Using Interactions to Improve Translation Dictionaries: UNC, Yahoo! and ciQA

Diane Kelly[1], Vanessa Murdock[2], & Xin Fu[1]

[1]School of Information and Library Science
University of North Carolina
Chapel Hill, NC 27599-3360
[dianek | fu@email.unc.edu]

[2]Yahoo! Research Barcelona
Ocata 1
08003 Barcelona, Spain
vmurdock@yahoo-inc.com

Abstract

Sentence retrieval is an important step in many question-answering (QA) technologies. However, characteristics of sentences and of the question-answering task itself often make it difficult to apply document retrieval techniques to sentence retrieval. The use of translation dictionaries offers one potentially useful approach to sentence retrieval, but training such dictionaries using QA corpora often introduces noise that can negatively impact retrieval performance. In this study, we experiment with using data elicited from assessors during interactions as training data for a translation dictionary. We employ two different interactions that elicit two types of data: data about assessors' topics and data about retrieved sentences. Results show that using sentence-level relevance feedback to adjust the translation dictionary improved retrieval for about half the topics, but harmed it for the other half.

1 Introduction

Many question answering systems employ sentence or passage retrieval as a first step to finding answers to questions. This is a critical first step and if retrieval results are not good, then it is unlikely that useful answers will be found. Sentence retrieval is challenging for several reasons. Sentence retrieval is especially sensitive to vocabulary mismatch between question terms and sentence terms because sentences have so few terms to begin with, and most sentences are unlikely to contain both a word and its synonym. Furthermore, because the question usually does not contain the answer, the application of traditional document retrieval approaches to sentence retrieval is often ineffective. Such approaches may retrieve topically related sentences, but not sentences that contain actual answers.

To address these problems we propose the use of a translation dictionary to associate a sentence to a question containing related words. A translation dictionary learns related words from a parallel corpus of answered questions. Ideally, questions would be associated with answer terms in a translation table, which should improve the precision of the retrieval system. However, a parallel corpus of answered questions can produce a very noisy translation dictionary. In this study, we investigate the potential effectiveness of using data generated by users as an additional source of evidence for improving the quality of the translation dictionary. The data generated by users is done so through two different types of interactions: one that probes users' about their information problems and another that elicits sentence-level relevance feedback.

Previous research has shown that techniques employed by reference librarians to discover more information about users' information needs are also useful for eliciting additional information from users of online information systems. It has further been shown that this information can be used to improve retrieval performance for a document retrieval task (Kelly, Dollu and Fu, 2005). As part of our participation in the TREC ciQA task we explored the effectiveness of an interaction technique based on the standard reference interview for the complex, question-answering task where sentences were being retrieved rather than documents. Specifically, we propose to use the feedback from this interaction technique to refine the translation dictionary.

In the area of information retrieval, it is well documented that relevance feedback is an effective technique for improving retrieval performance (c.f., Ruthven and Llamas, 2003). However, in last year's ciQA task (Kelly and Lin, 2007) traditional relevance feedback approaches were not particularly effective for question-answering tasks. We suspect that one reason for this is because techniques that work well for document retrieval do not necessarily work well for sentence retrieval. Specific reasons for this were noted above: the item that is being retrieved is much smaller than a document and the question and the answer may share few (if any) of the same words. In this study, we were interested in investigating an alternative application of data generated through relevance feedback by applying it to the development of a translation dictionary instead of using it in a traditional way (i.e., for query expansion).

Although there are questions about whether real users would actually provide relevance feedback in operational settings, we thought we would take advantage of this experimental situation where assessors are asked to provide feedback, to investigate whether this is a potentially effective way to use relevance feedback for sentence retrieval. We use the relevance feedback to modify the translation dictionary, both by adding vocabulary to the dictionary from the relevant sentences, and by improving the translation probabilities for terms related to relevant sentences.

2 Interaction Techniques

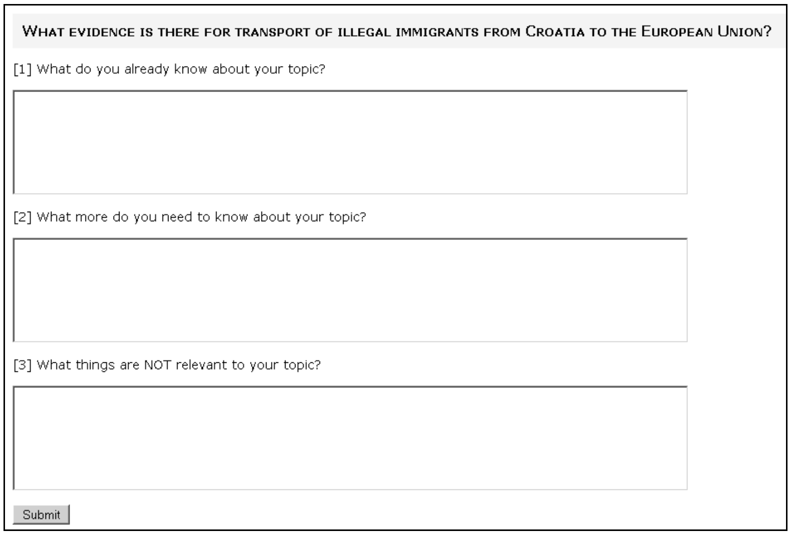
We studied two types of interactions both of which were embodied in static forms; thus, our approach to interaction was similar to that studied in previous versions of the ciQA task and in the HARD Track in that it was a single-cycle interaction. That is, the user presents an information need and the system asks the user to respond to a set of questions. The feedback provided by the user is then used to produce new results. We experimented with two types of interactions: open and closed. In the open interaction, assessors responded to a predefined set of questions that were designed to elicit more information from them about their questions. In the closed interaction, assessors provided relevance feedback about sets of sentences.

Eight assessors participated in this study. Each assessor was matched with 3-4 questions. Thus, assessors completed our interactions several times – once for each of their questions.

2.1 Open Interaction Technique

In the first interaction, we asked assessors to respond to a set of open questions which has been demonstrated in previous research to be effective at eliciting additional information from users beyond a simple query (Kelly, Dollu & Fu, 2005). These questions were inspired by questions typically asked by reference librarians when attempting to get a better understanding of a user's information need. Questioning techniques such as these are commonly employed by reference librarians and are a standard part of reference instruction (Katz, 2002).

The format of the interaction is displayed in Figure 1. The first question asked assessors to describe what they already know about their topics, the second question asked them to describe what else they need to know about their topics, and the third question asked them to describe things that are not relevant to their topic. Of course, responses to this last question could be numerous, but in the context of this study, it was expected that assessors would interpret this question in relation to their information problem. We asked this last question in hopes of getting some information that would allow us to experiment with negative feedback.



What evidence is there for transport of illegal immigrants from Croatia to the European Union?

[1] What do you already know about your topic?

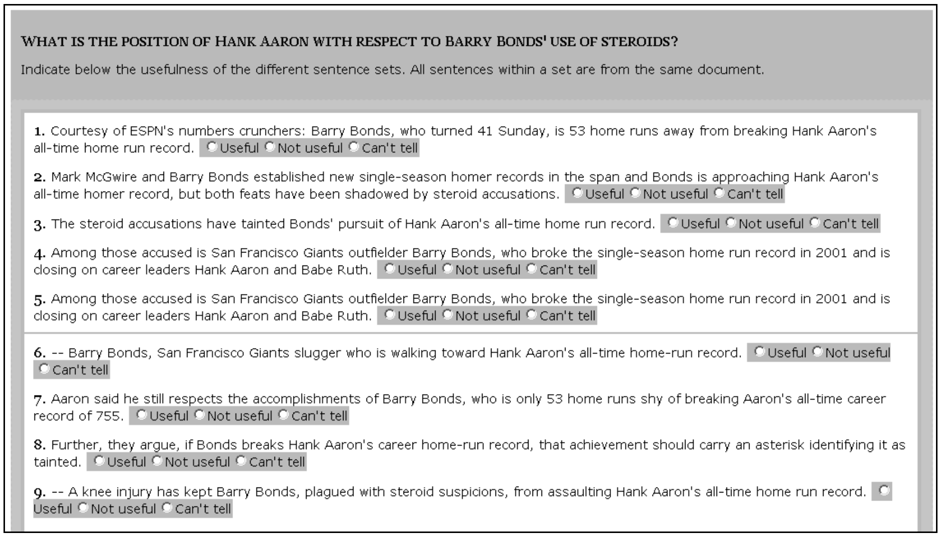
[2] What more do you need to know about your topic?

[3] What things are NOT relevant to your topic?

Figure 1. Open Interaction Technique

2.2 Closed Interaction Technique

Our second interaction technique consisted of sentence-level feedback (Figure 2). Assessors were presented with a list of 20 top-ranking sentences and asked to indicate whether sentences were useful or not useful, or if they couldn't tell. Our original intention was to present a limited number of sentences from a varying number of documents. Because some documents contain many relevant sentences, there is a possibility that top ranking sentences will be from the same document. We hoped our presentation method would potentially provide assessors with more diverse sentence sets from a larger number of documents. Unfortunately, we were not able to organize the sentences in this way and did not change the interface to reflect this.



WHAT IS THE POSITION OF HANK AARON WITH RESPECT TO BARRY BONDS' USE OF STEROIDS?

Indicate below the usefulness of the different sentence sets. All sentences within a set are from the same document.

1. Courtesy of ESPN's numbers crunchers: Barry Bonds, who turned 41 Sunday, is 53 home runs away from breaking Hank Aaron's all-time home run record. ○ Useful ○ Not useful ○ Can't tell

2. Mark McGwire and Barry Bonds established new single-season homer records in the span and Bonds is approaching Hank Aaron's all-time homer record, but both feats have been shadowed by steroid accusations. ○ Useful ○ Not useful ○ Can't tell

3. The steroid accusations have tainted Bonds' pursuit of Hank Aaron's all-time home run record. ○ Useful ○ Not useful ○ Can't tell

4. Among those accused is San Francisco Giants outfielder Barry Bonds, who broke the single-season home run record in 2001 and is closing on career leaders Hank Aaron and Babe Ruth. ○ Useful ○ Not useful ○ Can't tell

5. Among those accused is San Francisco Giants outfielder Barry Bonds, who broke the single-season home run record in 2001 and is closing on career leaders Hank Aaron and Babe Ruth. ○ Useful ○ Not useful ○ Can't tell

6. -- Barry Bonds, San Francisco Giants slugger who is walking toward Hank Aaron's all-time home-run record. ○ Useful ○ Not useful ○ Can't tell

7. Aaron said he still respects the accomplishments of Barry Bonds, who is only 53 home runs shy of breaking Aaron's all-time career record of 755. ○ Useful ○ Not useful ○ Can't tell

8. Further, they argue, if Bonds breaks Hank Aaron's career home-run record, that achievement should carry an asterisk identifying it as tainted. ○ Useful ○ Not useful ○ Can't tell

9. -- A knee injury has kept Barry Bonds, plagued with steroid suspicions, from assaulting Hank Aaron's all-time home run record. ○ Useful ○ Not useful ○ Can't tell

Figure 2. Closed Interaction Technique

3 Retrieval Techniques

We used the Lemur IR toolkit (http://www.lemurproject.org) to conduct our experiments. We did the retrieval in a two-step process: a document retrieval step, followed by a sentence retrieval step.

3.1 Document Retrieval

We built the document index using Lemur's basic defaults for indexing (BuildIndex function), with the Krovetz stemmer (Krovetz, 1993). Our baseline queries consisted of the template and narrative for each topic. Stopwords from the Inquery stoplist were removed from the text, and the default parameters for Jelinek-Mercer smoothing were used. We retrieved the top 1000 documents, sentence-segmented them with a rule-based segmenter, and indexed the sentences for each query separately.

Unfortunately, only documents from one source were indexed. We were not aware of this until close to the deadline for submitting the baseline results and despite our trouble-shooting efforts, we were unable to identify and fix the problem. Therefore, our official TREC baseline results (UNCYABL30) were only from AFP (> 99%) and APW sources (< 1%)

After the deadline passed, we gave up on the latest version of Lemur, reinstalled an earlier version (4.3.2) and re-indexed the corpus without any problem. This seems to indicate that the latest version of Lemur has some problem indexing this corpus, but we were unable to identify the problem. The new retrieval run, UNCYAEX1, was later submitted as one of our experimental runs, but will be analyzed hereafter as a baseline. It is important to note that the sentences we used to populate our closed interaction form were from UNCYAEX1 and not from our official baseline.

3.2 Sentence Retrieval

We used *Model S* for sentence retrieval. Model S is described in Murdock (2006) and uses translation probabilities learned from parallel corpora of answered questions to weight the language modeling probability in a query-likelihood retrieval framework. In addition, as the context of a sentence is a good indicator of its relevance, the sentences were smoothed from the documents they came from, as described in Murdock and Croft (2005). Model S used Jelinek-Mercer smoothing with the document language model, giving as little probability as possible to the corpus language model. Sentences were stemmed using the Krovetz stemmer, but only single-characters were removed as stopwords.

Model S relies on a translation table, which is trained from a parallel corpus of aligned sentences. In our case, we combined dictionaries from two different sources. The first source consisted of factoid questions and their known answer sentences from the TREC QA Track, questions 1 through 1893. The second source consisted of a parallel corpus of ciQA task questions, and their answer nuggets from the 2006 TREC ciQA task (Kelly and Lin, 2007). Both translation dictionaries were learned using GIZA++ (Al-Onaizan et al. 1999), an open-source implementation of the IBM translation models (Brown et al. 1990). The translation tables were learned using IBM Model 4, with the default parameter settings. Prior to learning the translation tables, sentences and questions were stemmed using the Krovetz stemmer, but only single-character stopwords were removed.

The two dictionaries were combined such that if a term appeared in both dictionaries, its probabilities were averaged otherwise the terms were simply added to the dictionary. Both runs UNCYABL30 and UNCYAEX1 were produced using this model, the difference in performance accounted for entirely by the faulty indexing of the baseline run. For our experimental run (UNCYAEX2), we used the sentences assessors indicated as useful or somewhat useful paired with the questions for which the sentences were retrieved to create a parallel corpus. The sentences were processed as described above. That is, single characters were removed and the sentences were stemmed using the Krovetz stemmer. A translation table was learned using GIZA++ as described above, and the dictionary was merged with the existing dictionary.

4 Results

4.1 Interaction form responses

We obtained quite a bit of feedback from assessors from our open-ended interaction. This feedback was illuminating in several ways. While the majority of it appears useful (we have not analyzed it thoroughly because we have not decided on a method for handling some of the data), some assessors' replies to these questions were unexpected and very different from the type of responses we received in our previous studies of this technique. For instance, at least one assessor provided what we consider spam – the responses entered were obviously and purposely not about the assessor's topic. In response to the first question asking assessors what more they'd like to know about their topics, several assessors responded by saying that they already knew everything about their topics and that providing a response to this question wouldn't be fair. Clearly, our understanding of how much assessors know about their topics *a priori* was not accurate. To be fair, we used this technique in a previous TREC study (Kelly, Dollu, and Fu, 2005) and found it to be effective. However, we note that the assessors in this previous Track were recruited by the Linguistic Data Consortium and were engaged in a document retrieval task. It may be the case that assessors for QA tasks necessarily have to know more about their topics, in which case the open interaction (or any interaction for that matter) may not have been a good fit. We'd like to emphasize that these types of responses were in the minority, but they definitely made us think about the potential limitations of the TREC framework for studying interactions of this type. Overall, some data was elicited by this technique; however, we have yet to determine whether it is useful. Furthermore, we need to identify a technique for automatically removing responses that were written to us as researchers (such as those described above), rather than in response to *the system's* questions and for dealing with spam.

With respect to sentence feedback, about 45% (274) of the 600 sentences shown to all assessors (30 topics * 20 sentences) were marked as useful, 36% (215) were marked as not useful, 16% (99) were marked as can't tell, and 3% (18) were not marked.

4.2 Retrieval results

Table 1 shows the mean F-scores and standard deviations for our official baseline run (UNCYACL30), unofficial baseline run (UNCYAEX1) and experimental run (UNCYAEX2). As expected, our official baseline run performed very poorly. However, our unofficial baseline run (which was the same technique employed in our unofficial run, expect with the entire corpus indexed) was on par with the median average baseline run score of 0.359.

Table 1. Means and standard deviations for F-scores of different runs

	Mean	Std. Deviation
Official Baseline (UNCYABL30)	0.062	0.114
Unofficial Baseline (UNCYAEX1)	0.355	0.177
Experimental (UNCYAEX2)	0.374	0.175

We conducted a paired sample t-test to investigate whether the difference between UNCYAEX1 and UNCYAEX2 was statistically significant. Results of this test indicated that UNCYAEX2 did not significantly improve performance over UNCYAEX1 [$\underline{t}(29)=0.892$, $\underline{p}=0.380$]. Table 2 shows the comparison of UNCYAEX2 to the unofficial baseline run (UNCYAEX1) at the topic level. UNCYAEX2 improved the performance of about 53% of the topics, but harmed it for 40% of the topics. We note that results remained unchanged for very few topics.

Table 2. Number of topics for which the experimental run (UNCYAEX2) outperformed the unofficial baseline (UNCYAEX1)

	Better	Same	Worse
UNCYAEX2	16	2	12

4.3 Cross site comparison

In this subsection, we report a comparison between our results and results of other sites. Figures 3 and 4 show the performance of our official and unofficial baseline runs relative to the best, median and worst baselines from other sites. The best and worst scores appear at the top and bottom of each line, respectively, while the median score for each topic is set to 0.000. As expected, our official baseline run was the worst performing run for most topics. Our unofficial baseline run fared better (Figure 4) and even produced the best run for a couple of topics.

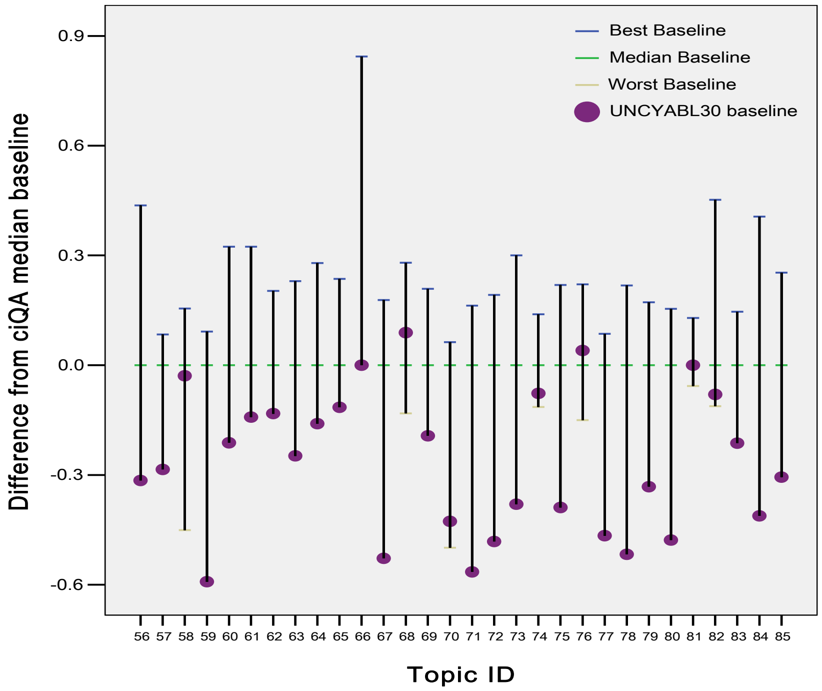


Figure 3. Comparison of official baseline (UNCYABL30) to ciQA07 baselines

Figure 5 shows the performance of our experimental run (UNCYAEX2) relative to the best, median and worst final runs from other sites. For 19 topics (more than 60%), the performance of UNCYAEX2 was equal to or above the ciQA07 medians. The average F-score of UNCYAEX2 was 0.374, which was also higher than the average final run performance of all sites 0.351, but the difference was not significant [$t(29)=0.704$, $p=0.487$].

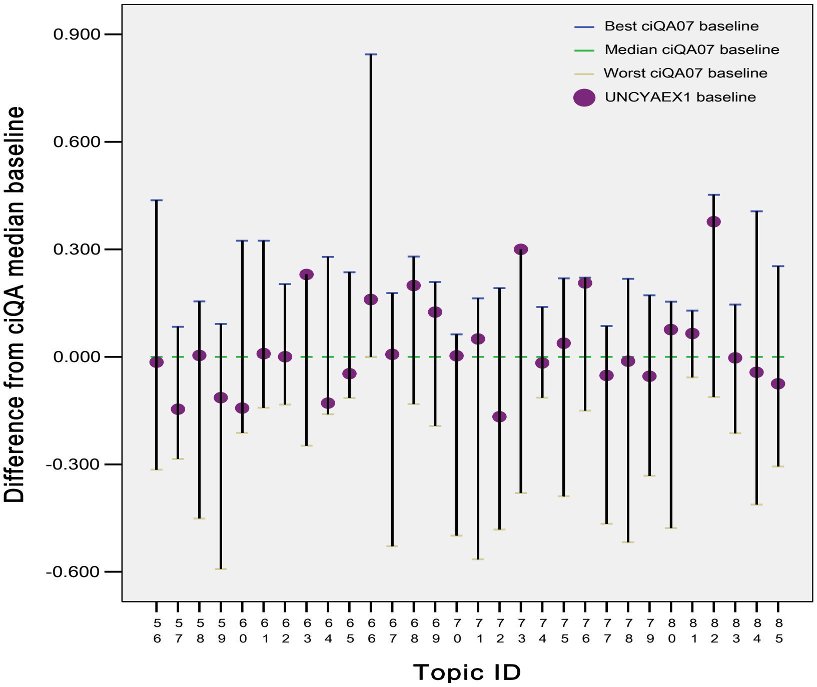


Figure 4. Comparison of unofficial baseline (UNCYAEX1) to ciQA07 baselines

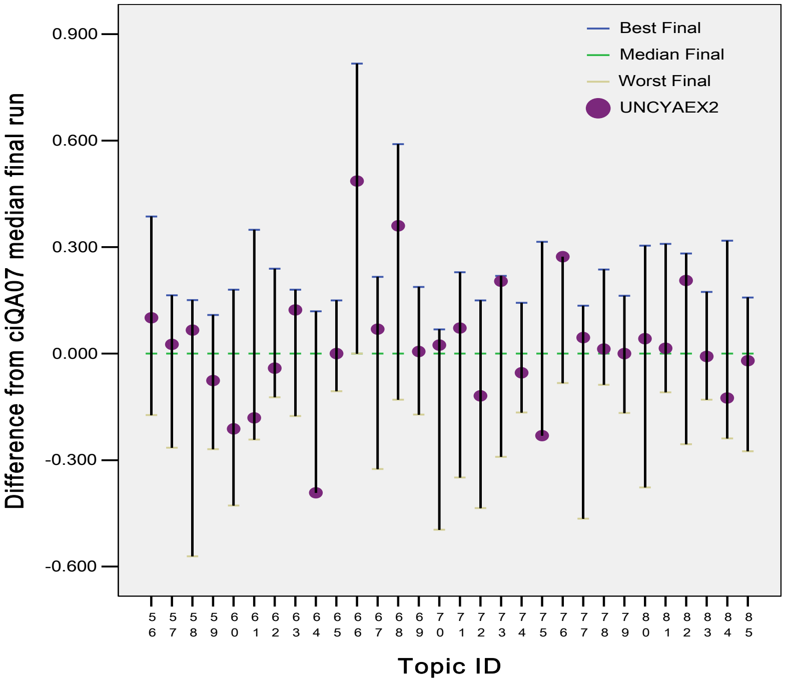


Figure 5. Comparison of UNCYAEX2 to ciQA07 final runs

5 Conclusions

In this study, we used sentence-level relevance feedback to modify a translation dictionary, both by adding vocabulary to the dictionary from the relevant sentences, and by adjusting the translation probabilities for terms related to relevant sentences. Our initial expectation was that the quality of the translation dictionary would improve slightly as term pairs not previously in the dictionary were added, and terms that existed already were adjusted by the translation probabilities from the feedback dictionary. We did not expect a drastic improvement because the original terms still existed in the dictionary. However, we showed that with even a small amount of feedback data, we can improve the quality of the translation table sufficiently to improve retrieval for about 50% of the topics. That the performance improved at all suggests that a better dictionary can improve performance, and that constructing this dictionary with feedback from users is a potentially useful way to construct a high-quality dictionary of related terms.

We experienced numerous problems in conducting this study, including problems with our initial baseline run. The experiment did not quite turn out as we had hoped, but we learned a lot about QA assessors and about the types of interactions that are or are not possible in this experimental setting. Despite not finding any substantial results, we believe there is some potential value in using user-generated data as evidence for improving the quality of translation dictionaries. We plan to continue to investigate how the data generated in this experiment can be used to improve sentence retrieval and the retrieval of answers for complex questions, and to investigate how we might collect data from other types of users for similar tasks.

6 References

Brown, Peter F., Cocke, John, Della Pietra, Stephen, Della Pietra, Vincent J., Jelinek, Fredrick, Lafferty, John D., Mercer, Robert L, and Roossin, Paul S. (1990). “A Statistical Approach to Machine Translation” in *Computational Linguistics*, vol. 16, no. 2, pages 79-85.

Katz, William A. (2002). *Introduction to reference work: reference services and reference processes, volume 2 (8th Edition).* NY: McGraw-Hill.

Kelly, Diane and Lin, Jimmy (2007). Overview of the TREC 2006 ciQA Task. *SIGIR Forum, 41*(1), 107-116.

Kelly, Diane, Dollu, Vijay. J., & Fu, Xin (2005). The loquacious user: A document-independent source of terms for query expansion. In *Proceedings of the 28th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brazil, 457-464.

Krovetz, Robert (1993). “Viewing Morphology as an Inference Process” in *Proceedings of the 16th Interational Conference on Research and Development in Information Retrieval*.

Murdock, Vanessa (2006). *Aspects of Sentence Retrieval*. PhD Thesis, University of Massachusetts.

Murdock, Vanessa and Croft, W. Bruce (2005). “A Translation Model for Sentence Retrieval” in *Proceedings of the Conference on Human Language Technologies and Empirical Methods in Natural Language Processing.*

Ruthven, Ian and Lalmas, Mounia (2003). “A survey on the use of relevance feedback for information access systems” in *Knowledge Engineering Review, 18*(2), 95-145.

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith and David Yarowsky (1999). “Statistical Machine Translation, Final Report, JHU Workshop.” Baltimore, Maryland.