Exploring Sentence Level Query Expansion in Language Modeling Based Information Retrieval

Debasis Ganguly Johannes Leveling Gareth J. F. Jones
Centre for Next Generation Localisation (CNGL)
School of Computing
Dublin City University
Dublin 9, Ireland
{dganguly, jleveling, gjones}@computing.dcu.ie

Abstract

We introduce two novel methods for query expansion in information retrieval (IR). The basis of these methods is to add the most similar sentences extracted from pseudo-relevant documents to the original query. The first method adds a fixed number of sentences to the original query, the second a progressively decreasing number of sentences. We evaluate these methods on the English and Bengali test collections from the FIRE workshops. The major findings of this study are that: i) performance is similar for both English and Bengali; ii) employing a smaller context (similar sentences) yields a considerably higher mean average precision (MAP) compared to extracting terms from full documents (up to 5.9% improvemnent in MAP for English and 10.7% for Bengali compared to standard Blind Relevance Feedback (BRF); iii) using a variable number of sentences for query expansion performs better and shows less variance in the best MAP for different parameter settings; iv) query expansion based on sentences can improve performance even for topics with low initial retrieval precision where standard BRF fails.

1 Introduction

A major problem in information retrieval (IR) is the mismatch between query terms and terms in relevant documents in the collection which satisfy the user's information need. Query expansion (QE) is a popular technique used to bridge this vocabulary gap. Query expansion techniques work by adding terms to the user's original query so as to enrich it to fully describe the information need either by including alternative terms which might have been used in the relevant documents or which augment the terms in the original query. If good expansion terms are selected then the retrieval system can fetch additional relevant documents or increase the retrieved rank of items already retrieved. The query expansion techniques aim to predict the most suitable candidate words to be added to the query so as to increase retrieval effectiveness. The various different methods for IR have corresponding different approaches to QE. In this paper we concentrate our investigation on the language modeling (LM) IR framework as proposed by (Hiemstra, 2000). The standard feedback techniques for IR assume that the top R documents as a whole are relevant, and do not take into consideration the fact that in some cases the documents as a whole might not be relevant to the query, but a particular subtopic of it may be highly relevant. The new method of QE introduced in this paper proposes that sorting sentences contained in relevant or pseudo relevant documents based on their similarities to the query, and then choosing a subset from this sorted set to add to the original query can help to prevent the addition of non relevant terms to the query. This approach to expansion by selecting sentences in this way introduces more context to the query as opposed to term based expansion.

The remainder of the paper is organized as follows: Section 2 overviews existing releated work, Section 3 presents two methods of sentence level query expansion one which adds a fixed number of sentences from each pseudo relevant document to the query and the other which adds a variable number of sentences, Section 4 describes our experi-

Proceedings of ICON-2010: 8th International Conference on Natural Language Processing,
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2010

mental setup and give experimental results, Section 5 contains a detailed analysis of the results, and finally Section 6 concludes the paper with directions for future work.

2 Related Work

2.1 Relevance Feedback

Query expansion (QE) is one output of the process of relevance feedback (RF). In standard RF the user indicates which of a number of documents from an initial retrieval run are relevant to their information need. Based on the assumption that further relevant documents will be similar to those identified so far, RF adjusts the query to increase the likelihood of retrieval of such documents. A key part of this process is QE where the query is expanded to incorporate terms appearing in the known relevant documents, the other principal component of RF is term reweighting which emphasizes the importance of terms found in relevant documents. As outlined in Section 1, blind RF (BRF) (or pseudo RF) assumes that the top ranked retrieved documents are relevant to the information need. BRF has been used in many IR studies over the years using many RF techniques. All the existing query expansion approaches are term based i.e. a subset of terms occuring in relevant documents are chosen based on some scoring function aiming to select the good expansion terms. The simplest scoring function which works well in practise uses term occurence statistics alone as advocated by Salton and Buckley (1994) where terms occuring in the largest number of relevant documents are added to the query. The score assigned to a term t is in this approach is shown in Equation 1 where r is the number of relevant documents that the term occurs in.

$$Occ(t) = r \tag{1}$$

Such a simple scoring function does not distinguish terms by their collection statistics and might end up adding too many common terms, thus not increasing IR effectiveness significantly, only because these terms are also abundant in the relevant documents. Scoring functions thus are augmented by incorporating the *idf* factor. The Robertson Selection Value (RSV) (Robertson et al., 1995) is one such term scoring formula defined as follows.

$$RSV(t) = r\ \log \frac{(r+.5)(N-R-n+r+.5)}{(n-r+.5)(R-r+.5)} \tag{2}$$

In Equation 2, r is the number of relevant documents that the term t occurs in, N is the total number of documents in the collection, n is the document frequency of the term and R is the number of relevant documents. A term selection score for LM as shown in Equation 3 was proposed in (Ponte and Croft, 1998).

$$L(t) = \sum_{d \in R} \log \frac{P(t|M_d)}{\frac{cf_t}{cs}} \tag{3}$$

In Equation 3, M_d denotes the query generation model from document d, cf_t denotes the collection frequency of term t and cs the collection size.

While BRF provides improvement in IR performance in a subsequent retrieval run for the query, it is not perfect due to the fact that some of the content assumed to be relevant in the BRF is in fact not relevant, thus leading to reduced IR effectiveness. Buckley's work (1994) which performed massive query expansion using the Vector Space Model of the SMART[1] retrieval system for ad-hoc retrieval experiments at TREC 3 involves employing Rocchio feedback (Rocchio, 1971) with 300 to 530 terms and phrases for each topic. An improvement in retrieval effectiveness between 7% and 25% in various experiments was observed. The problem with massive query expansion with VSM is that it favours retrieval of longer documents in the feedback step and thus if most of the relevant documents are of shorter length, retrieval effectiveness may be reduced after feedback. Instead, we investigate a sentence level expansion method (also a massive query expansion) using LM IR. Typically in LM, the query is represented as a *sequence* of independent words thus leading to a multinomial view of model M_d:

$$P(q|M_d) = \prod_{t \in q} P(t|M_d)^{\alpha(t)} \tag{4}$$

In Equation 4, $P(q|M_d)$ denotes the probability of generating query q from document d, $P(t|M_d)$ denotes the probability of generating the term t from d and $\alpha(t)$ denotes the number of times a term t appears in the query. Adding sentences from a document d to the original query q makes $P(q|d)$ more likely because the frequency of each term added is accurately reproduced. For example if we are adding a sentence "A B A A B" from d to q, we are adding "A" thrice and "B" twice thus adding $P(t = \text{``}A\text{''}|M_d)^3 + P(t = \text{``}B\text{''}|M_d)^2$

[1] ftp://ftp.cs.cornell.edu/pub/smart/

to $P(q|M_d)$ whereas if we add only the terms "A" and "B", the increment in $P(q|M_d)$ would be $P(t = \text{"}A\text{"}|M_d) + P(t = \text{"}B\text{"}|M_d)$. Simply speaking, adding sentences from a document to the query makes the query *look* more like the document and hence increases the likelihood of generating the expanded query from the given document.

Some of the existing works on using a smaller content instead of the whole document includes the work of Lam-Adesina and Jones (2001) which select QE terms from the summaries of a document, and Vechtomova and Karamuftuoglu (2007) which experiments with using smaller, lexically cohesive text units for BRF.

In seeking to address problems in BRF a range of questions can be explored, e.g. whether to use massive versus selective feedback, how to best obtain a ranking of feedback terms in order to select the most appropriate ones, or how to dynamically adapt feedback parameters. This paper focuses on exploring a novel method for BRF taking the standard Robertson term selection approach (1990) as our baseline.

2.2 FIRE Test Collections

FIRE is an evaluation forum for IR on Indian languages and till now it has been held twice in 2008 and 2010. The FIRE test collection for ad hoc IR contains newspaper articles on various topics including sports, politics, business, and local news (Majumdar et al., 2008). The articles are represented as structured XML documents in TREC format, using UTF-8 encoding. FIRE topics resemble topics from other retrieval campaigns such as TREC in format and content. They comprise a brief phrase describing the information need (topic title, T), a longer description (topic description, D), and a part with information on how documents are to be assessed for relevance (topic narrative, N). Retrieval queries are typically generated from the title and description fields of topics (TD). The official test topics for 2008 comprises of topics 26-75 whereas topics 76-125 had been used for 2010. The relevance of documents were assessed by pooling submissions from systems participating in the FIRE retrieval tracks.

3 Sentence Level Query Expansion

3.1 Motivation

We present two scenarios where sentence level query expansion can potentially outperform term based expansion.

Context Terms The deduction of the importance of the query terms by occurrence statistics only, fails to capture any context information. The idea is that in sentence expansion the constituents terms of the sentence can be useful to enrich the query with sufficient context information which is typically missed in term based expansion because all the constituent words might not be among the most frequent words occurring in the pseudo relevant set of documents. This is particularly a problem for languages rich in word compounding. Although word compounding is rare in English, for simplicity we cite an English example here. The word *farmland* can be split into two words *farm* and *land* to convey the same meaning. Often *farmland* can appear abbreviated to "land"after it has been introduced to convey the same meaning for subsequent appearance. So, in a retrieved list of say top R documents, the word *land* might be among the top most frequent words whereas the word *farm* might be not. But in this scenario, we would have liked to add both the terms *farm* and *land* to the original query if it contains the term *farmland*. A way to achieve this is sentence level expansion, because sentences offer an implicit way of capturing the context associated with a term. Returning to our example, the initial part might see an introduction to the term *farmland* which is followed by some discussions about *farming* for the rest of the sentence. The similarity of this sentence to the query sentence containing the term *farmland* would be high, and we would add this sentence to the query. But this also results in adding the term *farm* to the original query which we would have missed in conventional term based expansion.

Subtopic Relevance A common scenario in retrieval is that not all the documents from the top r are relevant to the query. But these documents might not be fully irrelevant to the query in the sense that a subtopic of these documents might actually be highly relevant to the query (Wilkinson, 1994). Hence term occurence statistics might introduce a huge query drift by adding highly frequent terms from the non relevant subsections of the pseudo relevant documents. Sentence level expansion can potentially be more effective in such scenarios because it only adds the sentences having maximum term overlap with the query sentences thus in effect defining a safe zone to choose the expansion terms from. This background motivation paves the path for a formal description of sentence expansion in the following section.

3.2 A formal description

Let r be the number of top ranked documents assumed to be relevant for a query. Each pseudo-relevant document d can be represented as a set comprising of the constituent sentences. Thus $d = \{d^1, \ldots d^{\eta(d)}\}$ where $\eta(d)$ is a function denoting the number of sentences of the document d and d^is are its constituent sentences. Each such sentence d^i is represented as a vector $d^i = (d^i_1, \ldots d^i_{\zeta(d)})$ where $\zeta(d)$ is the number of unique terms in d. The components of the vector are defined as follows.

$$d^i_j = \quad \textit{tf}(t_j, d^i)\phi(t_j) \quad \forall j \in [1, \zeta(d)] \qquad (5)$$

In Equation 5, ϕ is a weighting function which assigns a weight based on the Part of Speech (PoS) tag of term t_j and $\textit{tf}(t_j, d^i)$ denotes the term frequency of term t_j in sentence d^i. The weights for different word categories denoted by the function ϕ is defined as shown in Equation 6.

$$\phi(t_j) = \begin{cases} 1.0\text{:} & \text{if } t_j \text{ is a proper noun} \\ 0.8\text{:} & \text{if } t_j \text{ is a common noun} \\ 0.3\text{:} & \text{if } t_j \text{ is a verb} \\ 0.2\text{:} & \text{if } t_j \text{ is an adjective or adverb} \\ 0.1\text{:} & \text{otherwise} \end{cases} \qquad (6)$$

This way of defining the vectors for every sentence ensures that an overlap in the proper nouns between two sentences is given more importance than an overlap between the common nouns and so on. Previous research suggests that nouns are more informative than other types of terms and are better features for query expansion (Jing and Croft, 1994). Our initial experiments showed that for English, this weighted term overlap gives better results as compared to its unweighted counterpart. Hence in this paper for English, we employ PoS weighted term weighting for all the sentence level QE experiments. For the Bengali experiments, ϕ is a constant function set to 1.

Using the above notations we propose two variants of sentence level expansion algorithms. In the first variant, we fix the number of sentences to be added to the original query whereas in the second variant the number of sentences chosen from the pseudo relevant documents is proportional to the retrieval score of the document i.e. for top ranked documents we add more sentences whereas for bottom ranked ones we add fewer. The second variant ensures that as we go down through the ranked list we progressively become more selective in adding the sentences.

Algorithm 1 SentenceExpansion(q, m, vns).

```
1:  $q$ : The original query
2:  $m$ : Number of sentences
3:  $vns$ : Whether to use variable number of sentences
4:  $Q \leftarrow q$
5:  {For each pseudo-relevant document}
6:  for $i = 1$ to $r$ do
7:      $d \leftarrow i^{th}$ pseudo relevant document
8:      {For each sentence in this query}
9:      for $j = 1$ to $\eta(q)$ do
10:         $S \leftarrow \emptyset$
11:         {For each sentence in this document}
12:         for $k = 1$ to $\eta(d)$ do
13:             {Store the similarities of the $j^{th}$ query sentence
                vector with the $k^{th}$ document sentence vector}
14:             $s.sentence \leftarrow d^k$
15:             $s.sim \leftarrow d^k \cdot q^j$
16:             $S \leftarrow S \cup s$
17:         end for
18:         Reorder the set $S$ such that
19:         $S_\alpha.sim \geq S_\beta.sim \quad \forall \alpha < \beta$
20:         if $vns = false$ then
21:             $m_i \leftarrow m$
22:         else
23:             $m_i = \lfloor \frac{1-m}{r-1}(i-1) + m \rfloor$
24:         end if
25:         {Select the most similar $m_i$ sentences}
26:         for $k = 1$ to $m_i$ do
27:             $Q \leftarrow Q \cup S_k.sentence$
28:         end for
29:     end for
30: end for
31: return $Q$
```

Algorithm 1 provides the general functionality of sentence expansion. In the outer loop of line 6 of Algorithm 1 we iterate over the top r pseudo relevant documents. The inner loop of line 9 is used to iterate over every query sentence. The inner loop of line 12 iterates over the sentences of this document and adds a tuple s comprising of the sentence text itself and its similarity with the current query sentence in lines 14 and 15 to the set of sentences S. The similarity is measured by computing the inner product of the two sentence vectors the reason being that the cosine measure favours short text (Wilkinson et al., 1995). This set

S is then sorted in decreasing order of the stored similarities. Finally in lines 26-28, we select the top m_i tuples from the set S and add the sentence text to the original query. The value of m_i is constant if the parameter vns is set to false as can be seen in line 21 whereas m_i is obtained by a linear interpolation as shown in line 23 if vns is true. The slope of the interpolating line is uniquely determined from the fact that we use m number of sentences for the top ranked document and 1 sentence for the bottom ranked one.

The two variants of sentence expansion are realized by calling Algorithm 1 with the parameter vns set to $false$ and $true$ respectively.

4 Experimental results

4.1 Description and Setting

For our IR experiments, we used the two FIRE topic sets for monolingual retrievals in English and Bengali. For English, the terms have been stemmed with the Porter stemmer (Porter, 1980). A rule based stemmer was used to stem Bengali words (Leveling et al., 2010).

Sentence boundaries are detected using the Morphadorner package[2], which also implements PoS tagging for English using a trigram Hidden Markov Model.

SMART stopword list was used to remove the stopwords for English. Bengali common words were removed using the list of stopwords provided by the FIRE organizers [3]. All the IR experiments were done with TD queries only. We used the LM retrieval model implemented in SMART by one of the authors for all the experiments described in this section. The LM retrieval model aims to fetch the documents with maximum likelihood of generating the given query. The fundamental LM equation of generating q from a document d is given in Equation 7.

$$P(q|d) = P(d)\prod_{i=1}^{n} \lambda_i P(t_i|d) + (1-\lambda_i)P(t_i) \quad (7)$$

In Equation 7, λ_i is the probability of choosing the i^{th} query term from the document d whereas $(1-\lambda_i)$ is the probability of choosing the term from the collection . In all our experiments, we use $\lambda_i = 0.3$ for both the original and the expanded query terms.

Table 1: Best MAPs obtained by the three term selection methods for term based QE.

Topic set	BRF$_{Occ}$	BRF$_{RSV}$	BRF$_L$
26-75	**0.5682**	0.5614	0.5576
76-125	**0.4953**	0.4881	0.4767

4.2 Choosing a baseline

We used the retrieval performance without QE as a lower baseline (initial retrieval baseline). In order to choose a stronger baseline, we performed experiments using the three approaches to term selection (Occ, RSV, and L) as outlined in Section 2.1. The parameters to vary are the number of documents assumed to be pseudo relevant, hereafter refered to as D and the number of terms to be added to the original query denoted by T. We varied D and T in the range $[10, 40]$ in steps of 10 for the three approaches and performed retrieval experiments on FIRE 2008 and 2010 English topics. The best results are reported in Table 1. We observe from Table 1 that BRF$_{Occ}$ performs best for English and hence we select this approach for our subsequent experiments on term expansion. Hereafter we loosely refer to BRF$_{Occ}$ as BRF.

4.3 Sentence expansion results

Similar to term expansion, one of the parameters to vary is D. The other parameter to vary is m which is the number of sentences to add. For the algorithm BRF$_{cns}$ we vary m in the range $[2, 5]$ in steps of 1. For the algorithm BRF$_{vns}$ we vary m in the range $[4, 10]$ in steps of 2. This is done to ensure that we add the same number of sentences on an average for both the sentence expansion approaches. The IR runs with conventional term based expansion as against sentence expansion are reported in Tables 2-5. Statistically significant improvements in MAP, employing the Wilcoxon measure with confidence measure set to 95%, of the feedback retrievals over the initial retrievals are indicated by an adjacent $^\diamond$, whereas the significant improvements of the sentence level QE experiments with respect to both the initial retrievals and the best term based BRF experiments are indicated by *.

Both the approaches to sentence level QE show improvements in the best MAP obtained over a range of experiments for Bengali topic set 26-75 (see Table 2) and for the two English topic sets (see Tables 4 and 5). The algorithm BRF$_{cns}$ does

[2] http://morphadorner.northwestern.edu/morphadorner/

[3] http://www.isical.ac.in/~fire/

Table 2: Comparison of term expansion with two variants of sentence expansion on Bengali topic set 26-75 (Official test topics for FIRE 2008).

BRF			BRF$_{cns}$			BRF$_{vns}$		
D	T	MAP	D	m	MAP	D	m	MAP
10	10	0.3695$^\diamond$	10	2	0.3951$^\diamond$	10	4	0.3939$^\diamond$
10	20	0.3800$^\diamond$	10	3	0.4073$^\diamond$	10	6	0.3965$^\diamond$
10	30	0.3763$^\diamond$	10	4	0.4092*	10	8	0.4009$^\diamond$
10	40	0.3851$^\diamond$	10	5	0.4133*	10	10	0.4070*
20	10	0.3602$^\diamond$	20	2	0.4075*	20	4	0.4106*
20	20	0.3828$^\diamond$	20	3	0.4193*	20	6	0.4144*
20	30	0.3845$^\diamond$	20	4	**0.4226***	20	8	0.4222*
20	40	**0.3885**$^\diamond$	20	5	0.4223*	20	10	0.4251*
30	10	0.3372	30	2	0.4057$^\diamond$	30	4	0.4236*
30	20	0.3558	30	3	0.4177$^\diamond$	30	6	0.4168*
30	30	0.3843$^\diamond$	30	4	0.4213*	30	8	0.4257*
30	40	0.3812$^\diamond$	30	5	0.4216*	30	10	**0.4302***
40	10	0.3329	40	2	0.3932$^\diamond$	40	4	0.4160*
40	20	0.3619$^\diamond$	40	3	0.4099*	40	6	0.4189*
40	30	0.3497$^\diamond$	40	4	0.4082*	40	8	0.4274*
40	40	0.3641$^\diamond$	40	5	0.4101*	40	10	0.4300*
Initial retrieval baseline:								0.3084
Best official TD run (Dolamic and Savoy, 2008):								0.4134

Table 3: Comparison of term expansion with two variants of sentence expansion on Bengali topic set 76-125 (Official test topics for FIRE 2010).

BRF			BRF$_{cns}$			BRF$_{vns}$		
D	T	MAP	D	m	MAP	D	m	MAP
10	10	0.4339	10	2	0.4174	10	4	0.4388
10	20	0.4486	10	3	0.4400	10	6	0.4531
10	30	**0.4537**	10	4	0.4367	10	8	**0.4581**
10	40	0.4424	10	5	**0.4467**	10	10	0.4571
20	10	0.4250	20	2	0.4122	20	4	0.4365
20	20	0.4121	20	3	0.4021	20	6	0.4461$^\diamond$
20	30	0.4073	20	4	0.4144	20	8	0.4378$^\diamond$
20	40	0.4141	20	5	0.4198	20	10	0.4427$^\diamond$
30	10	0.4205	30	2	0.4231	30	4	0.4174
30	20	0.4160	30	3	0.4272	30	6	0.4380
30	30	0.3998	30	4	0.4355	30	8	0.4453
30	40	0.4065	30	5	0.4405	30	10	0.4482$^\diamond$
40	10	0.4070	40	2	0.4164	40	4	0.4274
40	20	0.4085	40	3	0.4172	40	6	0.4447
40	30	0.4001	40	4	0.4273	40	8	0.4404
40	40	0.3825	40	5	0.4352	40	10	0.4441$^\diamond$
Initial retrieval baseline:								0.4272
Best official TD run (Leveling et al., 2010):								0.4944

Table 4: Comparison of term expansion with two variants of sentence expansion on English topic set 26-75 (Official test topics for FIRE 2008).

BRF			BRF$_{cns}$			BRF$_{vns}$		
D	T	MAP	D	m	MAP	D	m	MAP
10	10	**0.5682**$^\diamond$	10	2	0.5778*	10	4	0.5725$^\diamond$
10	20	0.5609$^\diamond$	10	3	0.5863*	10	6	0.5809*
10	30	0.5490$^\diamond$	10	4	0.5904*	10	8	0.5823*
10	40	0.5442	10	5	0.5887*	10	10	0.5838*
20	10	0.5527$^\diamond$	20	2	0.5835*	20	4	0.5853*
20	20	0.5616$^\diamond$	20	3	0.5901*	20	6	0.5960*
20	30	0.5558$^\diamond$	20	4	0.5956*	20	8	0.5937*
20	40	0.5507$^\diamond$	20	5	0.5951*	20	10	0.5943*
30	10	0.5411$^\diamond$	30	2	0.5824*	30	4	0.5923*
30	20	0.5621$^\diamond$	30	3	0.5914*	30	6	0.5981*
30	30	0.5380	30	4	0.5962*	30	8	0.6007*
30	40	0.5364	30	5	**0.5964***	30	10	**0.6015***
40	10	0.5362$^\diamond$	40	2	0.5776*	40	4	0.5949*
40	20	0.5554$^\diamond$	40	3	0.5858*	40	6	0.6003*
40	30	0.5355	40	4	0.5863*	40	8	0.6007*
40	40	0.5383	40	5	0.5894*	40	10	0.6004*
Initial retrieval baseline:								0.5084
Best official TD run (Udupa et al., 2008):								0.5572

Table 5: Comparison of term expansion with two variants of sentence expansion on English topic set 76-125 (Official test topics for FIRE 2010).

BRF			BRF$_{cns}$			BRF$_{vns}$		
D	T	MAP	D	m	MAP	D	m	MAP
10	10	0.4694	10	2	0.4735	10	4	0.5016$^\diamond$
10	20	0.4780$^\diamond$	10	3	0.4848	10	6	0.5032$^\diamond$
10	30	**0.4953**$^\diamond$	10	4	0.5016	10	8	0.5034$^\diamond$
10	40	0.4891	10	5	0.4987	10	10	0.5063$^\diamond$
20	10	0.4515	20	2	0.4602	20	4	0.4948
20	20	0.4645$^\diamond$	20	3	0.4760	20	6	0.5024
20	30	0.4799	20	4	**0.5032**	20	8	**0.5102**
20	40	0.4790	20	5	0.4911	20	10	0.5057
30	10	0.4431	30	2	0.4551	30	4	0.4899
30	20	0.4453	30	3	0.4659	30	6	0.4939
30	30	0.4580	30	4	0.4779	30	8	0.5001
30	40	0.4588	30	5	0.4785	30	10	0.4958
40	10	0.4428	40	2	0.4513	40	4	0.4867
40	20	0.4452	40	3	0.4595	40	6	0.4944
40	30	0.4541	40	4	0.4763	40	8	0.4881
40	40	0.4373	40	5	0.4664	40	10	0.4872
Initial retrieval baseline:								0.4744
Best official TD run (Leveling et al., 2010):								0.4846

not outperform the best MAP of conventional term based QE for the topic set 76-125 for Bengali. But the algorithm BRF$_{vns}$, where we use a variable number of sentences, does outperform the term based QE. The improvements are statistically significant for the first topic sets in both the languages. Although we do not notice any statistically significant improvements of the sentence level QE methods as compared to BRF in Tables 3 and 5, we do observe that we get a higher number of statistically significant improvements with respect to the initial retrieval using the method BRF$_{vns}$, i.e. 5 vs. 0 cases as seen in Table 3 and 4 vs. 3 cases as seen in Table 5.

5 Analysis

It has been found that traditional QE techniques degrade performance for many topics (Billerbeck and Zobel, 2004). If most of the top ranked pseudo-relevant documents are actually irrelevant to the query, then QE can add a lot more unimportant terms which drifts the query vector further away from the centroid of the relevant documents and as a result the feedback retrieval can result in a worse precision.

We explore the relative query drifts caused by the term based and the sentence based query expansion approaches for Bengali topics. To study how query drift is affected by initial precision, we partition the 100 queries from both the topic sets into categories, hereafter refered to as bins, corresponding to the number of documents which are relevant in top 20 of the ranked list obtained by the initial retrieval. Thus the first bin contains the topics which retrieve no relevant document in the initial retrieval, the second bin consists of topics which fetch only one relevant document in the top 20 and so on. For each bin, the Mean Average Precision is computed by considering only the queries of the current bin. A similar analysis has been presented in (Mitra et al., 1998). In Table 6 we report the number of queries for which average precision decreases as compared to initial retrieval and in Figure 1 we report the percentage changes in the average precision as compared to the initial retrieval for the three expansion techniques.

Figure 1 reveals that the sentence level QE techniques work particularly well for queries with low P@20 for initial retrieval. The reason can be attributed to subtopic relevance i.e. even if the top ranked documents are not fully relevant to the query, some parts of these documents are highly relevant and the sentence level QE exploits this fact by adding only the most similar sentences from these documents to the original query. But term based expansion adds many unimportant terms to the query based on occurence statistics over the whole document and as a result suffers from heavy query drift in the low P@20 topic categories. Table 6 also shows that the total number of queries for which the average precision decreases with respect to the initial retrieval, is less for the two sentence level approaches.

Table 6: Query drift caused by expansion over the combined FIRE 2008 and FIRE 2010 topics in Bengali.

		# queries hurt		
Bin #	# Queries	BRF	BRF$_{cns}$	BRF$_{vns}$
0	5	2	1	1
1	2	1	1	1
2	12	3	3	2
3	7	0	1	2
4	16	8	4	3
5	9	5	4	2
6	8	4	2	2
7	9	6	3	2
8	9	4	2	4
9	4	1	0	0
10	2	0	0	0
11	3	0	0	1
12	6	1	1	1
13	4	3	4	3
16	2	0	0	0
17	1	0	0	0
19	1	1	0	0
Total	100	39	26	24

It is also observed that the algorithm BRF$_{vns}$ performs better that BRF$_{cns}$. The reason can be attributed to the fact that BRF$_{cns}$ suffers from the tendency of adding some non important terms from sentences belonging to documents which are lower ranked in the initial retrieval list. But BRF$_{vns}$ is more selective in adding sentences in the sense that it adds fewer sentences from documents ranked lower in the initial retrieval list and hence reduces the chance of adding non important query terms to the original query.

The term expansion based approach suffers from degradation of retrieval effectiveness even for a query for which 19 retrieved documents in top 20 were actually relevant (refer to the righmost bar of the term expansion histogram of Figure 1). Again this result suggests that occurence statistics of terms at the whole document level

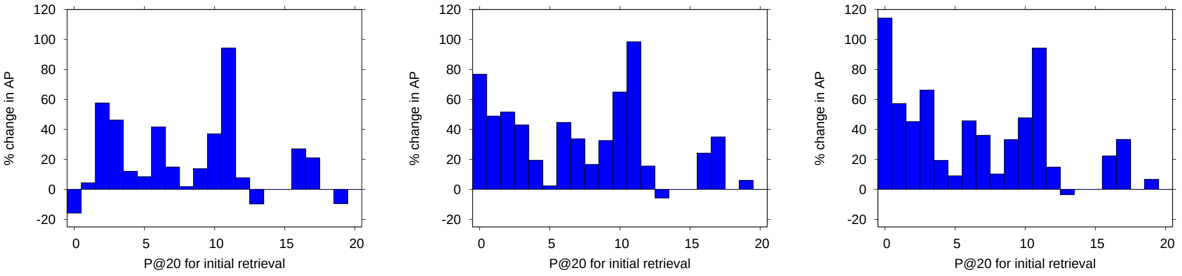

Figure 1: Query drift over the combined topic sets in Bengali for term-based QE, expansion with a constant and a variable number of sentences (from left to right).

might not always be a good estimator of the importance of a query term in improving precision. Both sentence expansion approaches work better for this query showing that subtopical relevance works better even for topics for which higher precision is achieved during the initial retrieval step.

It can also be observed that the best MAP that can be obtained with a particular number of documents used for blind feedback is higher for the sentence level expansion. Figure 2 shows the best MAPs using the various number number of documents for pseudo relevance on the first topic set for both the languages. The figures show that with increasing number of documents assumed as pseudo relevant, the performance of the term based QE falls off steeply whereas it can be seen that there is no drastical degradation of performance in the sentence level expansions. Thus the two sentence expansion algorithms are relatively less sensitive to the parameter changes.

For both the languages it is observed that the improvement in MAP for the sentence level QE approaches is not statistically significant for the second topic sets. The reason is that for the FIRE 2010 topic sets the initial retrieval precsion is better and the benefit of feedback is less for topics where the initial retrieval precision is high (Lynam et al., 2004). Though the improvements in the best MAPs are not statistically significant for FIRE 2010 topics, Table 3 shows that for the term based expansion, 12 cases give a MAP value lower than the baseline retrieval, for BRF$_{cns}$ 8 cases yield a lower MAP and for BRF$_{vns}$ we have only one such case. Similar observations can be made from Table 5 which shows that there are 11 worse MAPs for BRF, 7 for BRF$_{cns}$ and none for BRF$_{vns}$. Thus we see that although term based expansion works well for the scenario where the initial retrieval precision is high but the method is very sensitive to parameter changes and can lead to worse retrieval effectiveness as compared to the initial retrieval quite often.

To verify the hypothesis that sentence expansion adds more context to the original query as described in Section 3.1, we manually checked the first two FIRE 2008 English topics (topic ids 26 and 27). Topic 26 requires finding documents on Singur land dispute. The BRF approach adds the word *farm* but not the words like *farmland*, *Trinamool* or *Nandigram* which are good expansion terms candidates as judged by one of the authors who is familiar with the news. Topic 27 asks for documents on relationship between India and China. Term expansion adds the word *Delhi* but fails to add words like *Bangalore* and *Beijing*. We would like to emphasise on the fact that even if the expanded query contains the word Bangalore but the query does not drift away completely towards documents only on Bangalore because of frequent occurrences of the words India and China, which are added many times. Thus the query still is about India and China but at least gives a chance for documents about IT relationships between these two countries to be retrieved because of the presence of the word Bangalore in such documents.

To see if sentence level QE is indeed able to add the *important* query terms to the original query we run a series of true relevance feedback (TRF) experiments which involves selecting terms only from the top r actually relevant documents of the initial ranked list. We do the TRF experiments for both topic sets on English and Bengali. The number of terms used for query expansion using the BRF methods is set identical to the number of terms which yields the best MAP as observed from Tables 2, 3, 4 and 5.

These TRF experiments can be considered as the strongest possible baseline and cardinality of

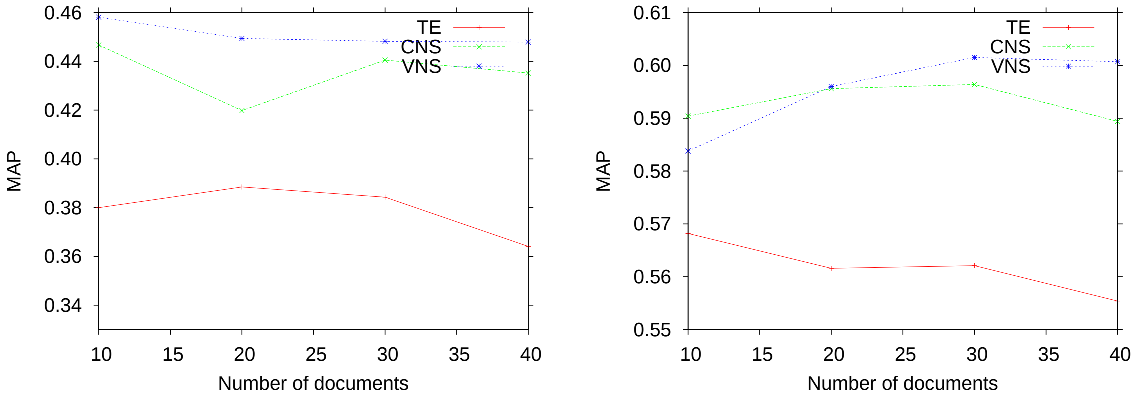

Figure 2: Best MAPs plotted against the number of documents used for pseudo relevance on topics 26-75 for Bengali and English (left to right).

intersection of the set of terms obtained by a BRF appraoch with the set of terms obtained by the TRF method indicates the effectiveness of the blind relevance approach. In Table 7 we report the intersection of the set of terms obtained by the best performing baseline BRF and BRF$_{vns}$ approaches with the TRF approach. T_X denotes the set of terms obtained by method X where X is either BRF (standard term based expansion) or BRF_{vns} (sentence level query expansion using variable number of sentences). We observe that the sentence level QE is able to add more number of *important* query terms as compared to standard term based QE.

6 Conclusions and Future Work

The main contribution of the paper is the proposal of a novel method of query expansion by adding sentences in contrast to the traditional approach of adding terms. Our experiments on sentence level QE show that it can significantly increase MAP compared to conventional term based QE. The results show that the variable number of sentences variant of the algorithm is better than the constant number of sentences variant. More work can be pursued in this direction regarding the type of the interpolation function used to obtain the value of the number of sentences to be added for the intermediate documents between the top and the bottom ranked ones. The linear function of line 23 of Algorithm 1 can be generalised to higher degree polynomials. We would also run the same set of experiments on other standard collections and topic sets e.g. the TREC and the INEX adhoc tasks.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project.

References

Bodo Billerbeck and Justin Zobel. 2004. Questioning query expansion: An examination of behaviour and parameters. In Klaus-Dieter Schewe and Hugh E. Williams, editors, *Proceedings of the Fifteenth Australasian Database Conference (ADC 2004)*, volume 27, pages 69–76, Dunedin, New Zealand. Australian Computer Society, Inc.

Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. 1994. Automatic query expansion using SMART: TREC 3. In Donna K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 69–80, MD, USA. National Institute for Standards and Technology (NIST).

Ljiljana Dolamic and Jacques Savoy. 2008. UniNE at FIRE 2008: Hindi, Bengali, and Marathi IR. In *Working Notes of the Forum for Information Retrieval Evaluation, December 12–14, 2008*, Kolkata, India.

Djoerd Hiemstra. 2000. *Using Language Models for Information Retrieval*. Ph.D. thesis, Center of Telematics and Information Technology, AE Enschede, The Netherlands.

Yufeng Jing and W. Bruce Croft. 1994. An association thesaurus for information retrieval. In *RIAO'1994*, pages 146–160.

Adenike M. Lam-Adesina and Gareth J. F. Jones. 2001. Applying summarization techniques for term selection in relevance feedback. In W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel,

Table 7: Intersection of the hyothesized sets of BRF terms with the reference set of TRF terms.

Language	Topic set	TRF		BRF		BRF$_{vns}$	
		MAP	$\|T_{TRF}\|$	MAP	$\|T_{TRF} \cap T_{BRF}\|$	MAP	$\|T_{TRF} \cap T_{BRF_{vns}}\|$
Bengali	26-75	0.5554	979	0.3885	744 (75.99%)	0.4302	955 (97.54%)
Bengali	76-125	0.7510	991	0.4537	728 (73.46%)	0.4581	933 (94.14%)
English	26-75	0.7322	937	0.5682	743 (79.29%)	0.6015	912 (97.33%)
English	76-125	0.6083	433	0.4953	407 (93.99%)	0.5102	432 (99.76%)

editors, *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 1–9. ACM.

Johannes Leveling, Debasis Ganguly, and Gareth J. F. Jones. 2010. DCU@FIRE2010: Term conflation, blind relevance feedback, and cross-language IR with manual and automatic query translation. In *Second Workshop of the Forum for Information Retrieval Evaluation (FIRE 2010), Working Notes*, pages 39–44.

Thomas R. Lynam, Chris Buckley, Charles L. A. Clarke, and Gordon V. Cormack. 2004. A multi-system analysis of document and term selection for blind feedback. In *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, November 8–13, 2004*, pages 261–269, Washington, DC, USA. ACM.

Prasenjit Majumdar, Mandar Mitra, Dipasree Pal, Ayan Bandyopadhyay, Samaresh Maiti, Sukanya Mitra, Aparajita Sen, and Sukomal Pal. 2008. Text collections for FIRE. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, pages 699–700. ACM.

Mandar Mitra, Amit Singhal, and Chris Buckley. 1998. Improving automatic query expansion. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–214, New York, NY, USA. ACM.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA. ACM.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In Donna K. Harman, editor, *Overview of the Third Text Retrieval Conference (TREC-3)*, pages 109–126, Gaithersburg, MD, USA. National Institute of Standards and Technology (NIST).

Stephen E. Robertson. 1990. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364.

Joseph J. Rocchio. 1971. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART retrieval system – Experiments in automatic document processing*. Prentice Hall, Englewood Cliffs, NJ, USA.

Raghavendra Udupa, Jagadeesh Jagarlamudi, and K. Saravanan. 2008. Microsoft Research at FIRE2008: Hindi-English cross-language information retrieval. In *Working Notes of the Forum for Information Retrieval Evaluation, December 12–14, 2008*, Kolkata, India.

Olga Vechomova and Murat Karamuftuoglu. 2007. Query expansion with terms selected using lexical cohesion analysis of documents. *Information Processing and Management*, 43(4):849–865.

Ross Wilkinson, Justin Zobel, and Ron Sacks-Davis. 1995. Similarity measures for short queries. In *In Fourth Text REtrieval Conference (TREC-4)*, pages 277–285.

Ross Wilkinson. 1994. Effective retrieval of structured documents. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 311–317, New York, NY, USA. Springer.