JHU Ad Hoc Experiments at CLEF 2008

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

Abstract

For CLEF 2008 JHU conducted monolingual and bilingual experiments in the ad hoc TEL and Persian tasks.

The TEL task involved focused on searching electronic card catalog records in English, French, and German using data from the British Library, the Bibliotheque Nationale de France, and the Österreichische Nationalbibliothek (Austrian National Library). The approach we adopted for TEL was to strip out non-content sections of records and to treat the task as ordinary full-text search using character n-grams and stemmed words.

For the Persian task, which is based on the *Hamshahri* corpus, several different forms of textual normalization were compared. Using the provided training topics we compared character n-grams, n-gram stems, ordinary words, words automatically segmented into morphemes, and a novel form of n-gram indexing based on n-grams with character skips. On the training topics we found that character 5-grams and skipgrams performed the best and this was borne out in our official submissions.

We also did some post hoc experiments using previous CLEF ad hoc tests sets in 13 languages.

In all three tasks we explored alternative methods of tokenizing documents including plain words, stemmed words, automatically induced segments, a single selected n-grams for each words, and all n-grams from words (*i.e.,* traditional character n-grams). Character n-grams demonstrated consistent gains over ordinary words in each of these three diverse sets of experiments. Using mean average precision, relative gains of of 50-200% on the TEL task, 5% on the Persian task, and 18% averaged over 13 languages from past CLEF evaluations, were observed.

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, Farsi language retrieval

1 Introduction

As a tokenization scheme character n-grams possess many advantages. They work in every language, require no training, and are more effective than plain words. It also appears that n-grams are beneficial for normalizing morphological variation, particularly in languages where words have many related surface forms.

Using test sets in the 13 languages used in the ad hoc tracks at previous CLEF evaluations, we compared n-grams to several tokenization alternatives, including a rule-based stemmer (*Snowball*), an unsupervised morphological segmenter (*Morfessor*), and a synthetic form of stemming based on selecting a single character n-gram from each 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.

Accordingly n-grams were used in official submissions to this year's ad hoc tasks. For this year's participation at CLEF we used the JHU HAIRCUT retrieval system, employing a statistical language model similarity metric with a smoothing constant of 0.5. The similarity calculation combines document term frequencies and corpus frequencies (for smoothing) using linear interpolation with a smoothing constant of 0.5 [9]. For retrieval of Farsi text, we explored a variant of n-gram indexing, *skipgrams*, which are n-gram sequences that omit some letters. Farsi has root and template morphology and it was thought that skipgrams might prove effective.

In Section 2 we describe our experiments for the TEL subtask. In Section 3 we analyze our training experiments and official results for the Persian subtask. In Section 4 some recent experiments on previous CLEF collections are described.

2 TEL task

Our approach to TEL was to treat the collection as unstructured documents. Fields that did not appear to contain good indexable content were removed, including: publisher, rights, format, description, indentifier, contributor, type, language, coverage, issued, available, extent, spatial, and created. Text from the following fields was retained: ispartof, edition, alternative, tableofcontents, abstract, bibliographiccitation, subject, title, abstract, date, creator, source, and relation. All SGML tags were removed.

Some of these choices were probably harmful. For example, queries that specified a particular language or document type (i.e., maps) might have benefitted from some of the deleted metadata. The aim of removing these fields was to increase the coherence of each document's indexable terms.

2.1 Indexing Schemes

The tokenization methods explored were:

- **words**: space-delimited tokens.
- **snow**: output of the Snowball stemmer.
- **morf**: the set of morphemes for each word identified by the *Morfessor* algorithm. Morfessor is available online at http://www.cis.hut.fi/projects/morpho/. 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].
- **lcn4/5**: least common n-gram stem (*i.e.,* rarest word-internal character n-gram) of length $n = 4$ or $n = 5$ [3].
- **4-grams**: overlapping, word-spanning character 4-grams produced from the stream of words encountered in the document or query.
- **5-grams**: length $n = 5$ n-grams created in the same fashion as the character 4-grams.

Common to each tokenization method was conversion to lower case letters, removal of punctuation, and truncation of long numbers to 6 digits.

Table 1: Monolingual Results

	English	French	German	Run designation
words	0.2719	0.2019	0.1073	not submitted
snow	0.3480	0.2290	0.1757	aplmoxxs
morf	0.3171	0.2332	0.1989	not submitted
lcn4	0.3086	0.2223	0.1565	not submitted
lcn5	0.2993	0.2270	0.1810	not submitted
4-grams	0.3382	**0.2950**	**0.3377**	aplmoxx4
5-grams	0.3190	0.2800	0.3102	aplmoxx5
4-grams + RF	**0.3531**	0.2861	0.3176	aplmoxx4rf

2.2 Monolingual Results

Our official submissions were based on 4-grams, both with and without automated relevance feedback, 5-grams (no RF), and stemmed words. Table 1 lists mean average precision for these runs, as well as for several unsubmitted runs. In the official run names *xx* indicates one of de (German), en (English), or fr (French).

While performance did not vary dramatically in English, except for the unnormalized word run which performed the worst, 4-grams were dominant with the French and German collections. Large gains were observed with 4-grams compared to plain words – more than a 50% relative gain in French and over 200% in German.

2.3 Bilingual Results

We considered the following bilingual pairs:

- Dutch to English
- French to English
- German to English
- Spanish to English
- Dutch to French
- English to French
- German to French
- English to German
- French to German

For each language pair the source side query was tokenized using only character 5-grams and those n-grams were 'translated' to the target language using a large aligned parallel corpus (content from the Official Journal of the European Journal). The methodology in query term translation was like that in [5]; however, here no pre-translation query expansion was performed. In Table 2 results are presented using mean average precision to compare performance.

Source language did not make a large difference in performance across the three collections. Bilingual performance was approximately 60% of the highest performing monolingual run, which is a bit lower than we have customarily observed in bilingual retrieval against news corpora at CLEF.

Table 2: Official Bilingual Runs

Source	English	French	German
Dutch	0.2024	0.1746	x
English	x	0.1669	0.1899
French	0.2087	x	0.1829
German	0.2111	0.1608	x
Spanish	0.1856	x	x

3 Persian language task

We made submissions for both the monolingual and bilingual subtasks. The bilingual submissions were based on online machine translation software[1] applied to the queries, so only one set of indexes was required. In addition to the methods in Section X.Y we used skipgrams, 4- or 5-grams with and without one internal skip (denoted by *sk41 & sk51*). Snowball does not support Farsi so no stemming runs were attempted.

3.1 Skipgrams

Consider the present tense conjugation of the Spanish verb contar (*to count*): cuento, cuentas, cuenta, contamos, contáis, and cuentan. Such inflectional variation can cause lexical mismatches that would impair retrieval, and character n-grams are unlikely to be a total solution to this problem since the 1st and 2nd person plural forms do not share longer n-grams with the other forms. Similar problems can happen with nouns, for example, in Welsh plentyn (*child*) and its plural, plant (*children*). These two examples contain patterns that could enable matching, such as regular expressions *c⋆nt* and *pl⋆nt*, which would match all the related forms.

Pirkola et al. [8] have proposed n-grams with skips[2] to match terminology for cross-language information retrieval in languages sharing a common alphabet. For example, the English word *calcitonin* can be matched to its Finnish translation *kalsitoniini*, supported in part by matches like l⋆t and n⋆n. Mustafa [7] proposed a similar method for monolingual Arabic language processing, where infix morphological changes are common. He identified relevant dictionary terms using bigrams with and without a single skip character and a Dice coefficient to compare sets of bigrams. Järvelin et al. [2] formalized the notion of skipgrams and investigated methods of comparing lexical terms; however, they focused on the case where a single skip is formed by deleting contiguous letters. This makes sense when only bigrams are considered – then the only place to skip characters is between the first and last letters of the (skip) bigram.

But with longer n-grams there are multiple places where skips can occur, and character skipgram methods can be generalized even further by including the possibility of multiple non-adjacent skips within a single word (though no such experiments are reported here). In these experiments skipgrams are considered as an alternative method for tokenization that might support matches across morphologically related words. When a letter is skipped we replace that letter in the n-gram subsequence with a special symbol (*i.e.,* a dot character (•). This is done in an attempt to avoid unintended conflations with n-gram strings produced by unrelated words. Skipgram tokenization of length four for the word *cream* would include the regular n-grams *crea* and *ream* in addition to *c•eam*, *cr•am*, and *cre•m*.

3.2 Training Data

The various methods of tokenization were compared on the 50 training topics. In Table 3 runs without relevance feedback are presented along with runs that made use of automated feedback using various numbers of expansion terms.

[1] http://www.parstranslator.net/eng/translate.htm

[2] They use the term s-grams.

Table 3: Training results for Persian (mean average precision)

	No RF	50	100	200	400	800
4-grams	0.3883	0.4199	0.4231	0.4172	-	-
5-grams	0.3810	0.4225	0.4305	0.4280	-	-
words	0.4091	0.4175	0.3999	0.3905	-	-
morfessor	0.3784	0.3951	0.3801	0.3637	-	-
lcn4	0.3914	0.3975	0.3840	0.3730	-	-
lcn5	0.3978	0.3960	0.3779	0.3723	-	-
sk41	0.3886	0.4000	0.4156	0.4332	0.4372	0.4290
sk51	0.3613	0.3607	0.3817	0.4012	0.4216	0.4280

Table 4: Monolingual runs

	No RF	50	100	200	400
words	0.3617	0.4332	0.4299	0.4211	-
morf	0.3559	0.4250	0.4223	0.4156	-
lcn4	0.3629	0.4252	0.4256	0.4180	-
lcn5	0.3506	0.4225	0.4188	0.4085	-
4-grams	0.3986	0.4383	0.4530	0.4564	-
5-grams	0.3821	0.4288	0.4493	0.4558	-
sk41	0.3906	0.3732	0.4053	0.4384	0.4519
sk51	0.3512	0.3238	0.3595	0.4008	0.4250

3.3 Monolingual and Bilingual Results

In Table 4 mean average precision is reported for eight tokenization methods. The n-grams methods are the highest performing approach and the skipgrams perform slightly worse than traditional character n-grams. The highest performing run was character 4-grams using 200 expansion terms which got a MAP score of 0.4564; however the results on the training topics suggested 5-grams would outperform and we selected them instead. N-grams appear to need more query expansion terms than words to maximize performance, and skipgrams, being even more conflationary require more than regular 4- or 5-grams.

The results for our official monolingual and bilingual runs are given in Table 5. Tokenization method did not appear to drastically affect the outcome monolingually; however, words and the Morfessor-based runs did markedly worse on the bilingual task compared to the n-gram based methods.

Table 5: Official runs

	Task	Index	RF Terms	MAP
jhufa5r100	mono	5-grams	100	0.4493
jhufask41r400	mono	sk41	400	**0.4519**
jhufawr50	mono	words	50	0.4332
jhufamr50	mono	morf	50	0.4250
jhuenfa5r100	bi	5-grams	100	0.1660
jhuenfask41r400	bi	sk41	400	**0.1892**
jhuenfawr50	bi	words	50	0.0946
jhuenfamr50	bi	morf	50	0.1112

Table 6: 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**

4 Analysis from 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.[3] 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). In Table 6 results are presented using mean average precision to compare performance. The score for the highest performing technique in each language is emboldened.

4.1 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.

4.2 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.

4.3 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 [6]. 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 –

[3] These results are also reported in our Morpho Challenge 2008 paper in these working notes.

each is relatively low in morphological complexity. The differences in French and Spanish were 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.

4.4 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..

4.5 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.

5 Conclusions

We examined a variety of methods for lexical normalization, finding that the most effective technique was character n-gram indexing. N-grams achieved consistent gains in mean average precision over unlemmatized words. Relative gains of of 50-200% on the TEL task, 5% on the Persian task, and 18% averaged over thirteen languages from past CLEF evaluations, were observed. In languages such as 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] Anni Järvelin, Antti Järvelin, and Kalervo Järvelin. S-grams: Defining generalized n-grams for information retrieval. *Information Processing and Management*, 43(4):1005–1019, 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 and James Mayfield. Translating pieces of words. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, pages 643–644. ACM, 2005.

[6] 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.

[7] Suleiman H. Mustafa. Character contiguity in n-gram based word matching: the case for arabic text searching. *Information Processing and Management*, 41:819–827, 2004.

[8] Ari Pirkola, Heikki Keskustalo, Erkka Leppänen, Antti-Pekka Känsälä, and Kalervo Järvelin. Targeted s-gram matching: a novel n-gram matching technique for cross- and mono-lingual word form variants. *Inf. Res*, 7(2), 2002.

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