Multi Word Term Queries for Focused Information Retrieval.

Eric Sanjuan, Fidelia Ibekwe-Sanjuan

Multi Word Term queries for focused Information Retrieval

Abstract. In this paper, we address both standard and focused retrieval tasks based on comprehensible language models and interactive query expansion (IQE). Query topics are expanded using an initial set of Multi Word Terms (MWTs) selected from top n ranked documents. MWTs are special text units that represent domain concepts and objects. As such, they can better represent query topics than ordinary phrases or n-grams. We tested different query representations: bag-of-words, phrases, flat list of MWTs, subsets of MWTs. We also combined the initial set of MWTs obtained in an IQE process with automatic query expansion (AQE) using language models and smoothing mechanism. We chose as baseline the Indri IR engine based on the language model using Dirichlet smoothing. The experiment is carried out on two benchmarks: TREC Enterprise track (TRECent) 2007 and 2008 collections; INEX 2008 Ad-hoc track using the Wikipedia collection.

1 Introduction

Previous experiments carried out within the framework of TREC [1] tended to conclude that retrieval performance has not been enhanced by adding NLP, especially syntactic level of processing. The problem lies in determining the level of NLP needed, on which text units to implement it, whether to implement NLP on both queries and documents and at what stage (whole collection or only on an initial set of returned documents). Previous research also concluded that a deep syntactic representation of queries and documents is not useful to achieve a state-of-the-art performance in IR [2]. It may on the contrary degrade results. On the other hand, performance can be boosted by better representing queries and documents with longer phrases using shallow NLP. In some cases, even a well-tuned n-gram approach can approximate the extraction of phrases and may suffice to boost retrieval performance.

Up until 2004, the dominant model in IR remained the bag-of-words representation of documents which continued to show superior performances in IR. However, a series of experiments carried out on several document collections over the past years are beginning to show a different picture. Nothwithstanding the apparent success of the bag-of-word representation in some IR tasks, it is becoming clear that certain factors related mostly to query length and document genre (general vs technical) influence the performance of IR systems. For instance, [1, 3] showed that representing queries and document by longer phrases can improve systems' performances since these text units are inherently more precise and will better disambiguate the information need expressed in the queries than lone words.

Furthermore, [1] concluded that the issue of whether or not to use NLP and longer phrases would yield better results if focused on query representation rather

than on the documents themselves because no matter how rich and elaborate the document representation, a poor representation of the information need (short queries of 1-2 words) will ultimately lead to poor retrieval performance.

Based on these earlier findings, we wish to investigate the issue of representing queries with a particular type of phrase which are Multiword Terms (MWTs). MWTs is understood here in the sense defined in computational terminology [4] as textual denominations of concepts and objects in a specialized field. Terms are linguistic units (words or phrases) which taken out of context, refer to existing concepts or objects of a given field. As such, they come from a specialized terminology or vocabulary [5]. MWTs are thus terms of length >1. MWTs, alongside noun phrases, have the potential of disambiguating the meaning of the query terms out of context better than single word terms or statistically-derived n-grams and text spans. In this sense, MWTs cannot be reduced to words or word sequences that are not linguistically and terminologically grounded. An initial selection of MWTs from queries is used in an Interactive Query Expansion (IQE) process to acquire more MWTs from top n-ranked documents. The expanded set is submitted to standard IR Language Models for document ranking. Our approach is tested on two corpora: the TREC Enterprise track 2007 and 2008 collections, and INEX 2008 Ad-hoc track. We chose as baseline against which to compare our IQE approach, an IR engine based on the language model using Dirichlet smoothing. The Indri IR system[6] in its default mode applies this language model. Indri was also used as baseline in TREC terabyte[1]. The idea was to test our IQE approach against a strong baseline that competes favorably with the best systems in current IR evaluation campaigns. The results obtained on the Wikipedia corpus in the INEX Ad-hoc track are particulary promising.

The rest of the paper is structured as follows. Section §2 presents our language model and its application to the IR tasks. Section §3 describes the application of our IR model to the TREC Enterprise track 2007 and 2008 collections for document search task. Section §4 presents the focused retrieval tasks on the Wikipedia collection in the INEX 2008 Ad-hoc track. Finally, section §5 discusses lessons learned from these experiments.

2 Combining Automatic and Interactive Query Expansion

2.1 Language Model

Language models are widely used in NLP and IR applications.In the case of IR, smoothing methods play a fundamental role [7]. We shall first describe the probability model that we use.

Document Representation: probabilistic space and smoothing Let us consider a finite collection $\mathcal{D}$ of documents, each document D being considered as

[1] `http://stefan.buettcher.org/trec-tb/`

a sequence $(D_1, ..., D_{|D|})$ of $|D|$ terms D_i from a language $\mathcal{L}$, i.e. $\mathcal{D}$ is an element of $\mathcal{L}^\star$, the set of all finite sequences of elements in $\mathcal{L}$. Our formal framework is the following probabilistic space $(\Omega, \wp(\Omega), P)$ where Ω is the set of all occurrences of terms from $\mathcal{L}$ in some document $D \in \mathcal{D}$ and P is the uniform distribution over Ω LMs for IR rely on the estimation of the a priori probability $P_D(q)$ of finding a term $q \in \mathcal{L}$ in a document $D \in \mathcal{D}$. We chose the Dirichlet smoothing method because it can be viewed as a maximum *a priori* (MAP) document probability distribution. Given an integer μ, it is defined as:

$$P_D(q) = \frac{f_{q,D} + \mu \times P(q)}{|D| + \mu} \tag{1}$$

Query Representation and ranking functions Our purpose is to test the efficiency of MWTs in standard and focused retrieval compared to a classic bag-of-word model and statistically-derived phrases. For that, we shall consider phrases (instead of single terms) and a simple way of combining them. Given a phrase $s = (s_0, ..., s_n)$ and an integer k, we formally define the probability of finding the sequence s in the corpus with at most k insertions of terms in the following way. For any document D and integer k, we denote by $[s]_{D,k}$ the subset of $D_i \in D$ such that: $D_i = s_1$ and there exists n integers $i < x_1, ..., x_n \leq i+n+k$ such that for each $1 \leq j \leq n$ we have $s_j = D_{x_j}$.

We can now easily extend the definition of probabilities P and P_D to phrases s by setting $P(s) = P([s]_{.,k})$ and $P_D(s) = P_D([s]_{D,k})$. Now, to consider queries that are set of phrases, we simply combine them using a weighted geometric mean for some sequence $w = (w_1, ..., w_n)$ of positive reals. Unless stated otherwise, we shall suppose that $w = (1, ..., 1)$, i.e. the normal gometric mean. Therefore, given a sequence of weighted phrases $Q = \{(s_1, w_1), ..., (s_n, w_n)\}$ as query, we shall rank documents acording to the following scoring function $\Delta_Q(D)$ defined by:

$$\Delta_Q(D) \stackrel{\text{rank}}{=} \sum_{i=1}^{n} \left(\frac{w_i}{\sum_{j=1}^{n} w_j} \times \log(P_D(s_i)) \right) \tag{2}$$

This plain document ranking can easily be computed using any passage information retrevial engine. We chose for this purpose the Indri engine since it combines a language model (LM) with a bayesian network approach which can handle complex queries.

2.2 Query Expansion

We propose a simple QE process starting with an approximative short query $Q_{T,\mathcal{S}}$ of the form $(T, \mathcal{S})$ where $T = (t_1, ..., t_k)$ is an approximative document title consisting of a sequence of k words, followed by a possibly empty family of sets of phrases: $\mathcal{S} = \{S_1, ..., S_{|S|}\}$ where for each $1 \leq i \leq |S|$, S_i is of the form $\{S_{i,1}, ..., S_{i,l_i}\}$ for some $l_i \geq 0$. If $l_i = 0$ then S_i is considered to be the empty set. In our case, each $S_{i,j}$ will be a MWT.

Baseline document ranking function By default, we shall rank documents acording to $\Delta_{T,\mathcal{S}} = \Delta_T \times \prod_{i=1}^{|\mathcal{S}|} \prod_{j=1}^{|l_i|} \Delta_{S_{i,j}}$. Therefore, the larger $\mathcal{S}$ is, the less the title part T is taken into account. Indeed, $\mathcal{S}$ consists of coherent subsets of MWTs defined by the user. If the user can expand the query by finding coherent clusters of terms, then we are no more in the situtation of a vague information need and documents should be first ranked according to precise MWTs. For our baseline, we shall generally consider $\mathcal{S}$ to be empty or made of phrases automatically generated from T.

Interactive Multiword Term Selection The IQE process works in the following manner. We consider the top twenty ranked documents of Δ_Q ranking. The user selects a family $\mathcal{S}'$ of several subsets $S'_1, ..., S'_s$ of MWTs appearing in these documents. This leads to acquiring sets of synonyms, abbreviations, hypernyms, hyponyms and associated terms with which to expand the original query terms. We also let the user check that these terms do not introduce noise by adding them individually to the initial query and observing the top ranked documents. The selected multiword terms S'_i are added to the initial set $\mathcal{S}$ to form a new query $Q' = Q_{T,\mathcal{S}\cup\mathcal{S}'}$ leading to a new ranking $\Delta_{Q'}$ computed as previously in §2.2. We emphasize that $\mathcal{S}'$ is more than a flat list of MWTs. In our experiments we also evaluate if the structure of $\mathcal{S}'$ (i.e., grouping the MWTs into subsets) is relevant or not.

Automatic Query expansion We also experimented with the automatic query expansion (AQE). In our model, it consists in the following. Let $D_1, ..., D_K$ be the top ranked documents by the initial query Q. Let $C = \cup_{i=1}^{K} D_i$ be the concatenation of these K top ranked documents. Terms c occuring in D can be ranked according to $P_C(c)$ as defined by equation (1). We consider the set E of the N terms $\{c_1, ..., c_N\}$ having the highest probability $P_C(c_i)$. We then consider the new ranking function Δ'_Q defined by $\Delta'_Q = \Delta_Q^{\lambda} \times \Delta_E^{1-\lambda}$ where $\lambda \in [0, 1]$.

Unless stated otherwise we shall take $K = 4$, $N = 50$ and $\lambda = 0.1$. We now explore in which context IQE based on MWTs is efficient. Our baseline is automatic document retrieval based on equation 2 in §2.1.

3 Enterprise search

The goal of the TREC enterprise track (TrecEnt) was "*to conduct experiments with enterprise data that reflect the experiences of users in real organizations*" [8]. This track ran from 2004 to 2008. We participated in the 2008 edition but "trained" our search strategies beforehand on the 2007 data. Hence, we will indicate performances obtained on data from both years.

3.1 Document collection and Tasks

In 2007, the TrecEnt track chose the CSIRO Enterprise Research Collection (CERC) which is a crawl of all the `*.csiro.au` public websites performed in

march 2007[2]. The collection consists of 370,715 documents totaling 4.2 gigabytes. The search topics used in the TrecEnt tasks were furnished by employees of CSIRO in charge of science communication. These topics correspond to real world information needs received by the CSIRO staff from the public. Thus participating IR systems were judged on real life information needs and not on artificially contrived queries. The submitted runs were evaluated by the community based on the final answer furnished by CSIRO staff to the original requester. Figure 1 gives an example of a topic from TrecEnt 2008.

```
<top>
<num>CE-051</num>
<query>weatherwall</query>
<narr>Have been trying to access the CSIRO weatherwall site to check on weather in
Melbourne over the last 24 hours. It seems to be off line at present. Any idea why?
When might it be back on line? </narr>
</top>
```

Fig. 1. Example of a topic in the TRECEnt 2008 track.

We designed four basic search strategies, called "runs" in the TREC terminology. These four runs were applied on the 2007 and 2008 TrecEnt collections as well on the INEX Ad-hoc tasks albeit with some variations. The first run is the baseline defined in §2.2 using only the query fields. The second is a boosting of this baseline by simply repeating queries in the $\mathcal{S}$ component as phrases. Clearly, instead of leaving $\mathcal{S}$ empty,$\mathcal{S}$ is the singleton $\{\{q\}\}$ made of the query phrase q. The last two runs are based on the IQE process described in 2.2. We give below the precise details of each run:

- **baseline bag-of-words (baseline-B)**: we set $T = \{q_1, ..., q_n\}$ where the q_i are the terms in topic query field q. $\mathcal{S}$ is left empty. This is the usual multinomial bag-of-word approach.
- **baseline phrases (baseline-P)**: we keep the same T but $\mathcal{S}$ is set to the singleton $\{\{(q_1, ..., q_n)\}\}$ whenever the query contains at least two words, i.e. in addition to the bag-of-words approach, we also consider the query q as a phrase.
- **IQE MWT-groupings (IQE-C**): this run corresponds to the IQE approach described in §2.2 except that the user creates sub-groups of MWTs, hence providing a hierarchy of sorts among MWTs. We set $\mathcal{S}$ to $\mathcal{S}(t)$ for each topic. The T component is unchanged.
- **IQE MWTs flat list (IQE-L)**: we consider as $\mathcal{S}$ a flat version of each $\mathcal{S}_t$ where all the selected MWTs are considered at the same level, the internal structure of $\mathcal{S}(t)$ is ignored.

The $\mathbf{IQE-L}$ run evaluates the impact of $MWTs$ on document ranking while the **IQE-C** run, also based on MWTs, evaluates the impact on the retrieval

[2] the Australian 'Commonwealth Scientific and Industrial Research Organization'

effectiveness of forming subsets of $MWTs$ by the user. We illustrate these two representations of MWTs on the same topic as in figure 1. For the **IQE-C run**, the user formed these subsets of MWT queries:

1. {weatherwall}
2. {(weatherwall site), weather, Melbourne}
3. {(CSIRO weatherwall site), weatherwall, (weather in Melbourne)}

In this representation, the particular angle by which the MWT is sought is reflected by a facet term placed to the right of it, e.g. *((weatherwall site), weather, Melbourne))*. In the **IQE-L run**, the expanded query is represented by this flat list of MWTs: *((weatherwall site), (CSIRO weatherwall site), (weather in Melbourne), weatherwall, weather, Melbourne*). This is a simplified version of the same MWTs used in the IQE-C run in which the facet terms have been removed. All terms are weighted equally here.

3.2 Results based on usual Average Precision

The official measure for the TrecEnt 2007 edition was Average Precision (AP). This was changed to *inferred* Average Precision (*inf*AP) for TRECEnt 2008. However, we can compute AP on both tracks.

Document search on the TrecEnt 2007 collection 50 topics were provided and all were judged. On the resulting document qrels, our baseline reaches a mean average precision (MAP) of 0.441 which outperforms all reported runs in [8], the highest MAP being 0.422. However, based on the query by query average precision (AP) score, there is no statistical evidence (t-test with a 95% confidence interval) that our baseline has a true mean not equal to 0.422. Since TrecEnt queries were short phrases most of which had the appearance of MWTs like *"solve magazine, selenium soil"*, the question was to ascertain if our baseline can be boosted by considering phases as suggested by [3]. It seems the answer is yes, but only slightly since the *phrases* run reaches the MAP score of 0.448.

Document search on the TrecEnt 2008 collection 77 topics were made available to participants of which 67 were judged. Four had no judged relevant documents and were dropped. The same IQE process was implemented in which a user selected for each topic t, subsets $\mathcal{S}(t)$ of MWTs following the methodology described in §2.2.

We first computed the AP measures used in TrecEnt 2007 in other to compare our baseline to its performance on this data. Confirming its good perfomance in 2007, our *baseline-B* run implementing the bag-of-word approach outperformed all our other approaches. The good performance of our *baseline-B* here confirms that it is indeed a strong one since it reaches similar precision scores at 10% of recall and even higher at 20% of recall. The 2008 curves then drop because TrecEnt 2008 qrels are based on a more complex pooling process that handicaps low ranked documents in participant runs. In fact, it appears that our two

baseline runs ranked first the "easiest to find" relevant documents among these qrels. These are documents found by most participants.

3.3 Results based on Inferred Average Precision

The inferred AP (infAP) measure used in TRECEnt 2008 is similar to the original infAP used in the TREC Terabyte track, except that it has been modified to work on stratified samples. Both versions of infAP take into account the fact that the measurement is based on a pool of relevant documents and not on an exhaustive list of all relevant documents. Indeed, AP relies on the knowledge of the complete set of relevant documents which on a large corpus is not generally known. According to NIST organizers of the TrecEnt 2008, *"two runs were pooled out from each group to depth 100. The documents were selected for judging by taking a stratified sample of that pool based on document ranks: documents retrieved at ranks 1-3 were sampled at 100% depth, documents of ranks 4-25 at depth 20%, and document between 25-75 rank were sampled at 10% depth. The rank of a document for sampling purposes is the highest rank over all pooled runs. "* The evaluation script and relevance judgments are available from the TREC website[3]. The script also allows us to estimate the usual Normalized Discounted Cumulated Gain (NDCG) that gives more importance to elements at higher ranks. Figure 2 shows the inferred AP and NDCG of our baseline and IQE runs.

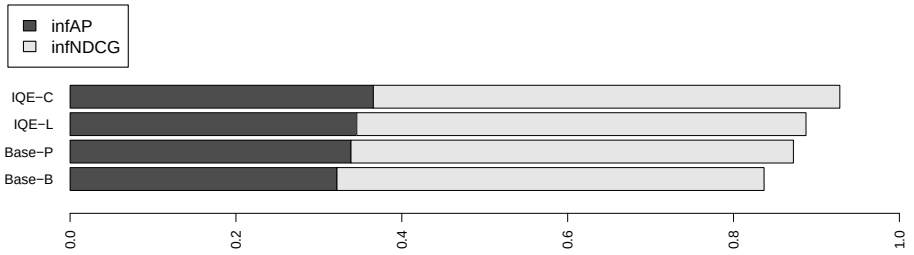

Fig. 2. Inferred Average Precision and Normalized Discounted Cumulated Gain on TrecEnt 2008 qrels using available sampling information.

On the resulting 2008 stratified qrels, our *baseline-B* run attains an infAP score of 0.3218 thus placing itself among the six best runs submitted to TrecEnt 2008. In contrast with previous results on absolute AP, the infAP goes up to 0.3387 when considering phrases in *baseline-P* run, 0.345 when considering *IQE-L* run based on the flat list of additional terms and 0.3657 for *IQE-C* run using the grouped set $\mathcal{S}(t)$ of MWTs. Therefore, using the infAP measure, our IQE-MWTs runs outperform the baseline bag-of-word and phrase runs.

[3] `http://http://trec.nist.gov/data/t17_enterprise.html/`

However, only the difference between the first *baseline-B* and other runs is statistically significant (t-test at 95% of confidence). Other differences are not significant. Since the *baseline-P* run is in fact the *baseline-B* boosted by adding the whole topic query as a phrase to the initial bag of words query, these results show that [3]'s observations that document retrieval performance can be boosted on large web collections by considering phrases, are also true on smaller enterprise web corpus.

4 Focused retrieval

The focused retrieval experiment was carried out in the framework of INEX 2008 Ad-hoc track which is the main forum for researchers working on the extraction of information from structured documents, mostly XML [9].

4.1 INEX 2008 Ad-hoc track

Corpus and topics The official INEX 2008 corpus was the 2006 version of the English Wikipedia comprising 659,388 articles without images [10]. On average, an article contains 161 XML nodes, where the average depth of a node in the XML tree of the document is 6.72. From this corpus, participants were asked to submit query topics corresponding to real life information needs. A total of 135 such topics were built, numbered from 544-678. 70 out of them were judged by the community and thus used in the official evaluation. A topic consists of four fields: content only field (<CO> or <Title>) with a multi-word 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.

Ad-Hoc Retrieval Tasks The 2008 Ad-Hoc track had 3 tasks: Focused retrieval, Relevant-in-Context (RiC), Best-in-Context (BiC).

1. The focused task requires systems to return a ranked list of relevant non-overlapping elements or passages. This is called the "fetching phase".
2. The Relevant-in-Context (RiC) task builds on the results of the focused task. This task is based on the assumption that a relevant article will likely contain relevant information that could be spread across different elements. This is called the "browsing phase". Systems are therefore 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. This task is based on the assumption that "even an article completely devoted to the topic of request will only have one best starting point from which to read (even if that is the beginning of the article)" [11].

Extended qrels and evaluation measures The evaluation procedure establishes an extended qrel file similar to those used in TREC against which all participating systems are evaluated. Like in TREC Terabyte and Ad-hoc tracks, the procedure consists in selecting for each query a pool of documents from participant runs. Topics and documents are then randomly distributed to assessors from the INEX community. Using an ergonomic java on-line interface, each assessor has to mark-up for each document, the relevant passages with regard to a topic. It is important to emphasize that query terms are highlighted in the display of documents. Moreover, in 2008, the interface offered the facility of selecting the whole document using a simple radio button. The assessor had also to point out the BEP. These result in a qrel file that gives for each evaluated pair of topic and document, the total length of relevant passages, the document length, the offset of the BEP and the list of relevant passages. Lengths are computed as number of characters in the text version of the corpus (without XML tags). The 2008 qrel file required the evaluation of $36,605$ articles. Among them, only $4,773$ were judged to contain at least one relevant passage for at least one topic. However, it appears that 40% of these $4,773$ documents have at least 95% of their content marked as relevant by assessors. These highly relevant documents only cover 0.02% of the total length of evaluated documents but almost 25% of the total length of relevant passages. These facts are important to estimate the upper AP bound for systems retrieving full document instead of passages or XML elements.

The RiC and BiC are also evaluated based on these qrels but using graded document scores whereas in the focused task, scores are based on the sole relevant passages no matter their co-occurrence in documents. Given a document score function S into $[0,1]$, both RiC and BiC evaluations are based on generalized precision gP at some rank r which is the average score S over the r scores documents. Given a document d, the score $S(d)$ is in the case of:

- RiC, the F-score of the retrieved passages from d by the system among all relevant passages in d.
- BiC, a normalized distance in number of caracters between the BEP found by the system and the real one.

The consequence is that these measures favour even more full document retrival strategies against passage retrieval since for 40% of relevant documents, full document retrieval strategies will obtain the maximal score whenever they retrieve relevant documents. We refer to [11] forfurther discussion of these measures.

4.2 Results

We first present our search strategies, then analyze results by tasks in the INEX Ad-hoc track.

Runs We consider the same four basic strategies as in the TREC Enterprise search track: *baseline bag-of-words (baseline-B), baseline phrases (baseline-P),*

IQE MWTs subsets (IQE-C) and *IQE MWT flat list (IQE-L)*. Like in the TrecEnt experiment, the two first runs are automatic, the last two rely on the sets of MWTs manually gathered when browsing the top ranked 20 documents based on an initial query. Table 1 gives an example of such expansion.

IQE-LC with subsets of MWTs	resulting flat list for IQE-C
{(dna testing) disease}	(dna testing)
{(dna testing ancestry)}	(dna testing ancestry)
{(genetic disease), (dna testing) ancestry}	(genetic disease)
{(hereditary disease) (dna testing) ancestry}	(hereditary disease)

Table 1. Selected multiword terms for the INEX 2008 topic "dna testing forensic maternity paternity".

Compared to the TrecEnt runs, there are two differences in the way that we apply these runs here: 1) we do not use any stemmer, nor lemmatization and we index all the text (no stop word list), 2) we systematically apply AQE to all runs.

Indeed, Wikipedia articles are well written, with very few spelling errors, thus any stemming will induce a loss of information whereas on the CSIRO web pages, stemming tended to reduce the noise. AQE on the non lemmatized Wikipedia corpus was able to automatically capture synonyms and some grammatical variants of the query term. On the CSIRO corpus used in TrecEnt, AQE just added more noise.

Focused task The INEX 2008 official measure for focused task was average interpolated Precision at 1% of recall (iP[0.01]). Figure 3 shows the Recall/Precision curves of our baseline and IQE runs. The best score for all runs in the official evaluation was 0.6896. Our *baseline-B* score (automatic run with AQE) obtains a significantly much lower score at 0.5737. The *baseline-P* run did not benefit from the same boosting effect as in TRECEnt experiment, hence its much lower score of 0.5732. The *IQE-L* run obtained a much higher score of 0.7016, even higher than the best participating system. This score is further improved to 0.7137 when we consider the *IQE-C* run in which MWTs had been grouped to reflect more complex query representations (see table 1 for an example).

The differences between IQE-based runs are not statistically significant, whereas the difference between *baseline* runs and the IQE runs is this time clearly significant. Indeed, using the Welch Two Sample paired t-test, we find a p-value of 0.02302. Moreover, other participants' best runs submitted at INEX 2008 were optimal for very low recall values but then drop down fast for higher recall values. One might put forward the argument that the good score of our IQE runs may be due to the fact that the user found one or two completely relevant documents with some specific MWTs which were then re-introduced in the expanded

query. The Precision/Recall curves in Figure 3 show that this was not the case. In fact, mean average iP for the *baseline* runs is only 0.28 while that of both both IQE runs reach 0.34. The difference is again statistically significant at 95% of confidence with an estimated p-value of 0.03966. Therefore, this experiment clearly demonstrates that representing queries with MWTs corresponding to real concepts instead of n-grams or bag-of-words, can dramatically improve IR when dealing with a high quality collection such as the Wikipedia. We now present results for the other two tasks of the Ad-hoc track.

Relevant-in-Context and Best-in-Context tasks The official measure for these tasks was MAgP (Mean Average generalized Precision). By considering that we only retrieve articles that are completely relevant, and that the best entry point is the first character of the document, the same four runs can be evaluated with regard to the RiC and BiC measures.

Our runs maintained the same order as it can be observed in figure 3. Among all submitted runs to INEX 2008, the best score was 0.228 for RiC and 0.224 for BiC. Our *baseline* already reaches a score of 0.197 for RiC and 0.20 for BiC. This places our baseline among the six best runs and our group among the three best teams. The baseline is slightly improved by considering *phrases*: 0.2 for RiC, 0.206 for BiC. The scores of IQE outperform the best scores in the official evlauation. Indeed, the *IQE-L* run reaches a score of 0.236 for RiC and 0.248 for BiC. Surprisingly, *IQE-C* run does not improve these score since it obtains a score of 0.235 for RiC and 0.246 for BiC. However, none of these differences are statiscally significant at 95% of confidence, the Welch Two Sample t-test p-value between the *baseline* and the *IQE-L* runs being 0.08739 for RiC and 0.05981 for BiC. Classical MAP was also computed at INEX 2008 by considering as relevant any document involving at least one relevant passage, whatever its length. There, we also find that IQE runs also outperform all other runs, but the difference with the baseline is even less significant.

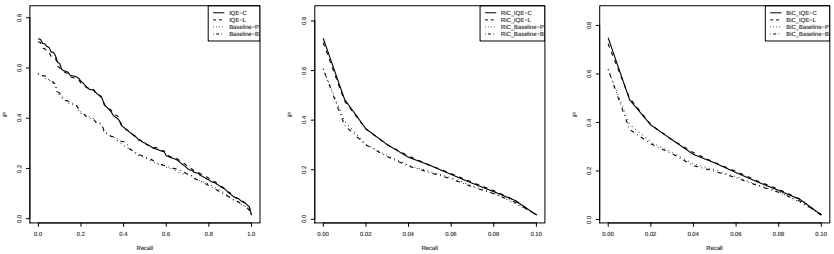

Fig. 3. Interpolated generalized precision curves on INEX 2008 topics for Focused (left) Relevant in Context (center) and Best in Context (right)

5 Conclusions

We have presented in this paper a methodology that relies on meaningful text units (multiword terms) to represent queries. These multiword terms are used alternatively with interactive query expansion and automatic query expansion, the two are also combined in order to determine the combination that best boosts retrieval effectiveness. The experimentation has been carried out on two different document collections: a web collection consisting of the CSIRO domain and the Wikipedia corpus within TREC Enterprise track and INEX Ad-hoc track respectively. While the results obtained on the TrecEnt collection are not conclusive due perhaps to poor corpus quality and a change of evaluation measures in the TrecEnt campaigns, the results on the Wikipedia collection show that multiword term query representation and interactive query expansion are a promising combination for both standard document and focused retrieval. We have furthermore tested that the interactive query expansion process can be partially automated in the future by using existing term extraction and term variant identification programs which involve shallow NLP. We also plan to involve a much larger panel of users in order to evaluate the effect on the multiword term selection process.

References

1. Perez-Carballo, J., Strzalkowski, T.: Natural language information retrieval: progress report. Information Processing and Management **36**(1) (2000) 155 – 178
2. Smeaton, A.F.: Using nlp and nlp resources for information retrieval tasks. Kluwer Academic Publishers (1999) 99–109
3. Mishne, G., de Rijke, M.: Boosting web retrieval through query operations. Lecture Notes in Computer Sciences **3408** (2006) 502 – 516
4. Kageura, K.: The dynamics of Terminology: A descriptive theory of term formation and terminological growth. John Benjamins, Amsterdam (2002)
5. Ibekwe-SanJuan, F.: Constructing and maintaining knowledge organization tools: a symbolic approach. Journal of Documentation **62** (2006) 229–250
6. Metzler, D., Strohman, T., Turtle, H., Croft, W.B.: Indri at trec 2004: Terabyte track. (2005) electronic proceedings only
7. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst. **22**(2) (2004) 179–214
8. Bailey, P., de Vries, A.P., Craswell, N., Soboroff, I.: Overview of the trec 2007 enterprise track. In Voorhees, E.M., Buckland, L.P., eds.: TREC. Volume Special Publication 500-274., National Institute of Standards and Technology (NIST) (2007)
9. Lalmas, M., Tombros, A.: Evaluating xml retrieval effectiveness at inex. SIGIR Forum **41**(1) (2007) 40–57
10. L. Denoyer, P.G.: The wikipedia xml corpus. In: SIGIR Forum. (2006) 6
11. 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 (15-18th December 2008) 1–27