A Translation Model for Sentence Retrieval

Vanessa Murdock and **W. Bruce Croft**
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, MA 01003
{vanessa,croft}@cs.umass.edu

Abstract

In this work we propose a translation model for monolingual sentence retrieval. We propose four methods for constructing a parallel corpus. Of the four methods proposed, a lexicon learned from a bilingual Arabic-English corpus aligned at the sentence level performs best, significantly improving results over the query likelihood baseline. Further, we demonstrate that smoothing from the local context of the sentence improves retrieval over the query likelihood baseline.

1 Introduction

Sentence retrieval is the task of retrieving a relevant sentence in response to a user's query. Tasks such as question answering, novelty detection and summarization often incorporate a sentence retrieval module. In previous work we examined sentence retrieval for question answering (Murdock and Croft, 2004). This involves the comparison of two well-formed sentences, one a question, one a statement. In this work we compare well-formed sentences to queries, which can be typical keyword queries of 1 to 3 terms, or a set of sentences or sentence fragments. The TREC Novelty Track provides this type of data in the form of topic titles and descriptions, and sentence-level relevance judgments for a small subset of the collection.

We present a translation model specifically for monolingual data, and show that it significantly improves sentence retrieval over query-likelihood. Translation models train on a parallel corpus and in previous work we used a corpus of question/answer pairs. No such corpus is available for the novelty data, so in this paper we present four ways to construct a parallel corpus, to estimate a translation model.

Many systems treat sentence retrieval as a type of document or passage retrieval. In our data a sentence is an average of 18 words, most of which occur once. A document is an average of 700 words, many of which are multiples of the same term. It is much less likely for a word and its synonym terms to appear in the same sentence than in the same document.

Passages may be any length, either fixed or variable, but are somewhat arbitrarily designated. Many systems that have a passage retrieval module, on closer inspection have defined the passage to be a sentence. What is needed is a sentence retrieval mechanism that retains the benefits of passage retrieval, where a passage is longer than a sentence. We propose that smoothing from the local context of the sentence improves retrieval over the query likelihood baseline, and the larger the context, the greater the improvement.

We describe our translation model in section 2, along with our smoothing approach. In section 3 we discuss previous work in sentence retrieval for the Novelty task, and translation models for information retrieval tasks. Section 4 presents four ways to estimate a translation model, in the absence of a parallel corpus, and presents our experimental results. We

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 684–691, Vancouver, October 2005.

discuss the results in section 5, and present our conclusions and future work in section 6.

2 Methodology

Our data was provided by NIST, as part of the TREC Novelty Track[1]. The documents for the TREC Novelty Track in 2002 were taken from the TREC volumes 4 and 5, and consist of news articles from the Financial Times, the Foreign Broadcast Information Service, and the Los Angeles Times from non-overlapping years. In 2003 and 2004, the documents were taken from the Aquaint Corpus, which is distributed by the Linguistic Data Consortium[2] and consists of newswire text in English from the Xinhua News Service, the New York Times, and the Associated Press from overlapping years.

We retrieved the top 1000 documents for each topic from the TREC and Aquaint collections, and sentence segmented the documents using MXTerminator (Reynar and Ratnaparkhi, 1997), which is a freely available sentence boundary detector. Each topic was indexed separately and had an average of 30,000 sentences. It was impractical to do sentence-level relevance assessments for the complete set of 150,000 documents, so we used the relevance assessments provided as part of the Novelty task, recognizing that the results are a lower bound on performance, because the relevance assessments do not cover the collection. The relevance assessments cover 25 known relevant documents for each topic.

We evaluated precision at N documents because many systems using sentence retrieval emphasize the results at the top of the ranked list, and are less concerned with the overall quality of the list.

2.1 Translation Models

We incorporated a machine translation model in two steps: estimation and ranking. In the estimation step, the probability that a term in the sentence "translates" to a term in the query is estimated using the implementation of IBM Model 1 (Brown et al., 1990) in GIZA++ (Al-Onaizan et al., 1999) out-of-the-box without alteration. In the ranking step we incorporate the translation probabilities into the query-likelihood framework.

In Berger and Lafferty (1999), the IBM Model 1 is incorporated thus:

$$P(q_i|S) = \sum_{j=1}^{m} P(q_i|s_j)P(s_j|S) \qquad (1)$$

where $P(q_i|s_j)$ is the probability that term s_j in the sentence translates to term q_i in the query. If the translation probabilities are modified such that $P(q_i|s_j) = 1$ if $q_i = s_j$ and 0 otherwise, this is Berger and Lafferty's "Model 0", and it is exactly the query-likelihood model (described in section 2.2).

A major difference between machine translation and sentence retrieval is that machine translation assumes there is little, if any, overlap in the vocabularies of the two languages. In sentence retrieval we depend heavily on the overlap between the two vocabularies. With the Berger and Lafferty formulation in equation 1, the probability of a word translating to itself is estimated as a fraction of the probability of the word translating to all other words. Because the probabilities must sum to one, if there are any other translations for a given word, its self-translation probability will be less than 1.0. To accommodate this monolingual condition, we make the following improvement.

Let $t_i = 1$ if there exists a term in the sentence s_j such that $q_i = s_j$, and 0 otherwise:

$$\sum_{1 \le j \le n} p(q_i|s_j)p(s_j|S) \Longrightarrow$$
$$t_i p(q_i|S) + (1 - t_i) \sum_{1 \le j \le n, s_j \neq q_i} p(q_i|s_j)p(s_j|S) \qquad (2)$$

The translation probabilities still sum to one. We determined empirically that this adjustment improved the results over IBM model 1, and over Berger and Lafferty model 0.

[1] http://trec.nist.gov
[2] http://www.ldc.upenn.edu

2.2 Document Smoothing

Query likelihood is a generative model that assumes that the sentence is a sample of a multinomial distribution of terms. Sentences are ranked according to the probability they generate the query. We estimate this probability by interpolating the term distribution in the sentence with the term distribution in the collection:

$$P(Q|S) = P(S) \prod_{i=1}^{|Q|} \Big(\lambda P(q_i|S) + (1-\lambda) P(q_i|C) \Big) \quad (3)$$

where Q is the query, S is the sentence, $P(S)$ is the (uniform) prior probability of the sentence, $P(q_i|S)$ is the probability that term q_i in the query appears in the sentence, and $P(q_i|C)$ is the probability that q_i appears in the collection.

In the experiments with document smoothing, we estimate the probability of a sentence generating the query:

$$P(Q|S) = P(S) \prod_{i=1}^{|Q|} \Big(\alpha P(q_i|S) + \beta P(q_i|D_S) + \gamma P(q_i|C) \Big) \quad (4)$$

where $\alpha + \beta + \gamma = 1.0$ and $P(q_i|D_S)$ is the probability that the term q_i in the query appears in the document the sentence came from. In our case, since the sentences for each topic are indexed separately, the collection statistics are in reference to the documents in the individual topic index.

3 Previous Work

The TREC Novelty Track ran for three years, from 2002 to 2004. Overviews of the track can be found in (Harman, 2002), (Soboroff and Harman, 2003) and (Soboroff, 2004). A number of systems use traditional information retrieval techniques for sentence retrieval, using various techniques to compensate for the sparse term distributions in sentences. The University of Massachusetts (Larkey et al., 2002) and Carnegie Mellon University (Collins-Thompson et al., 2002) both ranked sentences by the cosine similarity of the sentence vector to the query vector of tf.idf-weighted terms. Amsterdam University (Monz et al., 2002) used tfc.nfx term weighting which is a variant of tf.idf term weighting that normalizes the lengths of the document vectors. Meiji University (Ohgaya et al., 2003) expanded the query with concept groups, and then ranked the sentences by the cosine similarity between the expanded topic vector and the sentence vector.

Berger and Lafferty (1999) proposed the use of translation models for (mono-lingual) document retrieval. They used IBM Model 1 (Brown et al., 1990), to rank documents according to their translation probability, given the query. They make no adjustment for the fact that the query and the document are in the same language, and instead rely on the translation model to learn the appropriate weights for word pairs. The models are trained on parallel data artificially constructed from the mutual information distribution of terms in the document. The results presented either were not tested for statistical significance, or they were not statistically significant, because no significance results were given.

Berger et al. (2000) used IBM Model 1 to rank answers to questions in call-center data. In their data, there were no answers that were not in response to at least one of the questions, and all questions had at least one answer. Furthermore, there are multiples of the same question. The task is to match questions and answers, given that every question has at least one match in the data. The translation models performed better for this task than the tf.idf baseline.

4 Experiments and Results

In this section we describe four methods for estimating a translation model in the absence of a parallel corpus. We describe experimental results for each of the translation models, as well as for document smoothing.

4.1 Mutual Information and TREC

As in Berger and Lafferty (1999), a set of documents was selected at random from the TREC collection, and for each document we

	Query Likelihood	MT (MI)	MT (TREC)
Prec@5	0.1176	0.1149	0.1392*
Prec@10	0.1115	0.1047	0.1095
Prec@15	0.1023	0.0928*	0.0977
Prec@20	0.0973	0.0882*	0.0936
Prec@30	0.0890	0.0865	0.0874
Prec@100	0.0733	0.0680*	0.0705
R-Prec	0.0672	0.0642*	0.0671
Ave Prec	0.0257	0.0258	0.0264

Table 1: Comparing translation model-based retrieval with description queries. "TREC" and "MI" are two ways to estimate a translation model. Results with an asterisk are significant at the .05 level with a two-tailed t-test.

	Event		Opinion	
	Query Lklhd	MT (TREC)	Query Lklhd	MT (TREC)
Prec@5	0.1149	0.1307	0.1234	0.1574
Prec@10	0.1089	0.1079	0.1170	0.1128
Prec@15	0.1036	0.1030	0.0993	0.0865
Prec@20	0.0985	0.0980	0.0947	0.0840
Prec@30	0.0901	0.0894	0.0865	0.0830
Prec@100	0.0729	0.0719	0.0743	0.0674
R-Prec	0.0658	0.0694	0.0701	0.0622
Ave Prec	0.0275	0.0289	0.0219	0.0211

Table 2: Comparing translation-based retrieval for description queries, using the relevance judgments provided by NIST. The translation model was trained from TREC topics.

constructed a distribution according to each term's mutual information with the document, and randomly generated five queries of 8 words according to this distribution. We were retrieving sentences rather than documents, so each sentence in the document was ranked according to its probability of having generated the query, and then the query was aligned with the top 5 sentences. We call this approach "MI".

The second approach uses the TREC topic titles and descriptions aligned with the top 5 retrieved sentences from documents known to be relevant to those topics, excluding topics that were included in the Novelty data. We call this approach "TREC".

Table 1 shows the results of incorporating translations for topic descriptions. Results in the tables with an asterisk are significant at the .05 level using a two-tailed t-test. The results for sentence retrieval are lower than those typically obtained for document retrieval. Manual inspection of the results indicates that the actual precision is much higher, and resembles the results for document retrieval. The lower results are an artifact of the way the relevance assessments were obtained. The sentence-level judgements from the TREC Novelty Track are only for 25 documents per topic.

The Novelty data from 2003-2004 consists of event and opinion queries. We observed that a number of the topic descriptions for event topics had a high degree of vocabulary overlap with the sentences in our data. This was not true of the opinion queries. The results of using a translation-based retrieval on description queries are given in table 2, broken down by the sentiment of the query. The Novelty queries from 2002 were included in the "event" set.

Not all of the sentences judged relevant to opinion topics express opinions. To assess opinion-relevance we evaluated the top 10 sentences, and marked sentences that expressed opinions. In our data approximately 10% of sentences in the top 10 express opinions. Table 3 shows the result of using a translation model trained on TREC data for description queries, broken down by sentiment, with the baselines evaluated for this particular set of relevance judgments. For opinion questions, the column labeled "topical" indicates topical relevance. The column labeled "opinion" indicates topical relevance that also expresses an opinion.

If we consider a sentence relevant to an opinion question only if it expresses an opinion, we see improvement in the results at the top of the ranked list for those queries, using a translation model trained on TREC data. Of the 150 topics, only 50 were opinion topics, so although the magnitude of the improvement in opinion queries is large the results are not statistically

	Topical Rel		Express Opin	
	Query Lklhd	MT (TREC)	Query Lklhd	MT (TREC)
Prec@5	.7289	.7111	.3300	.3900
Prec@10	.7089	.6867	.3125	.3775
Prec@15	.5363	.4919	.2350	.2717
Prec@20	.4300	.4033	.1875	.2188
Prec@30	.3170	.2970	.1408	.1617
Prec@100	.1236	.1131*	.0587	.0580
R-Prec	.4834	.4653	.2947	.3597*
Ave Prec	.4996	.4696	.2563	.3177

Table 3: Comparison of translation retrieval on opinion queries, using truth data we created to evaluate opinion questions. Translation models were trained with TREC data. Results with an asterisk are significant at the .05 level using a two-tailed t-test.

significant with respect to the baseline.

4.2 Lexicons

External lexicons are often useful for translation and query expansion. The most obvious approach was to incorporate a thesaurus into the training process in GIZA as a dictionary, which affects the statistics in the first iteration of EM. This is intended to improve the quality of the alignments over subsequent iterations. We incorporated the thesauri into the training process of the data generated from the artificial mutual information distribution. The dictionaries had almost no effect on the results.

4.2.1 WordNet

We created a parallel corpus of synonym-term pairs from WordNet, and added this data to the artificial mutual information data to train the translation model. The results of using this approach to retrieve sentences using title queries are in figure 1, labeled "MI_WN". Using WordNet alone, without the mutual information data, is labeled "WN_Only". The results are statistically significant using a Wilcoxon sign test at the .05 level for precision at .10, .20 and .60. Query likelihood retrieval is the baseline. The results for description queries are not shown, and were not significantly different from the baseline.

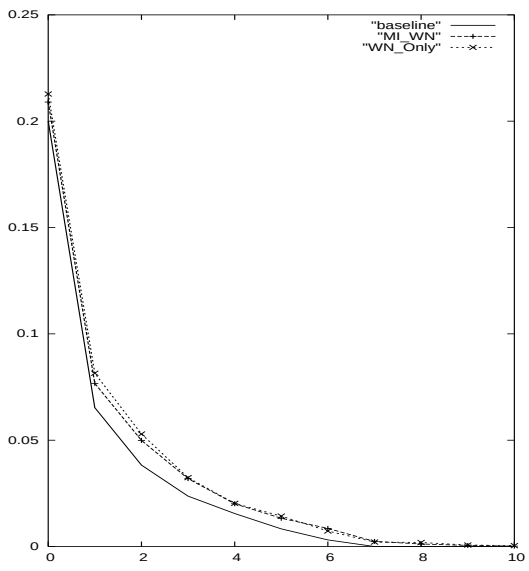

Figure 1: Comparing interpolated recall-precision for title queries using WordNet. The results are statistically significant using a Wilcoxon sign test at the .05 level, for precision at .10, .20 and .60.

4.2.2 Arabic-English corpus

Xu et al. (2002) derive an Arabic thesaurus from a parallel corpus. We derived an English thesaurus using the same approach, from a pair of English-Arabic/Arabic-English lexicons, learned from a parallel corpus. We assumed that if two English terms translate to the same Arabic term, the English terms are synonyms whose probability is given by

$$P(e_2|e_1) = \sum_{a \in A} P(e_2|a)P(a|e_1) \qquad (5)$$

Figure 2 shows the interpolated recall-precision of these results, for description queries. The English terms were not stemmed, and so the baseline query-likelihood results are also not stemmed. The results are statistically significant using a Wilcoxon sign test at the .05 level, for all retrieval levels. Not shown is the average precision, which is also significantly better for the Arabic-English lexicon than for the query-likelihood. The results for title queries are not shown, but are similar to those for descriptions.

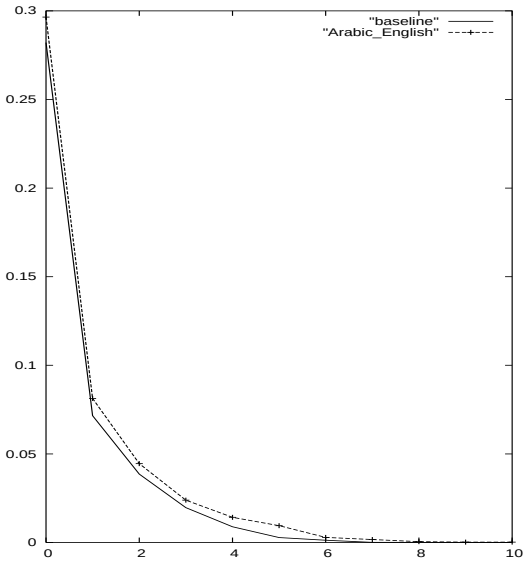

Figure 2: Comparing interpolated recall-precision for description queries using a pair of Arabic-English, English-Arabic lexicons. The results are statistically significant using a Wilcoxon sign test at the .05 level, for precision all recall levels.

4.3 Document Smoothing

Smucker and Allan (2005) demonstrated that under certain conditions, Jelinek-Mercer smoothing is equivalent to Dirichlet smoothing, and that the advantage of Dirichlet smoothing is derived from the fact that it smoothes long documents less than shorter documents. In our data there is much less variance in the length of a sentence than in the length of a document, thus we do not expect to see as great a benefit in performance from Dirichlet smoothing as has been reported in Zhai and Lafferty (2001). In fact we tried Absolute Discounting, Dirichlet, Jelinek-Mercer and Laplace smoothing and found them to produce equivalent results.

The vast majority of sentences in our data are not stand-alone units, and the topic of the sentence is also the topic of surrounding sentences. We took a context of the surrounding 5 sentences, and the surrounding 11 sentences (about 1/3 of the whole document). The sentences were smoothed from the surrounding context, backing-off to the whole document, using Jelinek-Mercer smoothing. Table 4 shows a comparison of the amount of context. Smoothing from the local context is clearly better than the baseline result.

	Query Lklhd	5 Sents	11 Sents
Prec@5	0.1203	0.1527*	0.1541*
Prec@10	0.1122	0.1446*	0.1419*
Prec@15	0.1018	0.1329*	0.1405*
Prec@20	0.0973	0.1311*	0.1345*
Prec@30	0.0890	0.1191*	0.1286*
Prec@100	0.0732	0.0935*	0.1006*
R-Prec	0.0672	0.0881*	0.0933*
Ave Prec	0.0257	0.0410*	0.0485*

Table 4: Comparison of smoothing context on description queries, retrieving sentences from the top 1000 documents. Results with an asterisk are significant at the .05 level using a two-tailed t-test.

We investigated the effect of smoothing from the entire document. Table 5 shows the results. Both topic titles and descriptions get significantly better results with document smoothing.

4.4 Novelty Relevance Task

In the TREC Novelty Track, participants are given a set of 25 documents most of which are relevant for each topic. If we believe that a document is relevant because it has relevant sentences in it, then a "good" sentence would come from a "good" document. This would suggest that smoothing from the document the sentence came from would improve retrieval. We found that for title queries document smoothing improved precision in the top 5 documents by 12.5%, which is statistically significant using a two-tailed t-test at the .05 level. Precision in the top 10 - 100 documents also improved results by an average of 5%, but the result is not statistically significant. For description queries, smoothing from the document had no effect.

For title queries, translation models improve the average precision, and R-Precision. For both title and description queries, the number of relevant documents that are retrieved is also improved with translation models.

	Title		Description	
	Query Lklhd	Doc Smth	Query Lklhd	Doc Smth
Prec@5	.0765	.2268*	.1203	.2362*
Prec@10	.0805	.2262*	.1122	.2128*
Prec@15	.0814	.2192*	.1018	.2000*
Prec@20	.0765	.2124*	.0973	.1893*
Prec@30	.0765	.2007*	.0890	.1743*
Prec@100	.0675	.1638*	.0732	.1335*
R-Prec	.0646	.1379*	.0672	.1226*
Ave Prec	.0243	.0796*	.0257	.0749*

Table 5: Comparison of document smoothing to query likelihood retrieving sentences from the top 1000 documents. Results with an asterisk are significant at the .05 level using a two-tailed t-test.

5 Discussion

The results for sentence retrieval are low, in comparison to results we would expect for document retrieval. We might think that although we show improvements, nothing is working well. In reality, the relevance assessments provided by NIST as part of the Novelty Track only cover 25 documents per topic. Evaluating the top 10 sentences by hand shows that the systems give a performance comparable to document retrieval systems, and the low numbers are the result of a lack of coverage in the assessments. Unfortunately, there is no collection of documents of significant size, where the relevance assessments at the sentence level cover the collection. Constructing such a corpus would be a major undertaking, outside of the scope of this paper.

The best performing method of constructing a parallel corpus used a bilingual lexicon derived from a sentence-aligned Arabic-English parallel corpus. This suggests that data in which sentences are actually translations of one another, as opposed to sentences aligned with key terms from the document, yield a higher quality lexicon. The model trained on the parallel corpus of TREC topics and relevant sentences performed better than the MI corpus, but not as well as the Arabic-English corpus. The TREC corpus consisted of approximately 15,000 sentence pairs, whereas the Arabic-English corpus was trained on more than a million sentence pairs. This may account in part for the higher quality results. In addition, the TREC corpus was created by retrieving the top 5 sentences from each relevant document. Even when the document is known to be relevant, the retrieval process is noisy. Furthermore, although there were 15,000 sentence pairs, there were only 450 unique queries, limiting the size of the source vocabulary.

Opinion topics have much less vocabulary overlap with relevant sentences than do event topics. Translation models would be expected to perform better when retrieving sentences that contain synonym or related terms. For sentences that have exact matches in the query, query likelihood will perform better.

We find that smoothing from the local context of the sentence performs significantly better than the baseline retrieval. The sentences are all about the same length, so there is no performance advantage to using Dirichlet smoothing, whose smoothing parameter is a function of the document length. The smoothing parameters gave very little weight to the collection. As sentences have few terms, relative to documents, matching a term in the query is a good indication of relevance.

6 Conclusions and Future Work

We have shown that translation models improve retrieval for title and opinion queries, and that a translation model derived from a high-quality bilingual lexicon significantly improves retrieval for title and description queries. Smoothing from the local context of a sentence dramatically improves retrieval, with smoothing from the document that contains the sentence performing the best.

We evaluated sentences based on lexical similarity, but structural similarity is also an important measure, which we plan to investigate in the future. The translation model we used was the most basic model. We used this model because it had been shown effective in document retrieval, and was easily incorporated in the query-likelihood framework, but we intend

to explore more sophisticated translation models, and better alignment mechanisms. Preliminary results suggest that sentence retrieval can be used to improve document retrieval, but we plan a more extensive investigation of evaluating document similarity and relevance based on sentence-level similarity.

7 Acknowledgements

The authors would like to thank Leah Larkey for her Arabic-English lexicons. This work was supported in part by the Center for Intelligent Information Retrieval, in part by Advanced Research and Development Activity and NSF grant #CCF-0205575 , and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, I. D. Melamed, F. J. Och, D. Purdy, N. A. Smith, and D. Yarowsky. 1999. Statistical machine translation, final report, JHU workshop.

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*.

Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of the 23rd Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*, pages 192–199.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelineck, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Kevyn Collins-Thompson, Paul Ogilvie, Yi Zhang, and Jamie Callan. 2002. Information filtering, novelty detection and named-page finding. In *Proceedings of the Eleventh Text Retrieval Conference (TREC)*.

Donna Harman. 2002. Overview of the TREC 2002 novelty track. In *Proceedings of the Eleventh Text Retrieval Conference (TREC)*.

Leah Larkey, James Allan, Margie Connell, Alvaro Bolivar, and Courtney Wade. 2002. UMass at TREC 2002: Cross language and novelty tracks. In *Proceedings of the Eleventh Text Retrieval Conference (TREC)*, page 721.

Christof Monz, Jaap Kamps, and Maarten de Rijke. 2002. The University of Amsterdam at TREC 2002. In *Proceedings of the Eleventh Text Retrieval Conference (TREC)*.

Vanessa Murdock and W. Bruce Croft. 2004. Simple translation models for sentence retrieval in factoid question answering. In *Proceedings of the Information Retrieval for Question Answering Workshop at SIGIR 2004*.

Ryosuke Ohgaya, Akiyoshi Shimmura, and Tomohiro Takagi. 2003. Meiji University web and novelty track experiments at TREC 2003. In *Proceedings of the Twelth Text Retrieval Conference (TREC)*.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*. http://www.cis.upenn.edu/ãdwait/statnlp.html.

Mark Smucker and James Allan. 2005. An investigation of dirichlet prior smoothing's performance advantage. Technical Report IR-391, The University of Massachusetts, The Center for Intelligent Information Retrieval.

Ian Soboroff and Donna Harman. 2003. Overview of the TREC 2003 novelty track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC)*.

Ian Soboroff. 2004. Overview of the TREC 2004 novelty track. In *Proceedings of the Thirteenth Text Retrieval Conference (TREC)*. forthcoming.

Jinxi Xu, Alexander Fraser, and Ralph Weischedel. 2002. Empirical studies in strategies for arabic retrieval. In *Proceedings of the 25th Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*.

ChengXiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342.