1883	0.2905	0.2964	0.3464	0.3805	0.3934
Lnu-ltc	0.1887	0.2913	0.2868	0.3357	0.3638	0.3793
dtu-dtn	**0.1980**	0.2857	0.2900	0.3357	0.3671	0.3826
atn-ntc	0.1794	0.2651	0.2755	0.3381	0.3653	0.3789
ltn-ntc	0.1919	0.2556	0.2567	0.3184	0.3421	0.3573
lnc-ltc	0.1616	0.2188	0.2153	0.2757	0.2983	0.3032
ltc-ltc	0.1675	0.2207	0.2183	0.2575	0.2773	0.2891
ntc-ntc	0.1713	0.2162	0.2079	0.2510	0.2649	0.2759

Table 4: MAP of various IR models applying different stemming strategies (Hungarian & German corpus)

stemming procedures always performing statistically better than the search done without stemming. When computing percentages of the MAP differences across the nine IR models, we found the UniNE stemmer would improve MAP by 22% on average for the Portuguese collection and by 8.4% for the German corpus. Using the same baseline, Porter's stemmer improved MAP by 7.7% on average for the Portuguese collection, and by 12.4% for the German corpus.

For the Hungarian corpus, both stemming approaches improved the MAP when compared to the nonstemming approach (on average by 42.8% for UniNE stemmer, and 42.2% for the light stemming scheme). Both stemmers did indeed improve MAP statistically compared to an indexing scheme that ignored stemming.

6.3 Comparing Different Stemmers

It is assumed that stemming usually improves retrieval performance (even though the difference might not always be statistically significant) on the one hand, and on the other, different stemmers tend to produce similar results. To investigate this issue we compared the retrieval effectiveness produced by the various stemmers.

Using the "S-stemmer" retrieval performance as a baseline, for the French collection Porter's stemmer improved by 2.5% on average (computed from the nine IR models). These differences are however not statistically significant. The UniNE stemmer showed an average enhancement of 6%, and this difference was statistically significant for the Okapi, GL2, and "dtu-dtn" IR schemes. While performance differences between Porter and UniNE always favored the second (+3.5% in average), these variations were not however statistically significant.

For Portuguese, the situation is relatively similar. Using the UniNE stemmer as a baseline, Porter's approach resulted in lower MAP (-11.8% in average across the nine IR models). Moreover, for the 5 IR models, the differences were also statistically significant. Thus for both French and Portuguese, different stemmers would result in IR performances with statistically significant differences. Moreover, for these languages at least a light stemming approach seemed to be more effective than a stemming approach that tried to remove more suffixes.

For German, Porter's stemmer provided better retrieval performance than did the UniNE scheme (average difference of 3.7% over all IR models). The difference between these two stemming schemes however was never statistically significant. Finally for Hungarian, the difference between the two suggested stemming methods is very small (0.3% on average), and not statistically significant.

When performing high precision searches, we assumed that the light stemming approach would produce better results. To verify this hypothesis, we computed the retrieval precision for five documents from the French corpus, and then compared the three stemming approaches (mean precision depicted in Table 5). This data did not show any enhancement over the light stemming approach (evaluation given under the label "S-stemmer") or a scheme ignoring stemming (under the label "None"). The other two stemming approaches

	Precision after 5 documents			
	None	S-stemmer	UniNE	Porter
Okapi	**0.5040**	0.5280	**0.5480**	**0.5400**
GL2	0.4840	0.5200	0.5280	0.5240
Lnu-ltc	0.4960	**0.5320**	0.5200	0.5160
dtu-dtn	0.4320	0.4720	0.4840	0.4720
atn-ntc	0.4800	0.5120	0.5040	0.5120
ltn-ntc	0.4560	0.4840	0.4600	0.4720
lnc-ltc	0.3960	0.4480	0.4280	0.4240
ltc-ltc	0.3240	0.3520	0.3480	0.3680
ntc-ntc	0.3360	0.3640	0.3600	0.3600

Table 5: Mean precision after 5 documents (French corpus)

did however seem to show better results. The differences in performance between the "S-stemmer" and the others were never statistically significant.

7 Conclusion

We have proposed and analyzed various stemming approaches based on four different languages, and our experiments have demonstrated that the Okapi probabilistic model produces the best retrieval performance. Moreover, the differences between the Okapi and other IR models are statistically significant for the French, Portuguese and German corpora.

A second set of experiments clearly shows that a stemming procedure improves retrieval effectiveness for those European languages belonging to either the Latin (French, Portuguese), Germanic (German) or Finno-Ugrian (Hungarian) families. For these same four European languages, differences in retrieval performance are significant from a statistical point of view and favor searches performed with a stemmer.

When comparing different stemming strategies, it seems that a light stemming approach (one that tries to automatically remove the most frequently used inflectional suffixes) produces better MAP than does a more aggressive stemmer. Moreover, for some IR models, the difference between these two stemming schemes could be statistically significant and in favor of a light stemming solution. For the German and the Hungarian languages, the performance difference between the stemmers is not statistically significant. Finally, based on our experiments we cannot confirm that a light stemmer would be more effective for high precision searches, at least for the French language.

Acknowledgments. This research was supported in part by the Swiss National Science Foundation under Grant #200020-103420.

References

[AvR02] G. Amati and C.J. van Rijsbergen. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM - Transactions on Information Systems*, 20:357–389, 2002.

[BR04] M. Braschler and B. Ripplinger. How Effective is Stemming and Decompounding for German Text Retrieval? *IR Journal*, 7:291–316, 2004.

[BS96] Singhal A. Mitra M. Buckley, C. and G. Salton. New Retrieval Approaches using SMART. In *Proceedings TREC-4*, pages 25–48, 1996.

[DNO04] Ferro N. Melucci M. Di Nunzio, G.M. and N. Orio. Experiments to Evaluate Probabilistic Models for Automatic Stemmer Generation and Query Word Translation. In Braschler M. Gonzalo J. Kluck M. Peters, C., editor, *Comparative Evaluation of Multilingual Information Access Systems*, Lecture Notes in Computer Science: Vol. 3237, pages 220–235, Heidelberg, 2004. Springer.

[Har91] D. Harman. How Effective is Suffixing? *Journal of the American Society for Information Science*, 42:7–15, 1991.

[Hul96] D. Hull. Stemming Algorithms: A Case Study for Detailed Evaluation. *Journal of the American Society for Information Science*, 47:70–84, 1996.

[KJ04] Laurikkala J. Jarvelin K. Korenius, T. and M. Juhola. Stemming and Lemmatization in the Clustering of Finnish Text Documents. In *Proceedings of the ACM-CIKM*, pages 625–633, 2004.

[Klu04] M. Kluck. The GIRT Data in the Evaluation of CLIR Systems - From 1997 Until 2003. In Braschler M. Gonzalo J. Peters, C. and M. Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems*, Lecture Notes in Computer Science: Vol. 3237, pages 376–390, Heidelberg, 2004. Springer.

[Lov68] J.B. Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[MM04] P. McNamee and J. Mayfield. Character N-gram Tokenization for European Language Text Retrieval. *IR Journal*, 7:73–97, 2004.

[PM05] Clough P.D. Gonzalo J. Jones G.J.F. Kluck M. Peters, C. and B. Magnini. *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*. Lecture Notes in Computer Science: Vol. 3491. Springer, Berlin, 2005.

[Por80] M.F. Porter. An Algorithm for Suffix Stripping. *Program*, 14:130–137, 1980.

[RB00] Walker S. Robertson, S. E. and M. Beaulieu. Experimentation as a Way of Life: Okapi at TREC. *Information Processing & Management*, 36:95–108, 2000.

[Sav93] J. Savoy. Stemming of French Words Based on Grammatical Category. *Journal of the American Society for Information Science*, 44:1–9, 1993.

[Sav97] J. Savoy. Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing & Management*, 33:495–512, 1997.

[Sav04] J. Savoy. Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic Models for Effective Monolingual Retrieval. In Braschler M. Gonzalo J. Peters, C. and M. Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems*, Lecture Notes in Computer Science: Vol. 3237, pages 322–336, Heidelberg, 2004. Springer.

[SP99] Choi J. Hindle D. Lewis D.D. Singhal, A. and F. Pereira. AT&T at TREC-7. In *Proceedings TREC-7*, pages 239–251, 1999.

[Spr92] R. Sproat. *Morphology and Computation*. The MIT Press, Cambridge, 1992.

[Tom04] S. Tomlinson. Lexical and Algorithmic Stemming Compared for 9 European Languages with Humminbird^{TM}SearchServer at CLEF 2003. In Braschler M. Gonzalo J. Kluck M. Peters, C., editor, *Comparative Evaluation of Multilingual Information Access Systems*, Lecture Notes in Computer Science: Vol. 3237, pages 286–300, Heidelberg, 2004. Springer.

[XC98] J. Xu and B. Croft. Corpus-Based Stemming Using Cooccurrence of Word Variants. *ACM-Transactions on Information Systems*, 16:61–81, 1998.