# Use of Multiword Terms and Query Expansion for Interactive Information Retrieval

Fidelia Ibekwe-SanJuan[1] and Eric SanJuan[2]

[1] ELICO, Université de Lyon 3
4, Cours Albert Thomas, 69008 Lyon, France
ibekwe@univ-lyon3.fr
[2] LIA & IUT STID, Université d'Avignon
339, chemin des Meinajaries, Agroparc BP 1228,
84911 Avignon Cedex 9, France
eric.sanjuan@univ-avignon.fr

**Abstract.** This paper reports our participation in the INEX 2008 Ad-Hoc Retrieval track. We investigated the effect of multiword terms on retrieval effectiveness in an interactive query expansion (IQE) framework. The IQE approach is compared to a state-of-the-art IR engine (in this case Indri) implementing a bag-of-word query and document representation, coupled with pseudo-relevance feedback (automatic query expansion(AQE)). The performance of multiword query and document representation was enhanced when the term structure was relaxed to accept the insertion of additional words while preserving the original structure and word order. The search strategies built with multiword terms coupled with QE obtained very competitive scores in the three Ad-Hoc tasks: Focused retrieval, Relevant-in-Context and Best-in-Context.

## 1 Introduction

The INEX Ad-Hoc track evaluates the capacity of IR systems to retrieve relevant passages from structured documents (XML elements) rather than whole documents. As this is our first participation in INEX, we tested two basic ideas: (i) evaluate the performance of a state-of-art IR engine designed for full document retrieval; (ii) evaluate the effectiveness of multiword terms for representing queries and documents coupled with query expansion (QE) and compare it to a bag-of-word approach coupled with the same QE mechanism. Here, a multiword term is taken to mean a syntactic construct usually associated with a noun phrase. Multiword terms are undeniably richer in information content and are less ambiguous than lone words. Moreover, recent experiments in IR in the biomedical domain, especially the TREC Genomic Track [1] showed that multiword terms and NLP processing hold promise for IR when applied to a corpus from a technical domain with a more homogeneous content. The hypotheses we wished to test were the following:

1. Can multiword terms gathered interactively from the from top $n$ ranked documents returned by an initial query improve retrieval effectiveness?

S. Geva, J. Kamps, and A. Trotman (Eds.): INEX 2008, LNCS 5631, pp. 54–64, 2009.
© Springer-Verlag Berlin Heidelberg 2009

2. More importantly, can a language model that preserves the structure of noun phrases coupled with a QE mechanism perform better than a bag-of-word model coupled with the same QE mechanism?

To implement our different search strategies, we used the Indri search engine in the Lemur package[1]. The rest of the paper is organized as follows: section 2 describes the Ad-Hoc retrieval tasks; section 3 presents our approach for multiword term selection and the different search strategies implemented; section 4 analyzes results and finally section 5 draws some conclusions from our research experiments.

## 2 Ad-Hoc Retrieval Tasks

The official corpus for Ad-Hoc retrieval is the 2006 version of the English Wikipedia comprising 659,388 articles without images [2]. Participants were asked to submit query topics corresponding to real life information need. A total of 135 such topics were collected, numbered from 544-678. A topic consists of four fields: content only field (<CO> or <Title>) with a multiword term expression of the topic; a content only + structure version of the topic (<CAS>) which is the title with indication of XML structure where the relevant elements may be found; a <description> field which is a slightly longer version of the title field; and a <narrative> field comprising a summary with more details about the expected answers. Typically, the narrative would indicate things to eliminate from relevant documents and draw boundaries that can be geographic, spatial, genre or historical in nature. Some title fields contained boolean signs that required systems to explicitly exclude (-) or include (+) certain terms in the relevant answer elements.

<topic id="546" ct_no="8">
<title> 19th century imperialism </title>
<castitle>article[about(., history)]
section[about(., 19th century imperialism)]</castitle>
<description>Describe the imperialism around the 19th century.</description>
<narrative>I am writing a thesis on 19th century imperialism. I am interested in which countries and why they practiced imperialism and how it affected the rest of the world. An element describing earlier or later than 19th century is acceptible if it supports the context of 19th century imperialism. But an element that describes post ww2 imperialism is far off. An element that describes about a history book/theory on the topic is also acceptable, but an element describing a person who is not directly related to the topic is not. E.g. An article about Hitler is acceptable, but not a novelist who fought in ww1.</narrative>
</topic>

**Fig. 1.** Example of a topic in the Ad-Hoc retrieval track

[1] http://www.lemurproject.org/lemur/IndriQueryLanguage.php

The Ad-Hoc track has 3 tasks

1. Focused retrieval: this requires systems to return a ranked list of relevant non-overlapping elements or passages.
2. The Relevant-in-Context (RiC) task builds on the results of the focused task. Systems are asked to select, within relevant articles, several non-overlapping elements or passages that are specifically relevant to the topic.
3. The Best-in-Context (BiC) task is aimed at identifying the best entry point (BEP) to start reading a relevant article.

## 3 Multiword Term Selection and Query Expansion

We first describe the document representation model in section 3.1, then the query representation (3.2) and finally our multiword term selection process (3.3). Section 3.4 describes the different search strategies we implemented using both automatic Indri search as a baseline and different parameters of the Indri QE feature.

### 3.1 Document Representation

The Wikipedia corpus was indexed using the Indri engine. No pre-processing was performed on the corpus. In particular, no lemmatization was performed and no stop word lists were used. The idea was to test the performance of an existing IR engine on raw texts without using any lexical resources. A nice feature of the Indri index is that word occurrences and positions in the original texts are recorded. A multiword term $t$ is represented as an ordered list of nouns, adjectives and/or prepositions, $t = w_n...w_0$, where $w_0$ is necessarily a noun. Thus, a multiword term is not simply a sequence of nominals (nouns and adjectives) but a syntactic construct corresponding to noun phrases in English where the last element is compulsorily a noun and the order of the words must be preserved. These noun phrases should ideally be interpretable out of context, thus correspond to concepts or objects of the real world. Multiword terms are encoded in Indri language using the "#4" operator. Therefore $t$ is encoded as $\#4(w_n...w_0)$. This operator will match any sequence of words in documents with at most 4 optional words inserted into it.

### 3.2 Query Representation

Given a query $Q$, the user selects some (possibly all) multiword terms in $Q$. If several terms are selected, we use the indri belief operator "#combine" to combine these terms. Hence, the initial query $Q$ is translated by the user in an indri query $Q'$ of the form

$$\#combine(\#4(w_{1,n_1}...w_{1,0})...\#4(w_{i,n_i}...w_{i,0}))$$

where:

- $i$ and $n_i$ are integers with $i > 0$.
- $w_{i,k}$ can be a noun, and adjective or a preposition.

We did not make use of the "+, -" boolean operators included in the initial topic description. We also tested the belief operators "#or" that is implemented as the complement of fuzzy conjunction, but its behavior appeared to be more confusing for the document ranking task. For more details on the Indri query language, see[2].

### 3.3 Interactive Multiword Term Selection and Query Expansion

Following an initial query $Q$ to the Indri search engine using only the title field, we consider the top 20 ranked documents based on $Q$ query. The user selects up to 20 multiword terms appearing in these documents. This leads to acquiring synonyms, abbreviations, hypernyms, hyponyms and associated terms with which to expand the original query term. The selected multiword terms are added to the initial Indri query $Q$ using the syntax described in §3.2. This gives rise to a manually expanded query $Q'$ which will be automatically expanded in a $Q''$ query using Indri QE feature with the following parameters: the number $N$ of added terms is limited to 50 and are all extracted from the $D = 4$ top ranked documents using the query $Q'$. Moreover, in the resulting automatic expanded query $Q''$, $Q'$ is weighted to $w = 10\%$. Figure 2 gives an example of multiword query terms used to expand topic 544. These multiword terms were acquired from the top 20 ranked document following the initial query from the title field. This interactive query expansion process required on the average 1 hour for each topic.

These three parameters ($D = 4, N = 50, w = 10$) were optimized on the TREC Enterprise 2007 data on the CSIRO website corpus[3]. Hence, the QE parameters were optimized on a different corpus than the one on which it is being tested now, i.e., Wikipedia.

### 3.4 Search Strategies

We first determined a baseline search which consisted in submitting the text in the title field of queries to Indri, without stop word removal, without attempting to extract any kind of terms, single or multiword. We then devised more elaborate search strategies, including the interactive multiword term selection process described in §3.3. The different search strategies mainly involved using the expanded set of multiword terms with other features of the Indri search engine such as QE and term weighting. These two features were combined with various possibilities of multiword term representation: bag-of-word, fixed structure, term relaxation (allowing insertion of $n$ additional words). The precise

[2] http://www.lemurproject.org/lemur/IndriQueryLanguage.php

[3] Australian Commonwealth Scientific and Industrial Research Organisation, http://www.csiro.au/

```
#combine( #band(#1(nature of life) philosophy)
#1(significance of life)
#1(meaning of life)
#combine(#1(meaning of life) #or(socrates plato aristotle))
#band(#1(meaning of life) philosophy)
#band(#1(meaning of life) existence)
#band(#1(meaning of life) metaphysics)
#band(#1(existence) existentialism)
#band(#2(purpose life) religion)
#band(#2(purpose life) philosophy)
#band(#3(purpose life) religion)
#band(#3(purpose life) philosophy)
#band(#1(reflection of life) philosophy)
#1(philosophy of life)
#1(philosophy of existence)
#combine(#1(philosopher of life) #or(socrates plato aristotle))
#band(#1(source of life) philosophy)
#band(#2(life wheel) philosophy)
#band(#1(center of life) philosophy)
#band(#1(direction of life) philosophy) )
```

**Fig. 2.** Example of an expanded query with multiword terms for topic 544 on the "Meaning of life"

parameters for each implemented search strategy is detailed hereafter. In the official INEX conference, we submitted five different runs for the three Ad-Hoc retrieval tasks. Thus our runs were not differentiated by task. We carried out additional experiments after the INEX's official evaluation in order to further test the effect of term relaxation on the performance of our search strategies. The different search strategies are summarized in table 1.

**Table 1.** Ad-hoc runs

| RunID | Approach |
|---|---|
| ID92_manual | multiword term with Indri with #1, #2 and #or operators |
| *manualExt* | *multiword term with Indri with #4 and #combine operators* |
| ID92_auto | automatic one word query with Indri #combine operator |
| *autoQE* | ID92_auto with automatic Indri Query expansion (QE) |
| ID92_manualQE | ID92_manual with QE |
| *manualExtQE* | *manualExt with QE* |
| ID92_manual_weighting | multiword term with Indri term weighting (TW) |
| ID92_manual_weightingQE | multiword term with Indri TW and QE |

Only strategies whose ID begin by "ID92..." were submitted to the official INEX Ad-Hoc Retrieval evaluation. The search strategies in italics were performed after the official evaluation.

**Baseline bag-of-word search.** We carried out two automatic search strategies labeled "ID92_auto" and "autoQE" respectively, using only the text from the title field of the topic, without stopword removal. These constitute our baseline. "ID92_auto" was submitted to INEX, meanwhile it appeared after evaluation that its scores could be slightly improved using the QE function with default parameters. We thus carried out the additional strategy labelled "*autoQE*".

**Multiword terms with Query Expansion.** In "ID92_manual", the multiword terms gathered during the process described in section 3.3 were combined with operators $\#n$ with $n \leq 2$ ($n = 1$ requires an exact match of the term, $n = 2$ allows for one insertion in the term) and linked by the "#or" operator. In "ID92_manualQE", we combined the above parameters with the QE mechanism. Note that only the selection of multiwords from the initial Indri ranked documents is manual. The QE function in Indri is automatic once the parameters are fixed. After the official evaluation, we ran additional experiments using the same principle but further relaxed the number of words that can be inserted into the multiword terms ($n = 4$). This gave rise to search strategies labeled "*manualExt*" and "*manualExtQE*" respectively. In both cases, we used the belief operator "#combine".

**Query term weighting.** Here, we experimented with "scrapping" the multiword term structure. In "ID92_manual_weighting", the multiword terms in "ID92_manual" were converted into a bag of weighted words in the following way:

1. each word $w$ occurring in at least one query term is used.
2. its weight is set to $c + 0.1 \times m$ where $c$ is the number of query terms with $w$ as head word (for example "*teacher*" in "*head teacher*") and $m$ the number of terms where it appears as a modifier word (for example "*head*" in "*head teacher*").
3. we then used the Indri operator "weight" to combine these words and their weights.

An additional strategy added the QE function to this vector space model representation of the queries thus giving rise to the "ID92_ manual_weightingQE" run.

## 4 Results

Two types of evaluation were provided in the Ad-Hoc retrieval tasks: (i) XML element or passage retrieval, (ii) full article retrieval.

### 4.1 Evaluation Protocol

For the focused task, the official measure is interpolated precision at 1% recall (iP[0.01]). However, results are also calculated for interpolated precision at other early recall points (0.00, 0.01, 0.05 and 0.10). Mean average interpolated precision [MAiP] over 101 standard recall points (0.00, 0.01, 0.02, ..., 1.00) is given as an overall measure.

### 4.2 Focused Retrieval Evaluation

Table 2 shows the scores obtained by all our runs in all three tasks. For each task, a first column shows the score obtained in the official measure while the second column gives the run's rank out of all submitted runs for that task. We will analyze the results of the focused search here. The analysis of the RiC and BiC results is done in sections 4.3 and 4.4 respectively. For the runs done after the evaluation, we can only provide the scores but not their ranks.

**Table 2.** Scores at INEX 2008 ad-hoc tasks

| Task | Focus | | RiC | | | BiC | | |
|---|---|---|---|---|---|---|---|---|
| Measure | iP[0.01] | Rank | gP[1] | MAgP | Rank | gP[1] | MAgP | Rank |
| *manualExtQE* | 0.693 | - | 0.61 | 0.215 | - | 0.61 | 0.225 | - |
| ID92_manualQE | 0.666 | 6th | 0.55 | 0.211 | 3rd | 0.56 | 0.220 | 2nd |
| ID92_manual | 0.642 | 13th | 0.55 | 0.158 | 24th | 0.55 | 0.166 | 14th |
| ID92_manual_weightingQE | 0.622 | 24th | 0.52 | 0.185 | 12th | 0.48 | 0.195 | 6th |
| ID92_manual_weighting | 0.589 | 30th | 0.47 | 0.148 | 32nd | 0.42 | 0.153 | 17th |
| *autoQE* | 0.574 | - | 0.46 | 0.197 | - | 0.43 | 0.201 | - |
| ID92_auto | 0.566 | 38th | 0.44 | 0.171 | 19th | 0.40 | 0.175 | 10th |

For the focused task, 61 runs from 19 different institutions were submitted. Three systems retrieving full articles, including ours were amongst the 10 topmost systems. Four of our search strategies were ranked in the first half of all submitted runs. Our "ID92_manualQE" strategy that combined manual multiword term selection with automatic QE was persistently better than the other four at all levels of recall. It was ranked 4th by institutions and 6th when considering all submitted runs. However one must be cautious when drawing any conclusion from these results as iP[0.01] corresponds roughly to the precision after 1 relevant document has been retrieved. The term weighting strategies which transformed the multiword query terms into a vector space model obtained lower scores although the variant with QE (ID92_manual_weightingQE) performed significantly better than the variant without QE (ID_manual_weighting). The lowest scores were observed for the baseline Indri on single words with or without automatic QE (autoQE, ID92_auto). The additional experiments carried out after official evaluation showed that multiword term relaxation (manualExt) improved our official scores, and that when QE is added (manualExtQE), the score significantly increases from an iP[0.01]=0.674 to iP[0.01]=0.693, slightly surpassing the score obtained by the best system in the focused task with an iP[0.01]=0.690.

**Relaxing the multiword term structure.** Figure 3 takes a closer look at the precision/recall for our search strategies implementing multiword terms with QE. More precisely, this figure compares:

1. a state of art automatic IR system (Indri) using automatic QE features (autoQE),

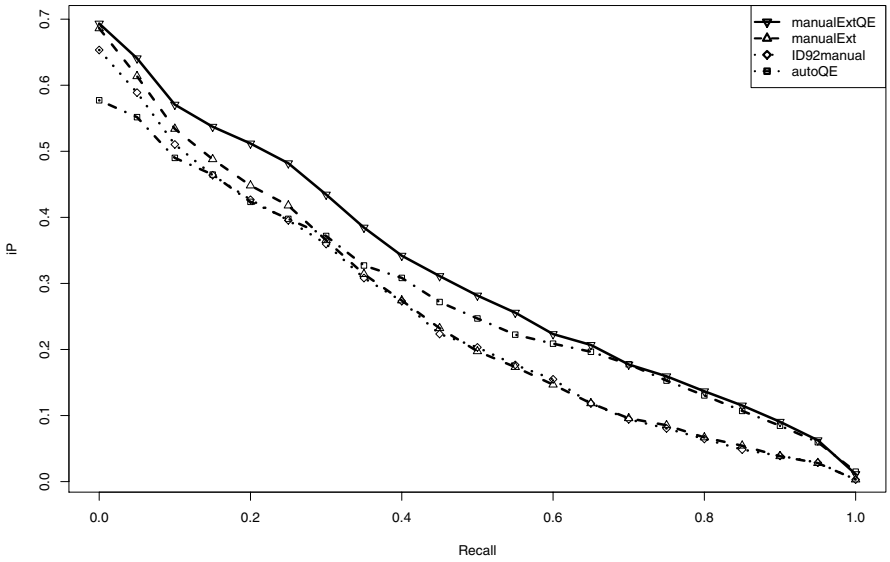



**Fig. 3.** Impact of multiword terms, query expansion and term relaxation on precision/recall for the focused task

2. IR with manually selected multiword terms where term structure and word order are preserved (ID92_manualQE),
3. the same strategy as in (2) but using a relaxed structure of terms by allowing insertion of additional words into the terms (manualExt).
4. the same strategy as in (3) but with automatic QE (manualExtQE).

For low recall levels (iP[0.05] and lower), all strategies with manually selected multiword terms have similar scores and clearly outperform their baseline counterpart. We can see from figure 3 that the two strategies using a more relaxed term structure (manualExtQE, manualExt) performed better than all the others. At iP[0.15], "manualExtQE" implementing the combination of the two features - QE with a relaxed term structure, clearly outperformed all other three runs and consequently all official INEX 2008 evaluated runs. In fact t-Tests with significance level $\alpha$=0.05 show that average score of manualExtQE between iP[0.0] and iP[0.25] is significantly higher than the average score of any of our other search strategies. It follows from these results that a relaxed multiword term structure combined with QE works better than a crisp one.

**Multiword *vs.* bag-of-words representation of queries.** We now study the behaviour of the strategies that implement a vector space model representation of multiword terms combined with term weighting. For that we plot in figure 4 the precision/recall for:

– "ID92_manual_weighting" where all multiword terms were represented by a bag of weighted words;
– its variant "ID92_manual_weightingQE" with automatic QE;
– the former two are compared with our best strategy (manualExtQE) and with the baseline run with QE (autoQE).

The best score for bag-of-word model was obtained by weighting the words according to their grammatical function in the term, i.e., head or modifier word. This is a way to project some of the multiword term structure onto the vector space model. However, even with this improvement, the strategies preserving the structure of multiword terms (manualExtQE, manualExt) significantly outperform the vector space model representation of queries. This is clearly visible in figure 4.

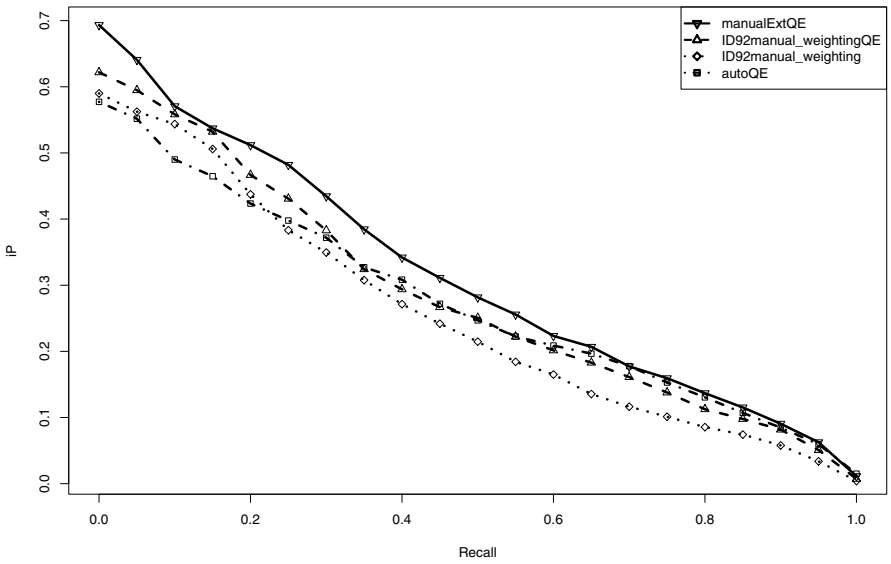



**Fig. 4.** Bag of word query representation *vs* multiword term structure.

It appears that the bag-of-word representation of multiword terms without QE (ID92_manual_weighting) is competitive with the scores obtained by the baseline run (autoQE) on top ranked documents. When we consider higher recall levels (0.25), it performs worse than the baseline.

### 4.3 Relevant-in-Context Task

A total of 40 runs were submitted for this task by all participating groups. The official INEX evaluation once again showed that systems retrieving full articles

instead of XML elements or passages were very competitive [3]. Table 2 shows the scores obtained by all our runs at different recall levels, their MagP and overall ranks.

Our "ID92_manualQE" run was ranked at the 3rd position out of all submitted runs and outperformed all our other runs. This is followed by the "ID92_manualQE". Surprisingly, the additional baseline approach with QE (autoQE) with a MAgP of 0.197 outperformed both the multiword term approach without QE (ID92_manual, MAgP=0.158) and the same approach with weighting and QE (ID92_manual_weightingQE, MAgP=0.185) whereas these two runs had higher precision values at early recall levels (gP[1-5]). It follows that for the Relevant-in-Context measure that combines several levels of recall, multiword terms used alone for queries are not sufficient. It is necessary to enrich them using top ranked documents to increase recall. In fact this phenomenon was also observed in the results of the focused task. Multiword terms queries without QE obtained lower scores than the baseline at higher levels of recall.

### 4.4 Best-in-Context Task

Our search strategies basically conserve the same order of performance as in the RiC task with all runs moving forward to higher ranks (see table 2). Particularly noticeable is the good performance of the "ID92_manualQE" run, ranked 2nd out of 35 submitted runs. The relaxed version "manualExtQE" does even better with a MAgP=0.225, thereby slightly outperforming the best system in the official evaluation (MAgP=0.224) at this task. Surprisingly again, the score of "ID92_auto" is among the 10 best systems (MAgP=0.175). When the QE mechanism is added (autoQE), it obtains a MAgP score of 0.201 thereby outperforming the system ranked 5th in the official evaluation (MAgP=0.120).

### 4.5 Document Retrieval Evaluation

INEX official evaluation also provided judgements full article retrieval. Retrieved elements or passages were ranked by descending order of relevance and judged on a first-come, first-served basis. Hence an element or passage represents the first occurrence of the document from which it was taken. For runs retrieving full articles, it was the classical case of document ranking. Evaluation was carried out over all submitted runs irrespective of task. A total of 163 submitted runs were ranked. Precision scores were calculated also at early recall levels of 5, 10 while mean average precision (MAP) was used as the official measure.

Table 3 shows the evaluation scores for our best strategies. Among the 163 runs that were submitted by participating groups, our "manual ID_92manualQE" strategy with a map of 0.3629 was ranked at the 3rd position. Also, this same strategy with relaxed term structure "manualExtQE" gives a score (map=0.3796) slightly better than the best ranked system (map=0.3789) and significantly outperforms our baseline "autoQE" (map=0.3700) from P5-P30 recall levels.

The reason for this very good performance of the "autoQE" run could be because qrels have been simply derived from those for focused task by considering

**Table 3.** Scores for full document retrieval. Total runs submitted: 163.

| Participant | Rank | P5 | P10 | 1/rank | map | bpref |
|---|---|---|---|---|---|---|
| *manualExtQE* | - | 0.6580 | 0.5942 | 0.8742 | 0.3796 | 0.4076 |
| *autoQE* | - | 0.6171 | 0.5471 | 0.8055 | 0.3700 | 0.3724 |
| p92-manualQEin | 3rd | 0.6371 | 0.5843 | 0.8322 | 0.3629 | 0.3917 |

that any document with a single relevant passage is relevant regardless of the size of the relevant passage within the document. On the contrary, the Focused and RiC measures takes the portion of the relevant passages into consideration.

## 5 Concluding Remarks

In this study, we tested the assumption that query and document representation with multiword terms, combined with query expansion (QE) can yield very competitive results. We tested this hypothesis against two baseline strategies implementing the bag-of-word representation using the Indri search engine with QE feature. The results obtained on the Wikipedia corpus in the three Ad-Hoc Retrieval tasks are very promising. All the search strategies implementing a multiword representation of queries and documents with QE were consistently ranked among the top five systems in the official INEX evaluation and outperformed the baseline strategies adopting a bag-of-word representation, even combined with QE. On the whole, our experiments have shown that using manually expanded multiword terms which are further expanded automatically with a query expansion mechanism is a promising research direction for IR when dealing with topically homogenous collection of texts such as Wikipedia articles. In the future, we intend to address how the interactive multiword term selection process may be automated.

## References

1. Ruch, F., Tbahriti, I., Gobeill, J., Aronson, A.: Argumentative feedback: A linguistically-motivated term expansion for information retrieval. In: Proceedings of the Joint Conference COLING-ACL 2006, Sydney, July 17-21 (2006)
2. Denoyer, L., Gallinari, P.: The wikipedia xml corpus. In: SIGIR Forum, p. 6 (2006)
3. Kamps, J., Geva, S., Trotman, A., Woodley, A., Koolen, M.: Overview of the inex 2008 ad hoc track. In: PreProceedings of the 15th Text Retrieval Conference (INEX 2008), Dagstuhl, Germany, December 15-18, pp. 1–27 (2008)