# JHU Experiments in Monolingual Farsi Document Retrieval at CLEF 2009

Paul McNamee

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

#### Abstract

At CLEF 2009 JHU submitted runs in the ad hoc track for the monolingual Persian evaluation. Variants of character n-gram tokenization provided a 10% relative gain over unnormalized words. A run based on skip n-grams, which allow internal skipped letters, achieved a mean average precision of 0.4938. Using traditional 5-grams resulted in a score of 0.4868 while plain words had a score of 0.4463.

## 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**

Farsi document retrieval

## 1 Introduction

For CLEF 2009 we participated in the ad hoc Persian task, submitting results only for the monolingual condition. Similar to our experiments at CLEF 2008, these experiments were based on comparing different tokenization methods. The JHU HAIRCUT retrieval system was used with a statistical language model similarity metric [2, 5]:

$$P(D|Q) \propto \prod_{t \in Q} \lambda P(t|D) + (1 - \lambda)P(t|C)$$
 (1)

Though performance might be improved slightly by optimizing choice of  $\lambda$  as a function of to-kenization, for simplicity a smoothing constant of 0.5 was used throughout these experiments. Automated relevance feedback was used for all submitted runs. From an initial pass of retrieval the 20 top-ranked documents were considered and depending on the method of tokenization a different number of expansion terms was used. We submitted five runs based on: (a) plain words; (b) words that were truncated to at most 5 characters; (c) overlapping character n-grams of lengths 4 & 5; and, (d) a variant of character n-gram indexing allowing some letters to be skipped. Common to each tokenization method was conversion to lower case letters, removal of punctuation, and truncation of long numbers to 6 digits.

The tokenization methods examined were:

Table 1: Monolingual Persian Performance

Terms	Query Fields	RF Terms	MAP	P@5	Rel. Found	Runid
words	TD	50	0.4463	0.6560	4072	-
trun5	TD	50	0.4511	0.6280	4100	jhufatr5r50td
4-grams	TD	100	0.4853	0.6560	4119	jhufa4r100td
4-grams	TDN	100	0.4825	0.6480	4082	jhufa4r100tdn
5-grams	TD	100	0.4868	0.7000	4121	jhufa5r100td
5-grams	TDN	100	0.4784	0.6840	4029	-
sk41	TD	400	0.4951	0.6760	4102	-
sk51	TD	400	0.4583	0.6480	4072	-
4-grams + sk41	TD	400	0.4938	0.6800	4108	jhufask41r400td
5-grams $+$ sk $51$	TD	400	0.4597	0.6600	4048	-

- words: space-delimited tokens.
- trun5: truncation of words to at most the first five letters.
- **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.
- sk41: skip n-grams of length n=5 that contain one internal skipped letter. The skipped letter is replaced with a special symbol to indicate the position of the deleted letter. Skipgram tokenization of length four for the word kayak would include the regular n-grams kaya and ayak in addition to  $k \bullet yak$ ,  $ka \bullet ak$ , and  $kay \bullet k$ .
- sk51: skip n-grams of length n=6 that contain one internal skipped letter.
- 4-grams+sk41: both regular 4-grams and sk41 skip n-grams.
- 5-grams+sk51: both regular 5-grams and sk51 skip n-grams.

N-grams are effective at addressing morphological variation, particularly in languages where words have many related surface forms. In recent experiments we have shown that in many languages more than 25% relative gains can be realized compared to unnormalized words when n-grams are used [3]. N-grams have been used previously in Middle Eastern languages. McNamee  $et\ al.$  used them for Arabic retrieval at the TREC-2001 and TREC-2002 evaluations [4] and AleAhmad  $et\ al.$  found character 4-grams to be effective in Farsi [1].

The results of our official runs are presented in Table 1. Compared to the 2008 evaluation we observe somewhat smaller gains with n-grams, about a 10% relative improvement over plain words. The best condition involved the use of character skipgrams, which were also found to be effective in our CLEF 2008 experiments. Despite their computational expense, we like skipgrams because of the slightly fuzzier matching they allow, which we think may be helpful with Farsi morphology. In Farsi some morphemes can be either bound or free, and thus there may, or may not be an intervening space character. There is also extensive derivational compounding where words from different parts of speech are combined.

## 2 Conclusions

Based on our experiments using the 50 queries from the CLEF 2008 evaluation, we expected n-gram based techniques to prove effective and they did. 4-grams, 5-grams, and skipgrams all provided about a 10% relative gain over plain words.

## References

- [1] Abolfazi AleAhmad, Parsia Hakimian, Farzad Mehdikhani, and Farhad Oroumchian. N-gram and Local Context Analysis for Persian Text Retrieval. In *Proceedings of International Symposium on Signal Processing and its Applications*, 2007.
- [2] Djoerd Hiemstra. Using Language Models for Information Retrieval. PhD thesis, University of Twente, 2001.
- [3] Paul McNamee, Charles Nicholas, and James Mayfield. Addressing Morphological Variation in Alphabetic Languages. In SIGIR '09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 75–82, New York, NY, USA, 2009. ACM.
- [4] Paul McNamee, Christine Piatko, and James Mayfield. JHU/APL at TREC 2002: Experiments in Filtering and Arabic Retrieval. In Proceedings of Eleventh Text REtrieval Conference (TREC 2002), 2002.
- [5] David R. H. Miller, Tim Leek, and Richard M. Schwartz. A hidden Markov model information retrieval system. In SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 214–221, New York, NY, USA, 1999. ACM.