Applying Query Expansion techniques to Ad Hoc Monolingual tasks with the IR-n system

Elisa Noguera and Fernando Llopis
Grupo de investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
elisa,llopis@dlsi.ua.es

Abstract

The paper describes our participation in Monolingual tasks at CLEF 2007. We submitted results for the following languages: Hungarian, Bulgarian and Czech. We focused on studying different query expansion techniques: Probabilistic Relevance Feedback (PRF) and Mutual Information Relevance Feedback (MI-RF) to improve retrieval performance. After an analysis of our experiments and of the official results at CLEF 2007, we achieved considerably improved scores by using query expansion techniques for different languages (Hungarian, Bulgarian and Czech).

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Experimentation, Performance, Query Expansion

Keywords

Information Retrieval

1 Introduction

Query expansion (QE) is a technique commonly used in Information Retrieval (IR) [5] [2] to improve retrieval performance by reformulating the original query adding new terms or re-weighting the original terms. Query expansion terms can be automatically extracted from documents or taken from knowledge resources.

In our seventh participation at CLEF, we focused on comparing two different query expansion strategies: Probabilistic Relevance Feedback (PRF) and Mutual Information Relevance Feedback (MI-RF). Specifically, we participated in tasks for the following languages: Hungarian, Bulgarian and Czech.

We used the IR-n system [3]. It is a Passage Retrieval (PR) system which uses passages with a fixed number of sentences. This provides the passages with some syntactical content.

This paper is organized as follows: the next section describes the task developed by our system and the training carried out for CLEF 2007. The results obtained are then presented. Finally, we present the conclusions and the future work.

2 Relevance Feedback

Query expansion techniques as Relevance Feedback (RF) can substantially improve retrieval effectiveness. Most of the IR systems commonly implemented query expansion techniques. RF is usually performed in the following way:

- A search using the original query is performed, selecting the n terms from top-ranked documents.
- The n terms are added to the original query to formulate a new query.
- The new query is performed to produce a new ranked list of documents.

An important factor is how to assign the weight to the selected terms with respect to the terms from the initial query. In this work, we compare two formulas in order to calculate this weight (w_t): Probabilistic and Mutual Information.

2.1 Probabilistic Relevance Feedback (PRF)

This is the term relevance weighting formula proposed by Robertson and Sparck Jones in [4]. The relevance weight of term t is given by:

$$w_t = \frac{(m_t + 0.5) \cdot (n - n_t - m + m_t + 0.5)}{(m - m_t + 0.5) \cdot (n_t - m_t + 0.5)} \quad (1)$$

where n is the number of documents in the collection, m is the number of documents considered as relevants (in this case 10 documents), n_t is the number of documents which the term t appears and m_t is the number of relevant documents in which the term t appears.

w_t will be better for those terms which have a higher frecuency in the relevant documents than the whole collection. Our system allows two query expansion techniques based on this formula: (1) query expansion based on the most relevant passages or (2) the most relevant documents.

2.2 Mutual Information Relevance Feedback (MI-RF)

This is based on the idea the co-ocurrence between two terms can determine the semantic relation that exists between them [1]. The mutual information score grows with the increase in frequency of word co-occurrence. If two words co-occur mainly due to chance their mutual information score will be close to zero. If they occur predominantly individually, then their mutual information will be a negative number. The standard formula for calculating mutual information is:

$$MI(x, y) = log(\frac{P(x, y)}{P(x) \cdot P(y)}) \quad (2)$$

where $P(x, y)$ is the probability that words x and y occur together; $P(x)$ and $P(y)$ are the probabilities that x and y occur individually. The relevance weight w_t of each term t is calculated adding the MI between t and each term of the query.

3 Experiments

This section describes the training process carried out in order to obtain the best features to improve the performance of the system. In CLEF 2007, our system participated in the following Monolingual tasks: Hungarian, Bulgarian and Czech.

The aim of the experimental phase was to set up the optimum value of the input parameters for each collection. CLEF 2005 and 2006 (Hungarian and Bulgarian) collections were used for training. Query expansion techniques were also evaluated for all languages. Here below, we describe the input parameter of the system:

- **Size passage (sp)**: We established two passage sizes: **8 sentences** (normal passage) or **30 sentences** (big passage).
- **Weighting model (wm)**: We use two weighting models: **okapi** and **dfr**.
- **Dfr parameters**: these are c and avg_{ld}.
- **Query expansion parameters**: If **exp** has value 1, this denotes we use PRF based on passages. If **exp** has value 2, the PRF is based on documents. And, if **exp** has value 3, MI-RF query expansion is used. Moreover, **np** and **nd** denote the k terms (nd) extracted from the best ranked passages or documents (np) from the original query.
- **Evaluation measure**: Mean average precision (**avgP**) is the evaluation measure used in order to evaluate the experiments. This value was obtained with the 2006 collections.

Table 1 shows the best configuration for each language:

Table 1: Best results obtained with training data CLEF 2006

language	sp	wm	C	avgld	exp	np	nd	avgP
Hungarian	8	dfr	2	300				0.3182
Hungarian	8	dfr	2	300	2	10	10	0.3602
Hungarian	8	dfr	2	300	3	10	10	0.3607
Bulgarian	30	dfr	1.5	300				0.1977
Bulgarian	30	dfr	1.5	300	2	10	10	0.2112
Bulgarian	30	dfr	1.5	300	3	10	10	0.2179

The best weighting scheme for Hungarian and Bulgarian was dfr. For Hungarian, we used 30 as passage size. For Bulgarian, we set up 8 as passage size. Finally, the configuration used for Czech was the same as for Hungarian (dfr as weighting scheme and 30 as passage size).

4 Results at CLEF 2007

We submitted four runs for each language in our participation at CLEF 2007. The best parameters, i.e. those that gave the best results in system training, were used in all cases. This is the description of the runs that we submitted at CLEF 2007:

- IRnxxyyyyN
  - xx is the language (BU, HU or CZ)
  - yyyy is the query expansion (*nexp*: not used, *exp2*: PRF, *exp3*: MI-RF).
  - N means the tag *narrative* was used.

The official results for each run are showed in Table 2. Like other systems which use query expansion techniques, these models also improve performance with respect to the base system. Our results are appreciably above baseline in all languages. The best percentage of improvement in AvgP is 40.09% for Hungarian.

5 Conclusions and Future Work

In this eighth CLEF evaluation campaign, we compared different query expansion techniques in our system for Hungarian, Bulgarian and Czech (see Table 1). Specifically, we compare two query

Table 2: CLEF 2007 official results. Monolingual tasks.

Language	**Run**	**AvgP**	**Dif**
Hungarian	IRnHUnexp (baseline)	33.90	
	IRnHUexp2	38.94	+14.88%
	IRnHUexp3	39.42	+16.29%
	IRnHUexp2N	40.09	+18.26%
Bulgarian	IRnBUnexp (baseline)	21.19	
	IRnBUexp2	25.97	+22.57%
	IRnBUexp3	26.35	+24.36%
	IRnBUexp2N	29.81	+40.09%
Czech	IRnCZnexp (baseline)	20.92	
	IRnCZexp2	24.81	+18.61%
	IRnCZexp3	24.84	+18.76%
	IRnCZexp2N	27.68	+32.36%

expansion techniques: Probabilistic Relevance Feedback (PRF) and Mutual Information Relevance Feedback (MI-RF).

The results of this evaluation indicate that for the Hungarian, Bulgarian and Czech our approach proved to be effective (see Table 2) because the results are above baseline.

For all languages, the best results were obtained using MI-RF (see Table 2).

In the future we intend to test this approach in other languages such as Spanish. We also intend to study ways of integrating NLP knowledge and procedures into our basic IR system and evaluating the impact.

Acknowledgements

This research has been partially supported by the framework of the project QALL-ME (FP6-IST-033860), which is a 6th Framenwork Research Programme of the European Union (EU), by the Spanish Government, project TEXT-MESS (TIN-2006-15265-C06-01) and by the Valencia Government under project number GV06-161.

References

[1] William A. Gale and Kenneth W. Church. Identifying word correspondence in parallel texts. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 152–157, Morristown, NJ, USA, 1991. Association for Computational Linguistics.

[2] Donna Harman. Relevance feedback revisited. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–10, New York, NY, USA, 1992. ACM Press.

[3] F. Llopis. *IR-n: Un Sistema de Recuperación de Información Basado en Pasajes*. PhD thesis, University of Alicante, 2003.

[4] Stephen E. Robertson and Karen Sparck Jones. *Relevance weighting of search terms*, pages 143–160. Taylor Graham Publishing, London, UK, 1988.

[5] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, New York, NY, USA, 1996. ACM Press.