Retrieval Experiments at Morpho Challenge 2008

Paul McNamee
JHU Human Language Technology Center of Excellence
paul.mcnamee@jhuapl.edu

Abstract

Morpho Challenge 2008 hosted an extrinsic evaluation of morphological analysis that explored whether unsupervised morphology induction could benefit information retrieval. This paper presents results in alternative methods for word normalization using test sets from the Cross-Language Evaluation Forum (CLEF) ad-hoc collections. Preliminary results for the Morpho Challenge 2008 evaluation are consistent with these data. We found that: (1) rule-based stemming is effective in less morphologically complicated languages; (2) alternative methods for stemming such as unsupervised learning of morphemes and least common n-gram stemming are helpful; and, (3) full character n-gram indexing is the most effective form of tokenization in more morphologically complex languages.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Experimentation

Keywords

Multilingual text retrieval, Character n-grams, Morphological normalization

1 Introduction

Over the past year we conducted experiments using alternative methods of lexical normalization. Using test sets in 13 languages used in Cross-Language Evaluation Forum (CLEF) evaluations between 2002 and 2007, we compared the use of a rule-based stemmer (*Snowball*), an unsupervised morphological segmenter (*Morfessor*), a synthetic form of stemming based on selecting a single character n-gram from each word, and the use of all n-grams for a given word. Character n-grams of length $n = 5$ were the most effective technique, performing 18% better than unnormalized words, averaged across the set of languages.

N-grams possess many advantages. They work in every language, require no training, and are more effective than plain words. Their main shortcoming is increased disk space requirements and slower query response times. To address the performance penalty we introduced a method of n-gram stemming that selects for each word its least common n-gram. By selecting only one n-gram per word, the inverted file becomes no larger than when words or stemmed words are used. And query responses times also improve since the number of query terms becomes equal to the number of words, not a function of the number of letters in the query string.

While our n-gram stems perform better than unlemmatized words, it was clear from the Morpho Challenge 2007 evaluation [2] that other techniques were even more effective. This year we wanted to compare full (*i.e.,* traditional) n-gram indexing against the more linguistically sophisticated approaches of other participants. For the Information Retrieval subtask at Morpho Challenge 2008 we provided official submissions using three techniques in each of English, Finnish, and German. The methods were:

- Character n-grams with length $n = 5$
- Character n-grams with length $n = 4$
- The rarest 5-gram from each word (measured using document frequency statistics)

In Section 2 we describe our tokenization experiments on the CLEF data sets. In Section 3 we analyze our Morpho Challenge 2008 results.

2 Analysis on Past CLEF Collections

We compared plain words, stems, induced morphemes, n-gram stems, and character n-grams using test sets from the CLEF ad hoc tasks between 2002 and 2007. In each of the 13 languages we used two years worth of data except for Czech where only one year was available. The number of test queries per language varied from 50 (Czech) to 107 (Spanish). The document collections for each language were indexed using the various methods of tokenization. Common to each method was conversion to lower case letters, removal of punctuation, and truncation of long numbers to 6 digits. The HAIRCUT retrieval engine was used with a statistical language model retrieval metric. The similarity calculation combines document term frequencies and corpus frequencies (for smoothing) using linear interpolation with a smoothing constant of 0.5 [6]. The methods of tokenization examined were:

- **words**: space-delimited tokens.
- **snow**: output of the *Snowball* stemmer[1], if available in the given language.
- **morf**: the set of morphemes produced by *Morfessor*[2] for each input word. A model was trained using the document collection's lexicon with digit-containing tokens omitted. The default parameters for the Morfessor algorithm were used [1].
- **lcn4**: the least common word-internal character 4-gram for each input word [3].
- **lcn5**: the least common word-internal character 5-gram for each input word.
- **4-gram**: overlapping, word-spanning character 4-grams produced from the stream of words encountered in the document or query.
- **5-gram**: length $n = 5$ n-grams created in the same fashion as the character 4-grams.

In Table 1 results are presented using mean average precision to compare performance. The score for the highest performing technique in each language is emboldened.

2.1 Unnormalized words

Not attempting to control for morphological processes can have harmful effects. In Bulgarian, Czech, Finnish, and Hungarian, more than a 30% loss is observed compared to the use of 4-grams as indexing terms.

[1] Available from http://snowball.tartarus.org/

[2] Available from http://www.cis.hut.fi/projects/morpho/

Table 1: Comparison of 7 Tokenization Alternatives (Mean Average Precision)

Language	Data	Queries	Words	Snow	Morf	LCN4	LCN5	4-gram	5-gram
Bulgarian	06-07	100	0.2195		0.2786	0.2937	0.2547	**0.3163**	0.2916
Czech	07	50	0.2270		0.3215	0.2567	0.2477	**0.3294**	0.3223
Dutch	02-03	106	0.4162	0.4273	0.4274	0.4021	0.4073	0.4378	**0.4443**
English	02-03	96	0.4829	**0.5008**	0.4265	0.4759	0.4861	0.4411	0.4612
Finnish	02-03	75	0.3191	0.4173	0.3846	0.3970	0.3900	0.4827	**0.4960**
French	02-03	102	0.4267	**0.4558**	0.4231	0.4392	0.4355	0.4442	0.4399
German	02-03	106	0.3489	0.3842	0.4122	0.3613	0.3656	0.4281	**0.4321**
Hungarian	06-07	98	0.1979		0.2932	0.2784	0.2704	**0.3549**	0.3438
Italy	02-03	100	0.3950	**0.4350**	0.3770	0.4127	0.4054	0.3925	0.4220
Portuguese	05-06	100	0.3232		0.3403	0.3442	0.3381	0.3316	**0.3515**
Russian	03-04	62	0.2671		0.3307	0.2875	0.3053	**0.3406**	0.3330
Spanish	02-03	107	0.4265	**0.4671**	0.4230	0.4260	0.4323	0.4465	0.4376
Swedish	02-03	102	0.3387	0.3756	0.3738	0.3638	0.3467	0.4236	**0.4271**
Average			0.3375		0.3698	0.3645	0.3604	0.3955	**0.3979**
Average (8 Snowball langs)			0.3504	0.3848	0.3608	0.3642	0.3632	0.3885	**0.3956**

Table 2: Word Normalization Examples

Word	Snowball	Morfessor	5-grams
authored	author	author+ed	_auth, autho, uthor, thore, hored, ored_
authorized	author	author+ized	_auth, autho, uthor, thori, horiz, orize, rized, ized_
authorship	authorship	author+ship	_auth, autho, uthor, thors, horsh, orshi, rship, ship_
afoot	afoot	a+foot	_afoo, afoot, foot_
footballs	footbal	football+s	_foot, footb, ootba, otbal, tbaall, balls, alls_
footloose	footloos	foot+loose	_foot, footl, ootlo, otloo, tloos, loose, oose_
footprint	footprint	foot+print	_foot, footp, ootpr, otpri, tprin, print, rint_
feet	feet	feet	_feet, feet_
juggle	juggl	juggle	_jugg, juggl, uggle, ggle_
juggled	juggl	juggle+d	_jugg, juggl, uggle, ggled, gled_
jugglers	juggler	juggle+r+s	_jugg, juggl, uggle, ggler, glers, lers_

2.2 Snowball stemming

Snowball does not support Bulgarian, Czech, or Russian and due to character encoding issues with the software we were not able to use it for Portuguese and Hungarian. In Table 1 performance for each technique is given averaged over eight remaining languages. Stemming, when available, is quite effective, and just slightly below the top-ranked approach of character n-grams.

2.3 Morfessor Segments

As it may be difficult to find a rule-based stemmer for every language, a language-independent approach can be quite attractive. The Morfessor algorithm only requires a lexicon (*i.e.,* wordlist) for a language to learn to identify morpheme boundaries, even for previously unseen words. Such automatically detected segments can be an effective form of tokenization [5]. Examples of the algorithm's output are presented in Table 2, along with results for Snowball and character 5-grams.

Compared to plain words the induced morphemes produced by Morfessor led to gains in 9 of 13 languages; 8 of these were significant improvements with $p < 0.05$ (Wilcoxon test). The languages where words outperformed segments were English (dramatically), French, Italian, and Spanish – each is relatively low in morphological complexity. The differences in French and Spanish were

Table 3: Preliminary Results for Morpho Challenge 2008

	English	Finnish	German
Plain words	0.3293	0.3519	0.3509
Snowball	**0.4081**	0.4275	0.3865
Morfessor	0.3861	0.4425	**0.4656**
LCN 5	0.3563	0.3688	0.3276
4-grams	0.3566	**0.4918**	0.4388
5-grams	0.3630	0.4515	0.4331

less than 0.004 in absolute terms. Segments achieved more than a 20% relative improvement in Bulgarian, Finnish, and Russian, and over 40% in Czech and Hungarian.

2.4 Least Common N-gram Stems

Another language-neutral approach to stemming is to select for each word, its least common n-gram. This requires advance knowledge of n-gram frequencies, but this is easily obtainable by constructing a regular n-gram index, or even by scanning a corpus and counting. Lengths of $n = 4$ and $n = 5$ appear about equally effective with a slight advantage for *lcn4*, but this is influenced primarily by the languages with greater morphological complexity, which see larger changes. An 8% relative improvement in mean average precision over words is obtained. As can be seen from Table 1, in languages where rule-based stemming is available its use is preferable. N-gram stemming achieves comparable performance with Morfessor segments..

2.5 Overlapping Character N-grams

N-grams achieve morphological regularization indirectly due to the fact that subsequences that touch on word roots will match. For example, "juggling" and "juggler" will share the 5-grams _jugg and juggl. While n-gram's redundancy enables useful matches, other matches are less valuable, for example, every word ending in 'tion' will share 5-gram tion_ with all of the others. In practice these morphological false alarms are almost completely discounted because term weighting de-emphasizes them. In fact, such affixes can be so common, that ignoring them entirely by treating them as "stop n-grams" is a reasonable thing to do.

Character n-grams are the most effective technique studied here, giving a relative improvement of 18%. Consistent with earlier work [4] lengths of $n = 4$ and $n = 5$ are equally effective averaged across the 13 languages; however there are some noticeable differences in particular languages. The data is suggestive of a trend that the most morphologically variable languages (*i.e.,* Bulgarian, Czech, Hungarian, and Russian) gain more from 4-grams than 5-grams, while 5-grams have a slight advantage in medium complexity languages.

Snowball stems are roughly as effective as n-grams, on average, but only available in certain languages (*i.e.,* 8 of 13 in this study). The other "alternative" stemming approaches, segments and least common n-grams, appear to gain about half of the benefit that full n-gram indexing sees compared to unnormalized word forms.

3 Morpho Challenge 2008 Submissions

This year's evaluation was conducted in English, Finnish, and German, as was the case in Morpho Challenge 2007. The contest was conducted by having participants submit an analysis for each word form, which would be used to create replacement indexing terms in an IR system. According to the contest website the LEMUR IR engine was used with Okapi BM 25 term weighting. Preliminary results reported by the organizers are presented in Table 3 using mean average precision.

An examination of the preliminary results shows the same basic trend identified in Section 2. Rule-based stemming is most effective in English, and in a morphologically rich language such as Finnish, n-grams have a strong advantage.

4 Conclusions

We examined a variety of methods for lexical normalization, finding that the most effective technique was character n-gram indexing which achieved a relative gain of 18% in mean average precision over unlemmatized words. In Czech, Bulgarian, Finnish, and Hungarian gains of over 40% were observed. While rule-based stemming can be quite effective, such tools are not available in every language and even when present, require additional work to integrate with an IR system. When language-neutral methods are able to achieve the same, or better performance, their use should be seriously considered.

References

[1] Mathias Creutz and Krista Lagus. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical report, Helsinki University of Technology, 2005.

[2] Mikko Kurimo, Mathias Creutz, and Ville Turunen. Overview of Morpho Challenge in CLEF 2007. In Alessandro Nardi and Carol Peters, editors, *Working Notes of the CLEF 2007 Workshop*, 2007.

[3] James Mayfield and Paul McNamee. Single n-gram stemming. In *SIGIR*, pages 415–416. ACM, 2003.

[4] Paul McNamee and James Mayfield. Character N-gram tokenization for european language text retrieval. *Information Retrieval*, 7(1-2):73–97, 2004.

[5] Paul McNamee, Charles Nicholas, and James Mayfield. Don't have a stemmer?: Be un+concern+ed. In *SIGIR '08*, pages 813–814, New York, NY, USA, 2008. ACM.

[6] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281. ACM, 1998.