Finding "Good Enough": A Task-Based Evaluation of Query Biased Summarization for Cross Language Information Retrieval

Jennifer Williams, Sharon Tam, Wade Shen, *
MIT Lincoln Laboratory Human Language Technology Group
244 Wood Street, Lexington, MA
{jennifer.williams, sharon.tam, swade} @ll.mit.edu

Abstract

In this paper we present our task-based evaluation of query biased summarization for cross-language information retrieval (CLIR) using relevance prediction. We describe our 13 summarization methods each from one of four summarization strategies. We show how well our methods perform using Farsi text from the CLEF 2008 shared-task, which we translated to English automtatically. We report precision/recall/F1, accuracy and time-on-task. We found that different summarization methods perform optimally for different evaluation metrics, but overall query biased word clouds are the best summarization strategy. In our analysis, we demonstrate that ROUGE scores cannot make the same distinctions as our evaluation framework does. Finally, we present our recommendations for creating much-needed evaluation standards and datasets.

1 Introduction

Despite many recent advances in query biased summarization for cross-language information retrieval (CLIR), there are no existing evaluation standards or datasets to make comparisons among different methods, and across different languages. Consider that creating this kind of summary requires familiarity with techniques from machine translation (MT), summarization, and information retrieval (IR). In this paper, we arrive at the intersection of each of these research areas. Query biased summarization involves automatically capturing relevant ideas and content from a document with respect to a given query, and presenting it as a condensed version of the original document. This kind of summarization is mostly used in search engines because when search results are tailored to a user's information need, the user can find texts that they are looking for more quickly and more accurately (Tombros and Sanderson, 1998; Mori et al., 2004). Query biased summarization is a valuable research area in natural language processing (NLP), especially for CLIR. Users of CLIR systems meet their information needs by submitting their queries in L_1 to search through documents that have been composed in L_2, even though they may not be familiar with L_2 (Hovy et al., 1999; Pingali et al., 2007). Cross-language query biased summarization is an important part of CLIR, because it helps the user decide which foreign-language documents they might want to read.

How do we know if a query biased summary is "good enough" to be used in a real-world CLIR system? There are no standards for objectively evaluating summaries for CLIR – a gap that we begin to address in this paper. We treat the actual CLIR search engine as a black box and instead we focus on finding out if the summaries themselves are useful. While extracted sentences or snippets of text may be acceptable for a typical monolingual IR system, we show that is not necessarily the case when summarizing for CLIR systems. The problem we explore in this paper is two-fold: what kinds of summaries are well-suited for CLIR applications, and how should we evaluate them?

In this work, we present 13 summarization methods. Each one of our methods belong to a summarization strategy: (1) unbiased machine translated text, (2) unbiased word clouds, (3) query biased word clouds, and (4) query biased sentence summaries. The methods and strategies that we present are fast, cheap, and language-independent. Our evaluation is based on a relevance prediction task: the user must decide if the

* This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

summary of a given document is relevant to a particular information need, or not (Hobson et al., 2007). To the best of our knowledge, we are the first to investigate an evaluation framework that would allow comparisons across languages and across summarization methods.

We approach our task as an engineering problem: our goal is to create summaries of foreign-language documents that are good enough to help CLIR system users find what they are looking for. We have simplified the task by assuming that such documents have already been retrieved from a search engine, as CLIR techniques are outside the scope of this paper. As a starting point, we begin with some principles that we expect to hold true when we evaluate. These principles provide us with the kind of framework that we need for a productive and judicious discussion about how well a summarization method is really working. We encourage the NLP community to consider the following concepts when developing evaluation standards for this problem:

- End-user intelligiblity
- Query-salience
- Retrieval-relevance

Summaries should be presented to the end-user in a way that is both concise and intelligible, even if the machine translated text is difficult to understand. Our notions of *query-salience* and *retrieval-relevance* capture the expectation that good summaries will be efficient enough to help end-users fulfill their information needs. For query-salience, we want users to positively identify relevant documents. Similarly, for retrieval-relevance we want users to be able to find as many relevant documents as possible.

This paper is structured as follows: Section 2 presents related work; Section 3 describes our data and pre-processing; Section 4 details our summarization methods and strategies; Section 5 describes our experiments; Section 6 shows our results and analysis; and in Section 7, we conclude and discuss some future directions for the NLP community.

2 Related Work

Automatic summarization is a well-investigated research area. The earliest methods for creating summaries of documents were based on describing which terms appear in a given document and the relationship of those terms to information content (Luhn, 1958; Edmunson, 1969; Salton and Yang, 1973; Robertson and Walker, 1994; Church and Gale, 1999; Robertson, 2004). Recent work has looked at creating summaries of single and multiple documents (Radev et al., 2004; Erkan and Radev, 2004; Wan et al., 2007; Yin et al., 2012; Chatterjee et al., 2012), as well as summary evaluation (Jing et al., 1998; Tombros and Sanderson 1998; Mani et al., 1998; Mani et al., 1999; Mani, 2001; Lin and Hovy, 2003; Lin, 2004; Nenkova et al., 2007; Hobson et al., 2007; Owczarzak et al., 2012), query and topic biased summarization (Berger and Mittal, 2000; Otterbacher et al., 2005; Daume and Marcu, 2006; Chali and Joty, 2008; Otterbacher et al., 2009; Bando et al., 2010; Bhaskar and Bandyopadhyay, 2012; Harwath and Hazen, 2012; Yin et al., 2012), and summarization across languages (Pingali et al., 2007; Orăsan and Chiorean, 2008; Wan et al., 2010; Azarbonyad et al., 2013).

There has been a lot of work on developing metrics for determining what makes a summary good. Evaluation metrics are either *intrinsic* or *extrinsic*. Intrinsic metrics, such as ROUGE, measure the quality of a summary with respect to gold human-generated summaries (Lin, 2004; Lin and Hovy, 2003). Generating gold standard summaries is expensive and time-consuming, a problem that persists with cross-language query biased summarization because those summaries must be query biased as well as in a different language from the source documents.

On the other hand, extrinsic metrics measure the quality of summaries at the system level, by looking at overall system performance on downstream tasks (Jing et al, 1998; Tombros and Sanderson, 1998). One of the most important findings for query biased summarization comes from Tombros and Sanderson (1998). In their monolingual task-based evaluation, they measured user speed and accuracy at identifying relevant documents. They found that query biased summarization improved the user speed and accuracy when the user was asked to make relevance judgements for IR tasks. We also expect that our evaluation will demonstrate that user speed and accuracy is better when summaries are query biased.

The previous work most closely related to

our own comes from Pingali et al., (2007). In their work, they present their method for cross-language query biased summarization for Telugu and English. Their work was motivated by the need for people to have access to foreign-language documents from a search engine even though the users were not familiar with the foreign language, in their case English. They used language modeling and translation probability to translate a user's query into L_2, and then summarized each document in L_2 with respect to the query. In their final step, they translated the summary from L_2 back to L_1 for the user. They evaluated their method on the DUC 2005 query-focused summarization shared-task with ROUGE scores. We compare our methods to this work also on the DUC 2005 task. Our work demonstrates the first attempt to draw at a comparison between user-based studies and intrinsic evaluation with ROUGE. However, one of the limitations with evaluating this way is that the shared-task documents and queries are monolingual.

Bhaskar and Bandyopadhyay (2012) tried a subjective evaluation of extractive cross-language query biased summarization for 7 different languages. They extracted sentences, then scored and ranked the sentences to generate query dependent snippets of documents for their cross lingual information access (CLIA) system. However, the snippet quality was determined subjectively based on scores on a scale of 0 to 1 (with 1 being best). Each score indicated annotator satisfaction for a given snippet. Our evaluation methodology is objective: we ask users to decide if a given document is relevant to an information need, or not.

Machine translation quality can also have an affect on summarization quality. Wan et al. (2010) researched the effects of MT quality prediction on cross-language document summarization. They generated 5-sentence summaries in Chinese using English source documents. To select sentences, they used predicted translation quality, sentence position, and sentence informativeness. In their evaluation, they employed 4 Chinese-speakers to subjectively rate summaries on a 5-point scale (5 being best) along the dimensions of content, readability, and overall impression. They showed that their approach of using MT quality scores did improve summarization quality on average. While their findings are important, their work did not address query biasing or objective evaluation of the summaries. We attempt to overcome limitations of machine translation quality by using word clouds as one of our summarization strategies.

Knowing when to translate is another challenge for cross-language query biased summarization. Several options exist for when and what to translate during the summarization process: (1) the source documents can be translated, (2) the user's query can be translated, (3) the final summary can be translated, or (4) some combination of these. An example of translating only the summaries themselves can be found in Wan et al., (2010). On the other hand, Pingali et al. (2007) translated the queries and the summaries. In our work, we used gold-translated queries from the CLEF 2008 dataset, and machine translated source documents. We briefly address this in our work, but note that a full discussion of when and what to translate and those effects on summarization quality, is beyond the scope of this paper.

3 Data and Pre-Processing

We used data from the Farsi CLEF 2008 ad hoc task (Agirre et al., 2009). The 50 queries (parallel English and Farsi) from this dataset each included a title, narrative, and description. For query-biasing, we used the title version. For our relevance prediction task on Mechanical Turk, we showed the narrative version. The dataset also included a ground-truth answer key indicating which documents were relevant to each query. For each query, we randomly selected 5 documents that were relevant as well as 5 documents that were not relevant. Our subset therefore included 500 original Farsi documents as well as the 50 parallel English-Farsi queries. Next we will describe our text pre-processing steps for both languages.

3.1 English Documents

All of our English documents came from automatically translating the original Farsi documents (Drexler et al., 2012). The translated documents were sentence-aligned with one sentence per line. For all of our summarization experiments (except showing full MT text), we processed the text as follows: removed extra spaces, removed punctuation, folded to lowercase, removed digits, and tokenized terms by splitting on whitespace. We also removed common English stopwords from the texts.

3.2 Farsi Documents

We used the original CLEF 2008 Farsi documents for two of our summarization methods. We stemmed each document using automatic morphological analysis with Morfessor CatMAP and we note that within-sentence punctuation was removed during this process (Creutz and Lagus, 2007). We also removed common Farsi stopwords as well as digits, and we tokenized terms by splitting on whitespace [1].

4 Summarization Strategies

All of our summarization methods, except for unbiased full machine translated text, were *extractive*. We used existing parts of a document to create a condensed version, or summary, of that document (Nenkova and McKeown, 2012). In this section, we present our 13 different summarization methods that we used for our task-based evaluation. Each of our summarization methods can be categorized into one of the following strategies: (1) unbiased full machine translated text, (2) unbiased word cloud summaries, (3) query biased word cloud summaries, and (4) query biased sentence summaries. Let t be a term in document d where $d \in D$ and D is the set of all documents in our experiments and $|D| = 1000$. Let q be a query where $q \in Q$ and Q is our set of 50 queries. Assume that log refers to log_{10}.

4.1 Unbiased Full Machine Translated English

Our first baseline approach was to make summaries from the raw machine translation output (no subsets of the sentences were used). Each summary therefore consisted of the full text of an entire document automatically translated from Farsi to English (Drexler et al., 2012). Although the full machine translated English text was not query biased, we highlighted words in yellow that also appeard in the query. Figure 1 shows an example full text document translated from Farsi to English and a gold-standard English query. Note that we use this particular document-query pair as an example throughout this paper (document: H-770622-42472S8, query: 10.2452/552-AH). According to our CLEF answer key, the example document is relevant to the example query.

[1]We used English and Farsi stopword lists from: http://members.unine.ch/jacques.savoy/clef/index.html

H-770622-42472S8 (document):
the price changes in the stock market tehran group economic: yesterday indicator of the entire stock price than he made the remarks 337 increased and 151664 . rate of also 30 increased and the last price yesterday 2102 each dollars. yesterday business in the stock market 693 and a hundred thousand and one part at a cost of 2 billion and 989 million and 452 and 839 by 361 of people like. yesterday stock price changes in the following. (the prices of is:

10.2452/552-AH (query, title):
Tehran's stock market

Figure 1: Full MT English summary and query.

We predicted that showing the full MT English text as a summarization strategy would not be particularly helpful in our relevance prediction task because the words in the text could be mixed-up, or sentences could be nonsensical, resulting in poor readability. We also predicted that it would take longer to arrive at a relevance decision for the same reasons.

4.2 Unbiased Word Clouds

For our second baseline approach, we ranked terms in a document and displayed them as *word clouds*. Word clouds are just a way to arrange a collection of words where each word can vary in text size and color. We used word clouds as a summarization strategy to overcome any potential disfluencies from the machine translation output and also to see if they are feasible at all for summarization. All of our methods for word clouds used words from machine translated English text. Each method described below generates a ranked list of terms. We created one word cloud per document using the top 12 ranked words. We used the raw term scores to scale text font size, so that words with a highter score appeared larger in the word cloud.

Term Frequency (TF) Term frequency is very commonly used for finding important terms in a document. Given a term t in a document d, the number of times that term occurs is:

$$tf_{t,d} = |t \in d|$$

Inverse Document Frequency (IDF) The idf term weighting is typically used in IR and other text categorization tasks to make distinctions between documents. Let N be the number of documents in the collection, such that $N = |D|$ and n_t is the number of documents that contain term t,

such that $n_t = |\{d \in D : t \in d\}|$, then:

$$idf_t = log \frac{N+1}{0.5 \times n_t}$$

While *idf* is usually thought of as a type of heuristic, there have been some discussions about its theoretical basis (Robertson, 2004; Robertson and Walker, 1994; Church and Gale, 1999; Salton and Yang, 1973). An example of this summary is shown in Figure 2.

Figure 2: Word cloud summary for inverse document frequency (IDF), for query "Tehran's stock market".

Term Frequency Inverse Document Frequency (TFIDF) We use $tfidf_{t,d}$ term weighting to find terms which are both rare and important for a document, with respect to terms across all other documents in the collection:

$$tfidf_{t,d} = tf_{t,d} \times idf_t$$

4.3 Query Biased Word Clouds

We generated query biased word clouds following the same principles as our unbiased word clouds, namely the text font scaling and highlighting remained the same.

Query Biased Term Frequency (TFQ) In Figure 3 we show a sample word cloud summary based on query biased term frequency. We define query biased term frequency tfQ at the document level, as:

$$tfQ_{t,d,q} = \begin{cases} 2tf_{t,d}, & \text{if } t \in q \\ tf_{t,d}, & \text{otherwise} \end{cases}$$

Query Biased Inverse Document Frequency (IDFQ) Inverse document frequency is known to help with identifying terms that discriminate among documents in a collection, so we would

Figure 3: Word cloud summary for query biased term frequency (TFQ), for query "Tehran's stock market".

expect that query biased idf can help to identify documents that are relevant to a query:

$$idfQ_{t,q} = \begin{cases} 2idf_t, & \text{if } t \in q \\ idf_t, & \text{otherwise} \end{cases}$$

Query Biased TFIDF (TFIDFQ) We define query biased $tf \times idf$ similarly to our TFQ and IDFQ, at the document level:

$$tfidfQ_{t,d,q} = \begin{cases} 2tf_{t,d} \times idf_t, & \text{if } t \in q \\ tf_{t,d} \times idf_t, & \text{otherwise} \end{cases}$$

Query Biased Scaled Frequency (SFQ) This term weighting scheme, which we call scaled query biased term frequency or sfQ, is a variant of the traditional $tf \times idf$ weighting. First, we project the usual term frequency into log-space, for a term t in document d with:

$$tfS_{t,d} = log(tf_{t,d})$$

We let $tfS_{t,d} \approx 0$ when $tf_{t,d} = 1$. We believe that singleton terms in a document provide no indication that a document is query-relevant, and treament of singleton terms in this way would have the potential to reduce false-positives in our relevance prediction task. Note that scaled term frequency differs from Robertson's (2004) *inverse total term frequency* in the sense that our method involves no consideration of term position within a document. Scaled query biased term frequency, shown in Figure 4, is defined as:

$$sfQ_{t,d,q} = \begin{cases} 2tfS_{t,d} \times idf_t, & \text{if } t \in q \\ tfS_{t,d} \times idf_t, & \text{otherwise} \end{cases}$$

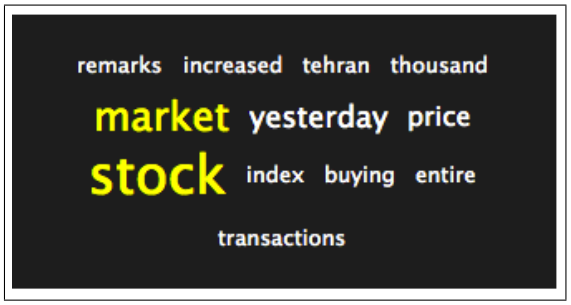

Figure 4: Word cloud summary for scaled query biased term frequency (SFQ) for query "Tehran's stock market".

Word Relevance (W) We adapted an existing relevance weighting from Allan et al., (2003), that was originally formulated for ranking sentences with respect to a query. However, we modified their originaly ranking method so that we could rank individual terms in a document instead of sentences. Our method for word relevance, W is defined as:

$$W_{t,d,q} = log(tf_{t,d} + 1) \times log(tf_{t,q} + 1) \times idf_t$$

In W, term frequency values are *smoothed* by adding 1. The smoothing could especially affect rare terms and singletons, when $tf_{t,d}$ is very low. All terms in a query or a document will be weighted and each term could potentially contribute to summary.

4.4 Query Biased Sentence Summaries

Sentences are a canonical unit to use in extractive summaries. In this section we describe four different sentence scoring methods that we used. These methods show how to calculate sentence scores for a given document with respect to a given query. Sentences for a document were always ranked using the raw score value output generated from a scoring method. Each document summary contained the top 3 ranked sentences where the sentences were simply listed out. Each of these methods used sentence-aligned English machine translated documents, and two of them also used the original Farsi text.

Sentence Relevance (REL) Our sentence relevance scoring method comes from Allan et al. (2003). The sentence weight is a summation over words that appear in the query. We provide their sentence scoring formula here. This calculates the relevance score for a sentence s from document d, to a query q:

$$rel_{(s|q)} = \sum_{t \in s} log(tf_{t,s}+1) \times log(tf_{t,q}+1) \times idf_t$$

Terms can occur in the sentence or query, or both. We applied this method to machine tranlsated English text. The output of this method is a relevance score for each sentence in a given document. We used those scores to rank sentences from our English machine translated text.

Query Biased Lexrank (LQ) We implemented query biased LexRank, a well-known graph-based summarization method (Otterbacher et al., 2009). It is a modified version of the original LexRank algorithm (Erkan and Radev, 2004; Page et al., 1998). The similarity metric, $sim_{x,y}$, also known as *idf-modified cosine similarity*, measures the distance between two sentences x and y in a document d, defined as:

$$sim_{x,y} = \frac{\sum_{t \in x,y} tf_{t,x} \times tf_{t,y} \times (idf_t)^2}{\sqrt{\sum_{t \in x} tfidf_{t,x}^2} \sqrt{\sum_{t \in y} tfidf_{t,y}^2}}$$

We used $sim_{x,y}$ to score the similarity of sentence-to-sentence, resulting in a similarity graph where each vertex was a sentence and each edge was the cosine similarity between sentences. We normalized the cosine matrix with a similarity threshold ($t = 0.05$), so that sentences above this threshold were given similarity 1, and 0 otherwise. We used $rel_{(s|q)}$ to score sentence-to-query. The LexRank score for each sentence was then calculated as:

$$LQ_{s|q} = \frac{d \times rel_{s|q}}{\sum_{z \in C} rel_{z|q}} + (1-d) \times \sum_{v \in adj[s]} \frac{sim_{s,v}}{\sum_{r \in adj[v]} sim_{v,r}} LQ_{v|q}$$

where C is the set of all sentences in a given document. Here the parameter d is just a damper to designate a probability of randomly jumping to one of the sentences in the graph ($d = 0.7$). We found the stationary distribution by applying the power method ($\epsilon = 5$), which is guaranteed to converge to a stationary distribution (Otterbacher et al., 2009). The output of LQ is a score for each sentence from a given document with respect to a query. We used that score to rank sentences from our English machine translated text.

Projected Cross-Language Query Biased Lexrank (LQP) We introduce LQP to describe a way of scoring and ranking sentences such that the L_1 (English) summaries are biased from the query and source document both in L_2 (Farsi). Our gold-standard Farsi queries came with the CLEF 2008 data, and they are therefore might be more reliable than what we could get from automatic translation. First, each of the Farsi sentences in a given document were scored and ranked using Farsi queries with LQ, described above. Then each LQ score was projected onto sentence-aligned English. By doing this, we simulated a user's English query translated into Farsi with the best possible query translation, before proceeding with summarization. We think this method could be of interest for CLIR systems that do query translation on-the-fly. It is also of interest for summarization systems that need to utilize previously translated source documents without the ability to translate the summaries from L_2 to L_1.

Combinatory Query Biased Lexrank (LQC) Anther twist on query biased LexRank that we introduce here is LQC, which combines LexRank scores from both languages. We did this by running LQ on Farsi and English separately, and adding the two scores together for each sentence. This combination of Farsi and English scores provided us with a different way to score and rank sentences, compared with LQ and LQP. The idea behind combinatory query biased LexRank is if a sentence has a high score in Farsi but not in English, then this should be reflected somehow. This method takes advantage of all available resources in our dataset: L_1 and L_2 queries as well as L_1 and L_2 documents.

5 Experiments

We tested each of our summarization methods and overall strategies in a task-based evaluation framework based on relevance prediction. We decided to use Mechanical Turk for our experiments since it has been shown to be useful for evaluating NLP systems (Callison-Burch 2009; Gillick and Liu, 2010). We obtained human judgments for whether or not a document was considered relevant to a query, or information need. We measured the relevance judgements by precision/recall/F1, accuracy, and also time-on-task based on the average response time per Human Intelligence Task (HIT).

5.1 Mechanical Turk

In our experiment, we used terminology from CLEF 2008 to describe a query as an "information need" for all of our HITs. All of the Mechanical Turk workers were presented with the following for their individual HIT: instructions, an information need and one summary for a document. Workers were asked to indicate if the given summary for a document was relevant to the given information need (Hobson et al., 2007). Workers were not shown the original source documents. We paid workers $0.01 per HIT. We obtained 5 HITs for each information need and summary pair. Workers on Mechanical Turk were provided with the following instructions

> "Instructions: Each image below consists of a statement summarizing the information you are trying to find from a set of documents followed by a summary of one of the documents returned when you query the documents. Based on the summary, choose whether you think the document returned is relevant to the information need. NOTE: It may be difficult to distinguish whether the document is relevant as the text may be difficult to understand. Just use your best judgment."

6 Results and Analysis

We present our experiment results and provide some additional analysis. First, we report the results of our relevance prediction task, showing performance for individual summarization methods as well as performance for our overall strategies. Then we provide some analysis of our results from the monolingual question-biased shared-task for DUC 2005, and we compare to previous work.

6.1 Results for Individual Methods

Our results for individual summarization methods are shown in Table 1. Performance for individual methods is shown along with performance for overall summarization strategies. Results for our overall summarization strategies are based on the arithmetic mean of the corresponding individual methods. We measured precision, recall and F1 to give us a sense of our summaries might influence document retrieval in an actual CLIR system. We also measured accuracy and time-on-task. For these latter two metrics, we distinguish between

Table 1: Individual method results: precision/recall/F1, time-on-task, and accuracy. Note that results for time-on-task and accuracy scores are distinguished for relevant (R) and non-relevant (NR) documents.

	Precision, Recall, F1			Time-on-Task		Accuracy	
Summarization Strategy	**Prec.**	**Rec.**	**F1**	**R**	**NR**	**R**	**N**
Unbiased Full MT English	**0.653**	**0.636**	**0.644**	**219.5**	**77.6**	**0.696**	**0.712**
TF	0.615	0.777	0.686	33.5	34.6	0.840	0.508
IDF	0.537	0.470	0.501	84.7	45.8	0.444	0.700
TFIDF	0.647	0.710	0.677	33.2	38.2	0.772	0.656
Unbiased Word Clouds	**0.599**	**0.652**	**0.621**	**50.5**	**39.5**	**0.685**	**0.621**
TFQ	0.605	0.809	0.692	55.3	82.4	0.864	0.436
IDFQ	0.582	0.793	0.671	23.6	31.6	0.844	0.436
TFIDFQ	0.599	0.738	0.661	37.9	26.9	0.804	0.500
SFQ	0.591	0.813	0.685	55.7	49.4	0.876	0.504
W	0.611	0.738	0.669	28.2	28.9	0.840	0.564
Query Biased Word Clouds	**0.597**	**0.778**	**0.675**	**36.4**	**34.2**	**0.846**	**0.488**
REL	0.582	0.746	0.654	30.6	44.3	0.832	0.548
LQ	0.549	0.783	0.646	64.4	54.8	0.868	0.292
LQP	0.578	0.734	0.647	28.2	28.0	0.768	0.472
LQC	0.557	0.810	0.660	33.9	38.8	0.896	0.292
Query Biased Sentences	**0.566**	**0.768**	**0.651**	**39.2**	**41.5**	**0.841**	**0.401**

summaries that were relevant (R) and non-relevant (NR). For many of the summarization methods, workers were able to positively identify relevant documents.

From Table 1 we see that Full MT performed better on precision than all of the other methods and strategies, but we note that performance on precision was generally very low. This might be due to Mechanical Turk workers overgeneralizing by marking summaries as relevant when they were not. Some individual methods preserve our principle of *retrieval-relevance*, as indicated by the higher recall scores for SQF, LQEF, and TFQ. That is to say, these particular query biased summarization methods can be used to assist users with identifying more relevant documents. The accuracy on relevant documents addresses our principle of *query-salience*, and it is especially high for our query-biased methods: LQEF, SQF, LQ, and TFQ. The results also seem to fit our intuition that the summary in Figure 2 seems less relevant to the summaries shown in Figures 3 & 4 even though these are the same documents biased on the same query "Tehran stock market".

Overall, query biased word clouds outperform the other summarization strategies for 5 out of 7 metrics. This could be due to the fact that word clouds provide a very concise and overview of a document, which is one of the main goals for automatic summarization. Along these lines, word clouds are probably not subject to the effects of MT quality and we believe it is possible that MT quality could have had a negative impact on our query biased extracted sentence summaries, as well as our full MT English texts.

6.2 Analysis with DUC 2005

We analysed our summarization methods by comparing our methods with peers from the monolingual question-biased summarization shared-task for DUC 2005. Even though DUC 2005 is a monolingual task, we decided to use it as part of our analysis for two reasons: (1) to see how well we could do with query/question biasing while ignoring the variables introduced by MT and cross-language text, and (2) to make a comparison to previous work. Pingali et al., (2007) also used this the same DUC task to assess their cross-language query biased summarization system. With this task, we used our methods: TF, IDF, TFIDF, TFQ, IDFQ, TFIDFQ, SFQ, W, REL, LQ. All of our methods produce a ranked list of words, except for LQ and REL which produce a list of ranked sentences. Systems from the DUC 2005 question-biased summarization task were evaluated automatically against human gold-standard summaries

Table 2: Comparison of peer systems on DUC 2005 shared-task for monolingual question-biased summarization, f-scores from ROUGE-2 and ROUGE-SU4.

Peer ID	ROUGE-2	ROUGE-SU4
17	0.07170	0.12970
8	0.06960	0.12790
4	0.06850	0.12770
Tel-Eng-Sum	0.06048	0.12058
LQ	0.05124	0.09343
REL	0.04914	0.09081
TFIDFQ	0.00069	0.01703
TFIDF	0.00068	0.01695
SFQ	0.00063	0.01706
TFQ	0.00058	0.01699
TF	0.00055	0.01698
W	0.00042	0.01625
IDFQ	0.00033	0.01627
IDF	0.00014	-

Table 3: Top 3 system precision scores for ROUGE-2 and ROUGE-SU4.

Peer ID	ROUGE-2	ROUGE-SU4
LQ	**0.08272**	**0.15197**
REL	**0.0809**	**0.15049**
15	0.07249	0.13129

using ROUGE (Lin and Hovy, 2003) . Our results from the DUC 2005 shared-task are shown in Table 2, reported as ROUGE-2 and ROUGE-SU4 f-scores, as these two variations of ROUGE are the most helpful (Dang, 2005; Pingali et al., 2007).

Table 2 also includes scores for several top peer systems, as well as results for the Tel-Eng-Sum method from Pingali et al., (2007). While we have reported f-scores in our analysis, we also note that our implementations of LQ and REL outperform all of the DUC 2005 peer systems for precision, as shown in Table 3. We begin to see that the ROUGE scoring method is not able to describe nuanced differences between our summarization methods in the same way that a task-based evaluation does.

7 Discussion and Future Work

ROUGE alone cannot make fine distinctions between different cross-language query biased summarization algorithms. Instead, we can see that our evaluation framework does make more distinctions between our methods than ROUGE does. We want to be able to say that we can "do query biased summarization" just as well for monolingual and cross-language IR systems. At minimum, the variables of translation quality and when to translate do not factor into monolingual summarization. But we would need relevance prediction experiments using humans who know L_1 and others who know L_2. Unfortunately in our case, we were not able to find Farsi speakers on Mechanical Turk. Access to these speakers would have allowed us to do further analysis.

Our results on the relevance prediction task tell us that query biased summarization strategies help users identify relevant document faster and with better accuracy, and our findings support the findings of Tombros and Sanderson (1998). Another important finding is that now we can weigh tradeoffs so that different summarization methods could be used to optimize over different metrics. For example, if we want to optimize for retrieval-relevance we might select a summarization method that tends to have higher recall, such as scaled query biased term frequency (SFQ). Similarly, we could optimize over accuracy on relevant documents, and use Combinatory LexRank (LQC) with Farsi and English together.

If the NLP community wants to make strides in cross-language query biased summarization for CLIR then we need some standards. First, researchers need to be using a parallel dataset consisting of documents in L_1 and L_2 with queries in L_1 and L_2 along with an answer key specifying which documents are relevant to the queries. We would also need sets of human gold-standard query biased summaries in L_1 and L_2. Only then could we begin to compare system-to-system across languages while teasing apart the variables, such as when to translate, translation quality, methods for biasing, summarization strategy. And of course it would be better if this standard dataset was multilingual instead of billingual, for obvious reasons.

We have approached cross-language query biased summarization as a stand-alone problem, treating the CLIR system and document retrieval as a black box. However, summaries need to preserve query-salience: summaries should not make it more difficult to positively identify relavant documents. And they should also preserve retrieval-relevance: summaries should help users identify as many relevant documents as possible.

References

E. Agirre, G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. Clef 2008: Ad hoc track overview. In *Evaluating systems for multilingual and multimodal information access*, pages 15–37. Springer, 2009.

J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, (SIGIR '03). ACM, New York, NY, USA, 314-321.

H. Azarbonyad, A. Shakery, and H. Faili. Exploiting multiple translation resources for english-persian cross language information retrieval. In P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, pages 93–99. Springer Berlin Heidelberg, 2013.

L.L Bando, F. Scholer, A. Turpin. Constructing Query-biased Summaries: A Comparison of Human and System Generated Snippets. In *Information Interaction in Context*, ACM, 2010.

A. Berger and V. O. Mittal. Query-relevant summarization using FAQs. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL '00)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 294-301.

P. Bhaskar and S. Bandyopadhyay. Cross lingual query dependent snippet generation. *International Journal of Computer Science and Information Technology (IJCSIT)*, 3(4), 2012.

P. Bhaskar and S. Bandyopadhyay. Language independent query focused snippet generation. In T. Catarci, P. Forner, D. Hiemstra, A. Peñas, and G. Santucci, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, volume 7488 of *Lecture Notes in Computer Science*, pages 138–140. Springer Berlin Heidelberg, 2012.

S. Borgatti, K. M. Carley, D. Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, (28):124–136, 2006.

F. Boudin, S. Huet, and J.-M. Torres-Moreno. A graph-based approach to cross-language multi-document summarization. *Polibits*, (43):113–118, 2011.

C. Callison-Burch. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 286–295, Singapore, August, 2009. ACL.

Y. Chali and S.R. Joty. Unsupervised Approach for Selecting Sentences in Query-based Summarization. In *Proceedings of the Twenty-First International FLAIRS Conference*, 2008.

N. Chatterjee, A. Mittal, and S. Goyal. Single document extractive text summarization using genetic algorithms. In *Emerging Applications of Information Technology (EAIT), 2012 Third International Conference*, pages 19–23, 2012.

K. Church and W. Gale. Inverse document frequency (idf): A measure of deviations from poisson. In *Natural language processing using very large corpora*, pages 283–295. Springer, 1999.

M. Creutz and K. Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1):3:1–3:34, Feb. 2007.

H. T. Dang. Overview of DUC 2005. In =*Proceedings of the Document Understanding Conference*, 2005.

H. Daumé III, D. Marcu. Bayesian Query-Focused Summarization In *Proceedings IWSLT 2012*, 2012.

J. Drexler, W. Shen, T. Gleason, T. Anderson, R. Slyh, B. Ore, and E. Hansen. The mit-ll/afrl iwslt-2012 mt system. In *Proceedings of 21st Annual Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 305–312, Sydney, Australia, July 2006.

H. P. Edmundson. New methods in automatic extracting. *J. ACM*, 16(2):264–285, Apr. 1969.

G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, Dec. 2004.

G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479, Dec. 2004.

D. Gillick and Y. Liu. Non-Expert Evaluation of Summarization Systems is Risky. In *Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148-151, Los Angeles, California, USA, June, 2010.

D. Harwath and T.J. Hazen. Topic identification based extrinsic evaluation of summarization techniques applied to conversational speech. In *Proceedings of ICASSP*, 2012: 5073-5076.

S.P. Hobson, B.J. Dorr, C. Monz, R. Schwartz. Task-based evaluation of text summarization using Relevance Prediction. In *Information Processing Management*, 43(6): 1482-1499, 2007.

H. Jing, R. Barzilay, K. McKeown, and M. Elhad. Summarization Evaluation Methods: Experiments and Analysis. In *Proceedings of AAAI*, 1998.

R. Karimpour, A. Ghorbani, A. Pishdad, M. Mohtarami, A. AleAhmad, H. Amiri, and F. Oroumchian. Improving persian information retrieval systems using stemming and part of speech tagging.

In *Proceedings of the 9th Cross-language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access*, CLEF'08, pages 89–96, Berlin, Heidelberg, 2009. Springer-Verlag.

C-Y. Lin. Looking For A Few Good Metrics: Automatic Summarization Evaluation - How Many Samples Are Enough? In *Proceedings of NTCIR Workshop 4*, June, 2004.

A. Louis and A. Nenkova. Automatic Summary Evaluation without Human Models. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, 2009.

H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, Apr. 1958.

I. Mani, E. Bloedorn, and B. Gates. Using Cohesion and Coherence Models for Text Summarization. In *AAAI Symposium Technical Report SS-989-06*, AAAI Press, 69–76, 1998.

I. Mani, D. House, G. Klein, L. Hirschman,T. Firmin, and B. Sundheim. The TIPSTER SUMMAC Text Summarization Evaluation. In *Proceedings of European Association for Coputational Linguistics*, (EACL), 1999.

I. Mani. Summarization Evaluation: An Overview. 2001.

T. Mori, M. Nozawa, and Y. Asada. Multi-answer focused multi-document summarization using a question-answering engine. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

A. Nenkova and K. McKeown. A survey of text summarization techniques. In C. C. Aggarwal and C. Zhai, editors, *Mining Text Data*, pages 43–76. Springer US, 2012.

C. Orăsan and O. A. Chiorean. Evaluation of a cross-lingual romanian-english multi-document summariser. In *Proceedings of Language Resources and Evaluation Conference*, LREC'08, 2008.

J. Otterbacher, G. Erkan, and D.R. Ravev. Using Random Walks for Question-focused Sentence Retrieval. In *Proceedings of Human Language Technology Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, Canada, pp 915-922, (2005).

J. Otterbacher, G. Erkan, and D.R. Ravev. Biased LexRank: Passage retrieval using random walks with question-based priors. In *Information Processing Management*, 45(1), January 2009, pp 42-54.

K. Owczarzak, J.M Conroy, H.T. Dang, and A. Nenkova. An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of the Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1-9, Montréal, Canada, June 2012. ACL.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

P. Pingali, J. Jagarlamudi, and V. Varma. Experiments in cross language query focused multi-document summarization Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies in IJCAI2007.

D. R. Radev, H. Jing, M. Styś, and D. Tam. Centroid-based summarization of multiple documents. In*Proceedings of Informaion Processing Management*, 40(6):919–938, Nov. 2004.

S.E. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241. Springer-Verlag New York, Inc., 1994.

G. Salton and C.S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372, 1973.

A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–10. ACM, 1998.

X. Wan and J. Xiao. Graph-Based Multi-Modality Learning for Topic-Focused Multi-Document Summarization. In *Proceedings of the 21st international jont conference on Artifical intelligence* (IJCAI'09), Hiroaki Kitano (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1586-1591.

X. Wan, H. Li, and J. Xiao. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 917-926.

X. Wan, H. Jia, S. Huang, and J. Xiao. Summarizing the differences in multilingual news. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 735–744, New York, NY, USA, 2011. ACM.

W. Yin, Y. Pei, F. Zhang, and L. Huang. SentTopicMultiRank: a novel ranking model for multi-document summarization. In *Proceedings of COLING*, pages 2977–2992, 2012.

J. Zhang, L. Sun, and J. Min. Using the web corpus to translate the queries in cross-lingual information retrieval. In *Proceedings in 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2005*, IEEE NLP-KE '05, pages 493–498, 2005.