Applying Query Formulation and Fusion Techniques For Cross Language News Story Search

Piyush Arora, Jennifer Foster and Gareth J. F. Jones
CNGL Centre for Global Intelligent Content
School of Computing, Dublin City University
Glasnevin, Dublin 9, Ireland
{parora, jfoster, gjones}@computing.dcu.ie

ABSTRACT

Cross Language News story search (CLNSS) is concerned with finding documents describing the same events in documents in different languages. As well as supporting information retrieval (IR), CLNSS has other applications in mining parallel and comparable data across different languages. In this paper, we present an overview of the work carried out for our participation in the Cross Language !ndian News Story Search (CL!NSS) task at FIRE 2013. In the CL!NSS task we explored the problem of cross language news search for the English-Hindi language pair. English news stories are used as queries to seek similar news documents from Hindi news articles. Hindi being a resource-scarce language offers many challenges towards retrieving relevant news articles. We investigate and contrast translation of input queries from English to Hindi using the Google and Bing translation services. To support translation of out-of-vocabulary words we use the Google transliteration service. A key challenge of the CL!NSS task is formation of search queries from the English news articles, since they are much longer than the much shorter queries typically used in IR applications. To address this problem, we explore the use of summarization to extract a query from the input news documents, and use these summarized queries as the input to the cross language IR system. We explore the use of query expansion using pseudo relevance feedback (PRF) in the IR process, since this has been shown to be effective for cross language IR in many previous investigations. We also explore in detail the use of data fusion techniques over different sets of retrieved results obtained using diverse query formulation techniques. For the CL!NSS task our team submitted 3 main runs. The results of our best run was ranked first among official submissions based on NDCG@5 and NDCG@10 values and second for NDCG@1 values. For the 25 test queries the results of our best main run were NDCG@1 0.7400, NDCG@5 0.6809 and NDCG@10 0.7268. We present our methodology, official results and results of a number of post-task experiments that were conducted to further examine the cross language search problem. Our experiments reveal that query formu-

lation plays a vital role in improving search results for news documents across different languages. Instead of using the complete news documents the summarized queries show better performance. Data fusion techniques also help to improve the performance of the system by boosting the rank of documents, thus improving the NDCG scores.

Keywords
Hindi Information Retrieval, Cross Language News Search, Query Translation, Query Summarization, Data Fusion, Pseudo Relevance Feedback

1. INTRODUCTION
Cross Language Information Retrieval (CLIR) has been a significant topic of research in information retrieval (IR) for many years. Over the years, CLIR has also become a popular technique for mining parallel and comparable data across different languages to support other natural language related tasks [5]. News documents pertaining to similar temporal and spatial information which talk about an event/activity/person are a potentially good source for extracting comparable corpora.

The paper describes details of our participation in the Cross Language !ndian News Story Search (CL!NSS) task at FIRE 2013 [7]. The CL!NSS task is an edition of the PAN@FIRE task [8] which focuses on addressing news story linking between English and Indian languages, in this case, Hindi. The task is to identify the same news story written in another language, and is thus a problem of cross language news story detection. It can also be interpreted as duplicate detection where the query is a news document and retrieved documents are equivalent news documents but in a different language, see Fig.1.

In our investigation of the CL!NS English-Hindi task, we explore the use of the Google and Bing translation services to translate English queries into Hindi, and use the Google transliteration service to handle cases where the translation fails. Since a long query created from an English news story

Figure 1: Cross Language News Story Detection

might add noise to the retrieval process, we examine formulation of effective queries using content summarization and then perform query expansion using PRF. To capture the diversity of results and ensure that we have high recall, we explore data fusion methods to combine information captured using different query formulations.

The remainder of this paper is structured as follows: Section 2 describes the related work in CLIR, Section 3 outlines the methodology we used to perform cross language news search, Section 4 summarizes the datasets used in the FIRE 2013 CL!NSS task, Section 5 provides a detailed description of our experimental work, Section 6 discusses our submitted runs and presents results of the CL!NSS'13 task, Section 7 presents an analysis of the effects of the various techniques we used to address the cross language search problem, and Section 8 concludes the paper with a summary of our work so far and future research directions.

2. RELATED WORK

CLIR systems aim to support IR across languages where queries are entered in one language to retrieve documents rendered in another one. Developments in this field have progressed due in large part to major evaluation forums: i) TREC (Text retrieval conference) from 1997-2002 (some western European languages and Arabic), ii) CLEF (Conference and Labs of the Evaluation Forum) (primarily European languages), iii) NTCIR Asian language evaluation forum (covering east Asian languages) and iv) FIRE (Forum for Information Retrieval Evaluation) (Indian languages). IR evaluation tasks run within these forums have led to great progress in the area of CLIR. Most of the earlier work within these tasks focused on participants building their own models for translation services often using dictionary-based approaches for converting input from one language to another [5] [11]. With progress in machine translation (MT) technologies the focus shifted to translation using available MT services such as the Google or Bing translation service. Attention has also been given to query analysis and formulation of better queries to account for the loss of information in translation. For example, in [11] the authors used a dictionary-based approach for translation and explored query expansion techniques for CLIR, their experiments showed that query expansion can help to compensate for the loss of information during translation. In [9] and [2], the authors explored different techniques for query combination, relevance feedback and fusion methods for performing CLIR. In [2] the authors have applied PRF using the Robertson selection value [10] to find relevant terms for query expansion. Various techniques for combining scores have been tried and compared against query modification approaches. Different query modification techniques and ranking models output different ranked lists. Combinations of these ranked lists has also been explored in detail [1, 4, 6].

Our work continues this strategy of investigating query translation, query modification and combination of methods in CLIR, applying this approach to the task of finding similar news articles across languages pairs. There has not been a lot of work in exploring and adapting traditional techniques of CLIR for our target language pair of English-Hindi. One exception is [3] which uses a word-aligned model to perform CLIR for Indian languages. In our work we use the Google and Bing translation services to translate queries from English to Hindi. We explore traditional approaches to query formulation using summarized queries and named entity transliteration. We perform PRF to expand queries by selecting the top terms from a list of results retrieved using input queries. We try different fusion techniques and their combination with query formulation techniques to improve CLIR performance.

3. METHODOLOGY

In this section, we provide an overview of our methodology for addressing the cross language news search problem for the CL!NSS task at FIRE 2013. Figure 2 presents an outline of our system used for this task. Of the two components (1 and 2) used for Query Formulation, component 2 (Pseudo Relevance Feedback) was integrated and experimented *after* the formal task submission and was not a part of our official submission for the task. The following subsections present the details of our system.

3.1 Data Procurement

We used the open source Lucene search engine library to perform IR, i.e. indexing the input documents and searching the queries over the target collection. We used Lucene version-4.4.0[1] for our experiments. While indexing the documents we used Lucene's inbuilt Hindi Analyzer which performs stopword removal and stemming over the documents. The stopword list we used was obtained by concatenating different standard stopwords list for the Hindi language: i) the FIRE Hindi stopword list[2], ii) the Lucene internal stopword list, and iii) a stopword list created by selecting all the words with document frequency (DF) greater than 5,000 in the target document collection.

The input queries as well as the target documents have the same structure. Each news document has a title, date and content field. We indexed all three fields of the documents using the Lucene engine. We used Lucene's default scoring function for our experiments. Lucene's scoring function[3]

---

[1] http://lucene.apache.org/core/
[2] http://www.isical.ac.in/~fire/resources.html
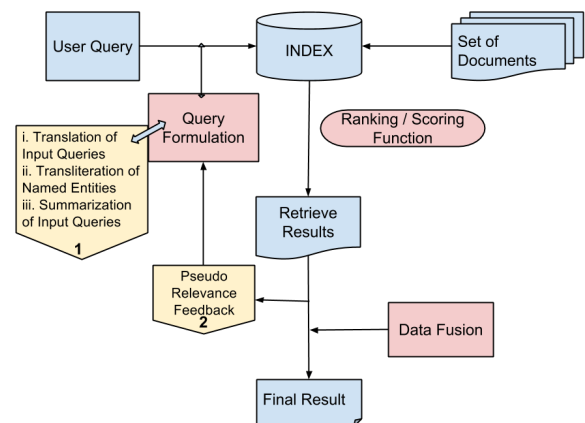[3] http://ipl.cs.aueb.gr/stougiannis/default.html

Figure 2: Our System Architecture

used in this work is a variant of a standard TF-IDF function. We explored query formulation and fusion techniques for cross language search for which the retrieval and ranking models are the same across all our experiments.

3.2 Query Formulation

Input queries are in the English language with the length of documents varying in the training and test set as shown in Table 1. The input queries were processed and analyzed to capture the information need effectively. We used different techniques to formulate queries as discussed below.

Dataset	Min Length	Max length	Avg Length
Training set	4.0	68.0	18.0
Test set	4.0	40.0	14.0

Table 1: No of sentences in the training and test set queries

3.2.1 Query Translation

We used MT systems to translate the queries since the source and target language of the query and target documents is different. Input query documents were translated from English to Hindi using the online Google[4] and Bing[5] translation services. We explored the use of two different translation services to minimize the bias of results and conclusions arising from the use of one translation service and to allow us to compare retrieval effectiveness with different translation services.

3.2.2 Transliteration

We observed that the Hindi target documents contained both words in the translated and transliterated forms of input queries as shown in Table 2. The use of the translated or transliterated forms in the documents was not predictable, and thus we hypothesized that it is advisable to include both forms in translated queries applied to the IR system.

News contains information pertaining to events, people or activities. To make sure all important parts of the news documents are captured, we ran a Named Entity Recognizer (NER) on the input queries and extracted a list of named entities, which were then transliterated to capture the language variants. We used the Stanford CoreNLP tool[6] to perform Named Entity extraction on the input English queries and extract the words which have a NER tag (*Person*, *Location* and *Organization*). The list of named entities for each query was transliterated using Google Transliteration[7]. Transliterated named entities were merged with the MT translated input queries.

Figure 3 shows the sequence of steps performed to form queries using a combination of query translation and transliteration of named entities. In (1) we have an English query with NE's identified by the Stanford tool, the input English query is translated using (Google/Bing) translation service

| English Word	Translated Word	Transliterated Word
Commonwealth Games	राष्ट्रमंडल खेल	कामनवेल्थ गेम्स

Table 2: Handling named entities

in (2) and the NE's identified in (1) are transliterated in (3). In (4) we merge the translated query and transliterated NE's to obtain the final combined query. Using transliteration, helps to capture both alternative forms of translation, but also words which are out of vocabulary for the MT systems which would otherwise remain untranslated from the input query. An example of how transliteration helps to capture variation is shown in Table 2.

3.2.3 Query Summarization

Not all parts of a query document are as important as others in describing the key themes of the document. In fact, some parts of the document can distract from the main topical content of the document. The paragraph/sentence content from a document which is more important to its main topic should be ranked higher in the retrieval process in seeking to find relevant similar documents between the query and target documents. We hypothesized that selecting the k sentences/paragraphs which are most important to the topic of the document and using these as the basis of our search query can prune noise and divergent content, hence yield a more effective query. The main challenge in exploring this approach comes in determining the selection criteria for these elements and the optimum size for the summary of the input documents.

In this investigation, we explored the sentence-based summarizer developed in our lab, at DCU, CNGL [13] to score and rank the sentences in a document. We used summarization over the input queries, and varied the length of the summaries to try to ensure that we do not lose relevant information by removing too many sentences and capture the main aspects of query document effectively. The length of the summary in our experiments is directly proportional to
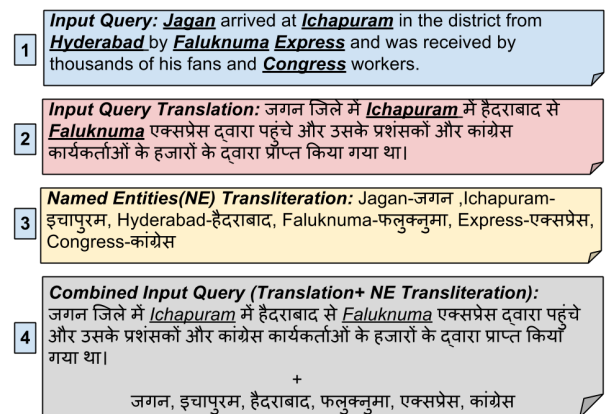
Figure 3: Combining Translation of Query and Transliteration of NE's

the length of the input query. As the query length parameter varies across the different queries, there is a trade off when selecting a particular fixed value of length of summaries for all input queries.

We used the following basic features of the summarizer to generalize our model:

- *Skimming*: This feature incorporates the position of a sentence in a paragraph. The underlying assumption is that sentences occurring early in a paragraph are more important for a summary.

- *NamedEntity*: This feature calculates the number of named entities that occur in each sentence. Any word (except the first in a sentence), that starts with a capital letter is assumed to be a named entity.

- *TSISF*: This is similar to a TF-IDF function, but works on the sentence level. Every sentence is treated like a document.

- *TitleTerm*: This feature scores the sentences by matching the overlap with the terms in the title.

- *ClusterKeyword*: This feature finds the relatedness between words in a sentence.

Using each of the above features the sentences in the documents are scored and ranked. The top k sentences constitute the summary of the input document [13].

3.2.4 Pseudo Relevance Feedback (PRF)

We use PRF to expand the input queries and then perform search. We used the Robertson selection value, RSV scores [10] to select the terms for expansion. The Robertson scale value is defined as:

$$rsv(i) = r(i) * rw(i) \qquad (1)$$

where $r(i)$ is number of relevant documents containing i, and $rw(i)$ is the standard Robertson/Jones relevance weight [10].

$$rw(i) = log \frac{(r(i) + 0.5)(N - n(i) - R + r(i) + 0.5)}{(n(i) - r(i) + 0.5)(R - r(i) + 0.5)} \qquad (2)$$

where $n(i)$ = total number of documents containing term i, R = total number of relevant documents for this query, N = total number of documents in the collection. We tried different combinations of the parameter R and the number of terms to be added to the original query in the query expansion, to select the optimum values.

3.3 Data Fusion

Data fusion is a well established technique in IR for merging results from multiple retrieval systems or merging results obtained by varying queries and searching over the same system [6]. The underlying motivation for adopting this approach is that documents retrieved using multiple approaches are more likely to be relevant to the information need underlying the search query. Using fusion techniques, the documents which occur in multiple retrieval results are given a boost and are ranked higher. Each retrieved list of documents has a rank and a score from the the search engine. Since the scores of documents retrieved using different methods or systems lie in different ranges due to the variations in the retrieval methods used in their creation, the scores of the retrieved documents list need to be normalized before they are combined with the scores retrieved by other systems. A standard technique for normalization in data fusion is referred to as the *min-max* method. This is defined as follows:

$$normalized\ \ score = \frac{unormalized\ \ score - minimum\ \ score}{maximum\ \ score - minimum\ \ score}$$

We used the CombMNZ [1] method to combine results across different ranked lists. To do this, we retrieved the top 200 results from each system and then used data fusion to combine results.

CombMNZ Score: For documents retrieved across different systems the average score of the document is calculated and then multiplied by the number of systems using which the document was found.

$$AverageScore = \frac{summation\ of\ individual\ retrieval\ results}{total\ systems}$$

$$Frequency = number\ of\ non\ zero\ retrievals$$

$$CombMNZ = AverageScore * Frequency$$

4. DATASETS
In this section, we give a brief overview of the CL!NSS task dataset.

- **Hindi Document Collection**: The target documents were 50,691 news documents in the Hindi language. All the news documents have 3 main fields: title of the news document, date when the news was published and the content of the news article.

- **English training dataset**: The training dataset had 50 documents in the English language. Each of these had 3 main fields (title, date and content) similar to the target documents. The variation in terms of the length of training documents is shown in Table 1.

- **English test dataset**: The test dataset had 25 documents in the English language. These documents also had 3 main fields: title, date and content. The variation in terms of the length of test documents is shown in Table 1.

5. EXPERIMENTAL DETAILS
In this section, we discuss the experimental details of our system. We present our experiments performed using varied parameters and combination settings. We describe in detail the results of different approaches used in our experiments as compared to our baseline system.

5.1 Baseline
The baseline system for our experiments was obtained using the raw queries translated using Google and Bing translation services in isolation, as shown in Table 3. Table 3 shows results for Lucene search using the translated input queries, and that these perform far better than the best run of the CL!NSS task at FIRE 2012 [8]. However, this performance

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Palkovski	0.322	0.326	0.339	0.362
Bing Baseline	0.520	0.477	0.498	0.514
Google Baseline	**0.581**	**0.518**	**0.521**	**0.549**

Table 3: Comparison of baseline runs with FIRE 2012 best result

discrepancy could be explained by the fact that, unlike previous year's participants in the CL!NSS task, we had training data for our method which we were able to use during system development.

5.2 Performing Query Modification

CL!NSS is quite different from normal CLIR tasks where generally the queries range from about 3-10 words. In the CL!NSS task queries are whole news documents with an average length of about 18 sentences as shown in Table 1. As queries are whole news documents, we reformulate them to capture the key information needed.

5.2.1 Transliteration

As discussed in Section 3.2.2, the Hindi target documents have words which are the translated and transliterated form of input queries, as shown in Table 2. Thus using just translation can lead to mismatch failures which can be addressed by performing transliteration of the named entities in the input queries as shown in Figure 3. To investigate this issue, we merged the transliterated NE's with the translated input query and tested the effect of incorporating the NE's transliteration with the translated input query.

Table 4: Combining transliteration and translation of queries

| System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Using Google				
Baseline	0.581	0.518	0.521	0.549
Translation+				
Transliteration	0.584	0.523	0.529	0.556
Using Bing				
Baseline	0.520	0.477	0.498	0.514
Translation+				
Transliteration	0.469	0.495	0.508	0.523

As shown in Table 4, incorporating transliterated words in addition to the translated queries slightly improves the system performance. For both Google and Bing translated input queries, (apart from NDCG@1 for Bing translated query), the NDCG scores improves when adding transliterated named entity information.

5.2.2 Summarizer

We performed summarization on the input queries. The summarizer used is discussed in detail in Section 3.2.3. We tried different combinations of summary length on the training set, plotting the graph of system performance with respect to summary length. The summarized query is translated using both Google and Bing translation.
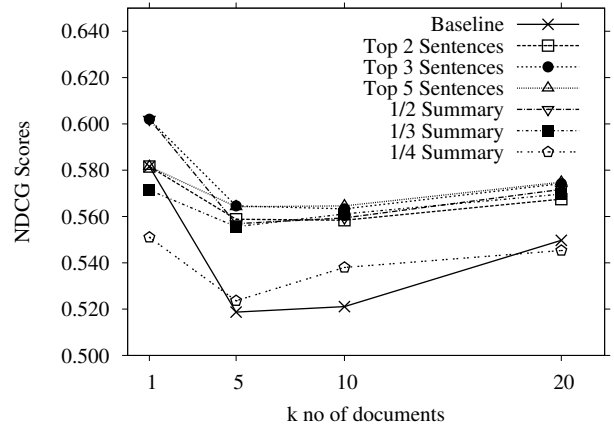
Figure 4: Varying summary length and comparing NDCG scores at different K values, using Google translation, where 1/2 means the size of summary is half the size of the input query and similarly for 1/3 and 1/4 summary.
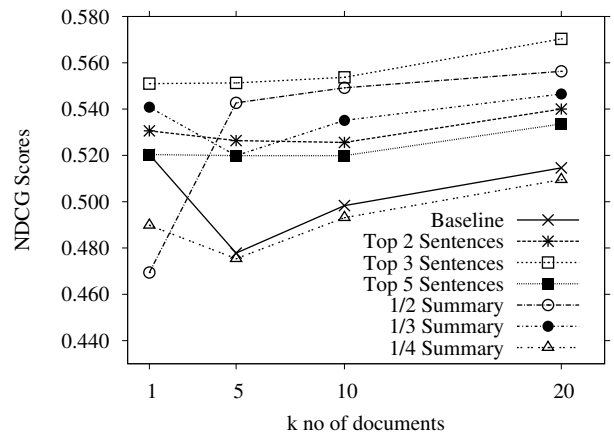
Figure 5: Varying summary length and comparing NDCG scores at different K values, using Bing translation.

Varying the length of summary directly changes the systems' performance as shown in Figure 4 and Figure 5. Thus it is important to select the length of summary carefully. The performance using a summary instead of the whole document as an input query is significantly better across both Google and Bing translation. To ensure diversity in terms of the summarized queries, we select, for further experiments, one with fixed length of summary (the top 3 sentences) and one with variable length which depends on the input length of the query (the one-third summary).

5.2.3 Using PRF

Initial retrieved results were used to expand the query as discussed in Section 3.2.4. The input parameters for performing PRF are: i) the number of relevant documents to be considered for populating terms for query expansion, and ii) the number of terms to be added to the original query in query expansion.
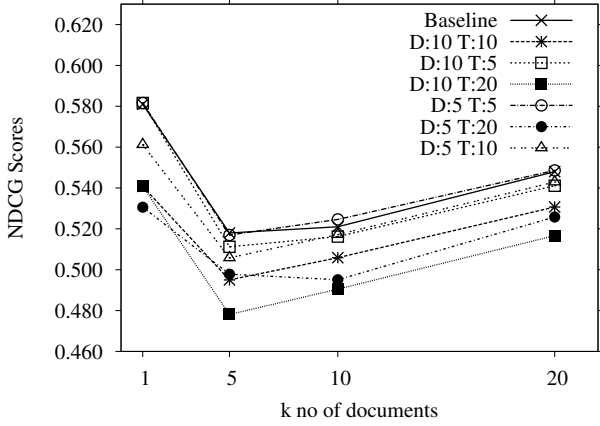
Figure 6: Varying the Number of Documents and Terms for calculating optimum values for PRF, using Google translation for queries, where D indicates number of documents and T indicates number of words
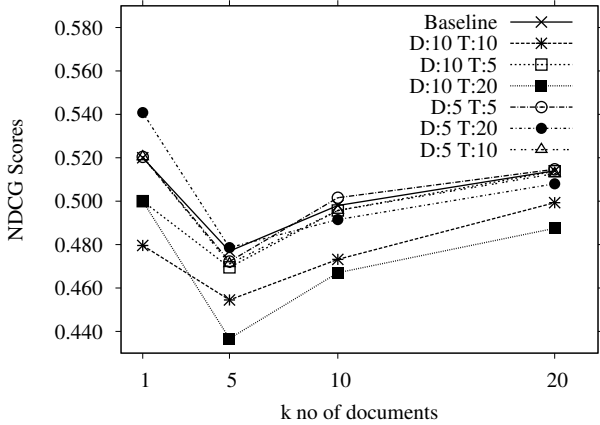

Figure 7: Varying the Number of Documents and Terms for calculating optimum values for PRF, using Bing translation for queries.

It is quite complex to determine the optimum value of the number of documents and number of words. Figures 6 and 7 indicate that not all the combinations of documents and words perform well. For further analysis and experiments, we chose the combination of 5 documents and 5 words which performs equally well or better than the baseline at all NDCG values for both Google and Bing translated queries.

5.2.4 Combining PRF and Summarizer Approaches

We wanted to explore how well the system performs when PRF is applied to the summary rather than the complete query. We performed experiments combining different query formulation approaches namely: i) summary of input queries (Top-3 sentences and one-third length summary), ii) transliteration of named entities, and iii) applying PRF on the input queries. Table 5 shows that performing PRF over search results retrieved using summarized queries improves the performance over the individual use of PRF on input

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Using Google				
Google Baseline	0.581	0.518	0.521	0.549
Raw+PRF	0.561	0.515	0.506	0.533
Top3+PRF	0.581	0.548	0.554	0.570
1/3+PRF	0.591	0.557	0.558	0.570
Raw+NE	0.581	0.518	0.521	0.549
Top3+NE	**0.653**	0.556	**0.563**	0.576
1/3+NE	0.602	0.557	0.557	0.571
Raw+NE+PRF	0.581	0.517	0.524	0.548
Top3+NE+PRF	0.632	**0.563**	0.559	**0.579**
1/3+NE+PRF	0.581	0.551	0.560	0.572
Using Bing				
Bing Baseline	0.520	0.477	0.498	0.514
Raw+PRF	0.449	0.459	0.483	0.504
Top3+PRF	0.520	0.531	0.542	0.556
1/3+PRF	0.520	0.481	0.497	0.519
Raw+NE	0.5204	0.4811	0.4972	0.5194
Top3+NE	0.551	**0.559**	**0.556**	**0.571**
1/3+NE	**0.561**	0.527	0.548	0.555
Raw+NE+PRF	0.540	0.482	0.495	0.517
Top3+NE+PRF	0.540	0.545	0.546	0.560
1/3+NE+PRF	0.540	0.526	0.539	0.550

Table 5: Combination of Query Formulation Approaches on the training data where *Raw* indicates normal translated query, *PRF* performed pseudo relevance feedback, *NE* named entities transliterated merged with the input query, *Top3* indicates summary using top-3 sentences and 1/3 indicates summary using 1/3 summary of input queries

queries or using just summarized queries for search. The best combination using the Google translation service with the top 3 sentence summary and transliterated named entities has the following scores: NDCG@1 0.653, NDCG@5 0.556, NDCG@10 0.563 and NDCG@20 0.576 and is statistically significant[8] better as compared to the baseline (p = 0.024). Each of the combination methods performs better than the baseline using just raw translated documents as input queries.

5.3 Data Fusion

We used data fusion to combine the results of multiple retrieval runs obtained using query formulation. As described earlier, the standard *CombMNZ* method was used to combine the retrieved ranked list of documents. We aimed to determine the potential utility of combining multiple search results obtained using alternative query formulations, and to compare this with results obtained using single queries formed using different formulation methods. The systems used for the combination experiments were those which performed well as discussed above in Section 5.2.

Instead of modifying the query we tried to combine the different ranked lists, selecting the best of all systems for the combination to see the relative effectiveness of fusing different information. We tried four different fusion combinations using both the Google and Bing translation services.

---

[8] We used the paired t-test to calculate the statistical significance of our results. We calculated MAP scores for each query and used it for finding the p-value.

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Using Google				
Google Baseline	0.581	0.518	0.521	0.549
CombSys	0.622	**0.574**	0.574	0.590
CombSys+PRF	0.622	0.550	0.562	0.573
CombSys+NE	**0.632**	0.573	**0.579**	**0.591**
CombSys+NE+PRF	0.602	0.552	0.560	0.577
Using Bing				
Bing Baseline	0.52	0.477	0.498	0.514
CombSys	**0.602**	0.526	0.548	0.558
CombSys+PRF	0.571	**0.558**	**0.575**	**0.583**
CombSys+NE	0.581	0.545	0.552	0.564
CombSys+NE+PRF	0.571	0.548	0.554	0.566

Table 6: Fusion Results on the training set

- *CombSys*: Default Query, Top 3 Sentence Summary and One Third Summary.

- *CombSys+PRF*: Default Query, Top 3 Sentence Summary and One Third Summary with PRF performed over the queries.

- *CombSys+NE*: Default Query, Top 3 Sentence Summary, One Third Summary with each query having NE's transliterated mer-ged with the translated query.

- *CombSys+NE+PRF*: Default Query, Top 3 Sentence Summary, One Third Summary all with PRF performed over the queries having NE's transliterated merged with translated query.

The results of the fusion approach are shown in Table 6. The combination approach using fusion techniques performs better than individual query formulation techniques. The best system on the training set *CombSys+NE* performs considerably better than the baseline. The difference is statistically significant ($p = 0.019$).

Based on the best parametric settings for the training set and the combination metric used over the training set, we conducted similar experiments with the test data. Table 7 shows the query formulation scores for the test set and Table 8 shows the fusion scores for the test data

Tables 7 and 8 indicates that NDCG scores using Bing translation are better than those obtained using Google translation. The query formulation results are better than fusion results for the test set. The approach of translating queries using Bing translation and using transliteration of named entity information and performing PRF shows the best performance over all other approaches applied to the test set.

6. CL!NSS'13 TASK SUBMISSION

For the submission of the CL!NSS task we wanted to capture the diversity in our runs[9]. We submitted a system which is a combination of different retrieved lists, using summary features with different lengths and translation services that performed well over the training set. The following features were selected for our final runs:

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Using Google				
Google Baseline	0.760	0.673	0.689	0.691
Raw+PRF	0.720	0.677	0.686	0.692
Top3+PRF	0.720	0.642	0.658	0.662
1/3+PRF	0.640	0.677	0.694	0.689
Raw+NE	0.760	0.673	0.689	0.691
Top3+NE	0.720	0.666	0.676	0.679
1/3+NE	0.680	**0.700**	**0.706**	**0.706**
Raw+NE+PRF	0.720	0.677	0.686	0.692
Top3+NE+PRF	**0.780**	0.651	0.665	0.667
1/3+NE+PRF	0.660	0.694	0.700	0.698
Using Bing				
Bing Baseline	0.780	0.734	0.748	0.747
Raw+PRF	0.760	0.739	0.745	0.748
Top3+PRF	0.720	0.638	0.667	0.676
1/3+PRF	0.720	0.677	0.684	0.706
Raw+NE	**0.780**	0.736	**0.749**	**0.751**
Top3+NE	0.760	0.669	0.691	0.704
1/3+NE	0.720	0.710	0.733	0.744
Raw+NE+PRF	0.760	**0.749**	0.747	0.750
Top3+NE+PRF	0.720	0.669	0.693	0.697
1/3+NE+PRF	0.720	0.713	0.723	0.740

Table 7: Combination of Query Formulation Approaches on the test set

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Using Google				
Google Baseline	0.760	0.673	0.689	0.691
CombSys	**0.760**	**0.682**	**0.708**	**0.700**
CombSys+PRF	**0.760**	0.669	0.695	0.697
CombSys+NE	**0.760**	0.681	0.706	**0.700**
CombSys+NE+PRF	0.740	0.680	0.701	**0.700**
Using Bing				
Bing Baseline	0.780	0.734	0.748	0.747
CombSys	0.720	0.709	0.724	0.732
CombSys+PRF	0.720	0.709	0.722	0.731
CombSys+NE	0.720	0.713	0.730	0.737
CombSys+NE+PRF	0.720	0.702	0.725	0.731

Table 8: Fusion Results on test set

- **Using Google Translation**

  - Using 1/3 summary of input query.

  - Using 3-sentence summary of input query.

  - Using 3-sentence summary of input query merged with all named entities transliterated using Google transliteration.

  - Using complete input query merged with all the named entities transliterated using Google Transliteration.

- **Using Bing Translation**

  - Using 1/3 summary of input query

  - Using 3-sentence summary of input query

  - Using complete input query merged with all the named entities transliterated using Google Transliteration.

- **Using Date Feature**: The date of publication of a news article gives an idea of the proximity of another news document. Under the assumption that closer

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Run-1	0.571	0.555	0.561	0.569
Run-2	**0.663**	**0.583**	**0.580**	**0.595**
Run-3	0.602	0.565	0.569	0.580

Table 9: System combinations results on training set

System	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Run-1	**0.740**	0.665	0.675	0.684
Run-2	**0.740**	0.670	0.704	0.704
Run-3	**0.740**	**0.680**	**0.726**	**0.724**

Table 10: Final results on test set

proximity means that documents are more likely to be related, we give a small boost to all the retrieved documents which appeared within a window of some threshold days before or after the query document. We conduct multiple experiments and empirically chose the $boost=0.04$ and $threshold=10$ for our final run.

The retrieved results of the above 7 features were combined using data fusion. The three runs we submitted were combined in the following way.

- Run-1: Using Google translation and one third summary of queries.

- Run-2: Using Google translation and combining one third summary of queries, 3-sentence summary of queries, one third summary of query merged with all named entities transliterated using Google transliteration and using whole query merged with named entities transliterated + incorporating the date factor

- Run-3: Combining all the 7 features, i.e including the queries translated using both Google and Bing. Using complete query as well as 1/3 summary and 3-sentence summary of the query with and without merging NE transliterated, all fused together.

Table 9 presents the results of three runs on the training set, while Table 10 shows the official results of the three runs on the test set.

The results of our runs, shown in Table 10, were ranked first out of the formal submissions for NDCG@5 and NDCG@10, and second for NDCG@1. We tried to use query information wisely in the form of combining different summaries to capture the information need. Fusion techniques improve the systems' recall. Incorporating named entity transliteration helps to handle out-of-vocabulary words.

7. ANALYSIS

We next discuss the effect of the various approaches and their combinations which we investigated for cross language search.

7.1 Translation

As discussed previously, we used the Bing and Google translation services to translate input English queries to the Hindi language. We observed for the experiments carried out on the training set that all the results with respect to Google translation outperformed the Bing translation results. However, for the test data, the behaviour is reversed, where results obtained using Bing translation outperform those for Google translation. Our best system combination on the training data is the combination using Google translation but post-submission experiments indicate that for test data, the best system combination is obtained by fusing Bing translation and its variants. A possible reason for these results could relate to the nature of the query documents in the training and test sets, where differences in the document language may lead to differences in the output quality of the translation service.

7.2 Transliteration

Transliteration of named entities appears to be useful for English Hindi cross language search, with improvements for both training and test queries. However, automatic transliteration may be incorrect leading to inappropriate matches or failures to match. Transliteration is sometimes a complex task, the transliterated word can have different representations for the same word based on its pronunciation. For example, the abbreviation $LTTE$ has two possible transliterations: एलटीटीई and लिट्टे. Both the transliterations are valid and are frequently used. Google transliteration output has some errors as it fails to handle the spelling variations in the Hindi language and wrongly maps characters. For example, $PLGA$ is transliterated as ब्रग by Google transliteration where the actual transliteration is पीएलजीए.

7.3 Summarization

For long documents such as news stories, it is advantageous to use a summary of the whole news documents as a query. As shown in Figure 4 and Figure 5, all the combinations of summary apart from one-fourth for Bing translation have NDCG scores higher than the baseline. Combination of the summary with other techniques boosted the performance for the training set. However, for the test set use of top-3 sentence and one-third summary degraded the results as compared to using the baseline full document queries, as shown in Table 7 and Table 8. As shown in Table 1, the average length of the test queries is considerably shorter than the training queries, suggesting that they may be better focussed for user queries, less in need of summarization and that application of summarization may even remove important topical terms which are better for search.

7.4 Pseudo-Relevance Feedback

In our experiments, we applied PRF for query expansion. In general, we find that using PRF has a positive effect on the performance of the system. Determining the optimum values for the number of documents and words for query expansion is a complex problem. For our training set the combination of five documents and five terms performed well, but the same combination did not perform as well on the test set, possibly the effectiveness of PRF for the test set might be improved by adjustments of the fixed parameter values.

7.5 Data Fusion

Our results indicate the using data fusion has the potential to improve cross language search effectiveness by combining multiple types of information together. However, similar to results observed for query summarization, for our training set, the fusion technique performs better than any single formulation of the input query, whereas for the test set, the performance is less convincing. One possible reason for this may be less diversity in the relevance set for the test set, meaning that there is less potential for data fusion to act in a beneficial manner.

8. CONCLUSIONS AND FUTURE WORK

The analysis of the experimental results in Section 7 highlights some interesting aspects of the cross language news search task and the FIRE 2013 data sets. It proved useful to explore the use of two different translation services, as we see that the performance of the translation service combined with the query formulation techniques is reversed in the test set as compared to the training set. The nature of the collection of query documents plays an important role in the performance of the system. The tuning of a system based on the training data may not always work well for the test collection.

Certain challenges remain unexplored in our current study. For example, abbreviations such as "MNIK","YSR", movie names and political party names should be handled in a systematic way. In addition, handling spelling variants is a significant challenge. Stemming takes care of the affixes. However the main problem for Hindi arises with handling the diacritic marks and vowel variations. With better normalization techniques we would be able to handle the erroneous cases and capture the missing information. We also plan to explore the use of alternative scoring functions such as BM25 and other variants.

9. REFERENCES

[1] J.H. Lee: Analyses of Multiple Evidence Combination, In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 97), pages 267-275, Philadelphia, Pennsylvania, USA (1997)

[2] A.M. Lam-Adesina and G.J.F. Jones, Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval, In proceeding of: Comparative Evaluation of Multilingual Information Access Systems, 4th Workshop of the Cross Language Evaluation Forum, CLEF 2003,

[3] J. Jagarlamudi and A.Kumaran, Cross-Lingual Information Retrieval System for Indian Languages,Advances in Multilingual and Multimodal Information Retrieval Lecture Notes in Computer Science Volume 5152, 2008,

[4] N.J. Belkin, P. Kantor, E.A. Fox and J.A. Shaw: Combining the evidence of multiple query representations for information retrieval, Information Processing and Management 31(3):431–448 (1995)

[5] J.Y. Nie, M. Simard, P. Isabelle and R. Durand, Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

[6] W.B. Croft: Combining Approaches To Information Retrieval, Advances Information Retrieval: Recent Research from the CIIR, Springer (2000)

[7] P. Gupta, P. Clough, P. Rosso, M. Stevenson and R. E. Banchs, PAN@FIRE 2013: Overview of the Cross-Language !ndian News Story Search (CL!NSS) Track. In Proceedings of the Fifth Forum for Information Retrieval Evaluation (FIRE 2013), New Delhi, India (2013)

[8] P. Gupta, P. Clough, P. Rosso and M. Stevenson, PAN@FIRE 2012: Overview of the Cross-Language !ndian News Story Search (CL!NSS) Track. In Proceedings of the Fourth Forum for Information Retrieval Evaluation (FIRE 2012), Kolkata, India (2012)

[9] G.J.F. Jones and A.M. Lam-Adesina, Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval, Proceedings of the CLEF 2001: Workshop on Cross Language Information Retrieval and Evaluation, Darmstadt, Germany, pages 59-77 (2002)

[10] S.E. Robertson, and K.S. Jones, Relevance weighting of search terms. Journal of the American Society for Information Science 27 (3): 129-146; 1976.

[11] L. Ballestems and W.B. Croft, Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval, SIGIR 97 Philadelphia PA, USA

[12] P. Arora, J. Foster, G.J.F. Jones, DCU at FIRE 2013: Cross Language !ndian news story search: FIRE-Forum for Information Retrieval Evaluation, 5-7 Dec 2013, Bangalore, India.

[13] L. Kelly, J. Leveling, S. McQuillan, S. Kriewel, L. Goeuriot, and G.J.F. Jones: Report on summarization techniques, Khresmoi project deliverable D4.4 (2013)