Statistical Machine Translation for Query Expansion in Answer Retrieval

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal and **Yi Liu**
Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043
{riezler|avasserm|ioannis|vibhu|yliu}@google.com

Abstract

We present an approach to query expansion in answer retrieval that uses Statistical Machine Translation (SMT) techniques to bridge the lexical gap between questions and answers. SMT-based query expansion is done by i) using a full-sentence paraphraser to introduce synonyms in context of the entire query, and ii) by translating query terms into answer terms using a full-sentence SMT model trained on question-answer pairs. We evaluate these global, context-aware query expansion techniques on *tfidf* retrieval from 10 million question-answer pairs extracted from FAQ pages. Experimental results show that SMT-based expansion improves retrieval performance over local expansion and over retrieval without expansion.

1 Introduction

One of the fundamental problems in Question Answering (QA) has been recognized to be the "lexical chasm" (Berger et al., 2000) between question strings and answer strings. This problem is manifested in a mismatch between question and answer vocabularies, and is aggravated by the inherent ambiguity of natural language. Several approaches have been presented that apply natural language processing technology to close this gap. For example, syntactic information has been deployed to reformulate questions (Hermjakob et al., 2002) or to replace questions by syntactically similar ones (Lin and Pantel, 2001); lexical ontologies such as Wordnet[1] have been used to find synonyms for question words (Burke et al., 1997; Hovy et al., 2000; Prager et al., 2001; Harabagiu et al., 2001), and statistical machine translation (SMT) models trained on question-answer pairs have been used to rank candidate answers according to their translation probabilities (Berger et al., 2000; Echihabi and Marcu, 2003; Soricut and Brill, 2006). Information retrieval (IR) is faced by a similar fundamental problem of "term mismatch" between queries and documents. A standard IR solution, query expansion, attempts to increase the chances of matching words in relevant documents by adding terms with similar statistical properties to those in the original query (Voorhees, 1994; Qiu and Frei, 1993; Xu and Croft, 1996).

In this paper we will concentrate on the task of answer retrieval from FAQ pages, i.e., an IR problem where user queries are matched against documents consisting of question-answer pairs found in FAQ pages. Equivalently, this is a QA problem that concentrates on finding answers given FAQ documents that are known to contain the answers. Our approach to close the lexical gap in this setting attempts to marry QA and IR technology by deploying SMT methods for query expansion in answer retrieval. We present two approaches to SMT-based query expansion, both of which are implemented in the framework of phrase-based SMT (Och and Ney, 2004; Koehn et al., 2003).

Our first query expansion model trains an end-to-end phrase-based SMT model on 10 million question-answer pairs extracted from FAQ pages.

[1] http://wordnet.princeton.edu

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 464–471,
Prague, Czech Republic, June 2007.

The goal of this system is to learn lexical correlations between words and phrases in questions and answers, for example by allowing for multiple unaligned words in automatic word alignment, and disregarding issues such as word order. The ability to translate phrases instead of words and the use of a large language model serve as rich context to make precise decisions in the case of ambiguous translations. Query expansion is performed by adding content words that have not been seen in the original query from the n-best translations of the query.

Our second query expansion model is based on the use of SMT technology for full-sentence paraphrasing. A phrase table of paraphrases is extracted from bilingual phrase tables (Bannard and Callison-Burch, 2005), and paraphrasing quality is improved by additional discriminative training on manually created paraphrases. This approach utilizes large bilingual phrase tables as information source to extract a table of para-phrases. Synonyms for query expansion are read off from the n-best paraphrases of full queries instead of from paraphrases of separate words or phrases. This allows the model to take advantage of the rich context of a large n-gram language model when adding terms from the n-best paraphrases to the original query.

In our experimental evaluation we deploy a database of question-answer pairs extracted from FAQ pages for both training a question-answer translation model, and for a comparative evaluation of different systems on the task of answer retrieval. Retrieval is based on the *tfidf* framework of Jijkoun and de Rijke (2005), and query expansion is done straightforwardly by adding expansion terms to the query for a second retrieval cycle. We compare our global, context-aware query expansion techniques with Jijkoun and de Rijke's (2005) *tfidf* model for answer retrieval and a local query expansion technique (Xu and Croft, 1996). Experimental results show a significant improvement of SMT-based query expansion over both baselines.

2 Related Work

QA has approached the problem of the lexical gap by various techniques for *question reformulation*, including rule-based syntactic and semantic reformulation patterns (Hermjakob et al., 2002), reformulations based on shared dependency parses (Lin and Pantel, 2001), or various uses of the WordNet ontology to close the lexical gap word-by-word (Hovy et al., 2000; Prager et al., 2001; Harabagiu et al., 2001). Another use of natural language processing has been the deployment of SMT models on question-answer pairs for *(re)ranking* candidate answers which were either assumed to be contained in FAQ pages (Berger et al., 2000) or retrieved by baseline systems (Echihabi and Marcu, 2003; Soricut and Brill, 2006).

IR has approached the term mismatch problem by various approaches to *query expansion* (Voorhees, 1994; Qiu and Frei, 1993; Xu and Croft, 1996). Inconclusive results have been reported for techniques that expand query terms separately by adding strongly related terms from an external thesaurus such as WordNet (Voorhees, 1994). Significant improvements in retrieval performance could be achieved by *global* expansion techniques that compute corpus-wide statistics and take the entire query, or query *concept* (Qiu and Frei, 1993), into account, or by *local* expansion techniques that select expansion terms from the top ranked documents retrieved by the original query (Xu and Croft, 1996).

A similar picture emerges for query expansion in QA: Mixed results have been reported for word-by-word expansion based on WordNet (Burke et al., 1997; Hovy et al., 2000; Prager et al., 2001; Harabagiu et al., 2001). Considerable improvements have been reported for the use of the local context analysis model of Xu and Croft (1996) in the QA system of Ittycheriah et al. (2001), or for the systems of Agichtein et al. (2004) or Harabagiu and Lacatusu (2004) that use FAQ data to learn how to expand query terms by answer terms.

The SMT-based approaches presented in this paper can be seen as global query expansion techniques in that our question-answer translation model uses the whole question-answer corpus as information source, and our approach to paraphrasing deploys large amounts of bilingual phrases as high-coverage information source for synonym finding. Furthermore, both approaches take the entire query context into account when proposing to add new terms to the original query. The approaches that are closest to our models are the SMT approach of Radev et al. (2001) and the paraphrasing approach

	web pages	FAQ pages	QA pairs
count	4 billion	795,483	10,568,160

Table 1: Corpus statistics of QA pair data

of Duboue and Chu-Carroll (2006). None of these approaches defines the problem of the lexical gap as a query expansion problem, and both approaches use much simpler SMT models than our systems, e.g., Radev et al. (2001) neglect to use a language model to aid disambiguation of translation choices, and Duboue and Chu-Carroll (2006) use SMT as black box altogether.

In sum, our approach differs from previous work in QA and IR in the use SMT technology for query expansion, and should be applicable in both areas even though experimental results are only given for the restricted domain of retrieval from FAQ pages.

3 Question-Answer Pairs from FAQ Pages

Large-scale collection of question-answer pairs has been hampered in previous work by the small sizes of publicly available FAQ collections or by restricted access to retrieval results via public APIs of search engines. Jijkoun and de Rijke (2005) nevertheless managed to extract around 300,000 FAQ pages and 2.8 million question-answer pairs by repeatedly querying search engines with "`intitle:faq`" and "`inurl:faq`". Soricut and Brill (2006) could deploy a proprietary URL collection of 1 billion URLs to extract 2.3 million FAQ pages containing the uncased string "faq" in the url string. The extraction of question-answer pairs amounted to a database of 1 million pairs in their experiment. However, inspection of the publicly available Web-FAQ collection provided by Jijkoun and de Rijke[2] showed a great amount of noise in the retrieved FAQ pages and question-answer pairs, and yet the indexed question-answer pairs showed a serious recall problem in that no answer could be retrieved for many well-formed queries. For our experiment, we decided to prefer precision over recall and to attempt a precision-oriented FAQ and question-answer pair extraction that benefits the training of question-answer translation models.

[2]http://ilps.science.uva.nl/Resources/WazDah/

As shown in Table 1, the FAQ pages used in our experiment were extracted from a 4 billion page subset of the web using the queries "`inurl:faq`" and "`inurl:faqs`" to match the tokens "faq" or "faqs" in the urls. This extraction resulted in 2.6 million web pages (0.07% of the crawl). Since not all those pages are actually FAQs, we manually labeled 1,000 of those pages to train an online passive-aggressive classificier (Crammer et al., 2006) in a 10-fold cross validation setup. Training was done using 20 feature functions on occurrences question marks and key words in different fields of web pages, and resulted in an F1 score of around 90% for FAQ classification. Application of the classifier to the extracted web pages resulted in a classification of 795,483 pages as FAQ pages.

The extraction of question-answer pairs from this database of FAQ pages was performed again in a precision-oriented manner. The goal of this step was to extract url, title, question, and answers fields from the question-answer pairs in FAQ pages. This was achieved by using feature functions on punctuations, HTML tags (e.g., `<p>, <BR>`), listing markers (e.g., `Q:, (1)`), and lexical cues (e.g., `What, How`), and an algorithm similar to Joachims (2003) to propagate initial labels across similar text pieces. The result of this extraction step is a database of about 10 million question answer pairs (13.3 pairs per FAQ page). A manual evaluation of 100 documents, containing 1,303 question-answer pairs, achieved a precision of 98% and a recall of 82% for extracting question-answer pairs.

4 SMT-Based Query Expansion

Our SMT-based query expansion techniques are based on a recent implementation of the phrase-based SMT framework (Koehn et al., 2003; Och and Ney, 2004). The probability of translating a foreign sentence **f** into English **e** is defined in the noisy channel model as

$$\arg\max_e p(\mathbf{e}|\mathbf{f}) \quad = \quad \arg\max_e p(\mathbf{f}|\mathbf{e})p(\mathbf{e}) \qquad (1)$$

This allows for a separation of a language model $p(\mathbf{e})$, and a translation model $p(\mathbf{f}|\mathbf{e})$. Translation probabilities are calculated from relative frequencies of phrases, which are extracted via various heuristics as larger blocks of aligned words from best word alignments. Word alignments are estimated by models similar to Brown et al. (1993). For a sequence of I phrases, the translation probability in equation (1) can be decomposed into

$$p(f_i^I|e_i^I) = \prod_{i=1}^{I} p(f_i|e_i) \quad (2)$$

Recent SMT models have shown significant improvements in translation quality by improved modeling of local word order and idiomatic expressions through the use of phrases, and by the deployment of large n-gram language models to model fluency and lexical choice.

4.1 Question-Answer Translation

Our first approach to query expansion treats the questions and answers in the question-answer corpus as two distinct languages. That is, the 10 million question-answer pairs extracted from FAQ pages are fed as parallel training data into an SMT training pipeline. This training procedure includes various standard procedures such as preprocessing, sentence and chunk alignment, word alignment, and phrase extraction. The goal of question-answer translation is to learn associations between question words and synonymous answer words, rather than the translation of questions into fluent answers. Thus we did not conduct discriminative training of feature weights for translation probabilities or language model probabilities, but we held out 4,000 question-answer pairs for manual development and testing of the system. For example, the system was adjusted to account for the difference in sentence length between questions and answers by setting the null-word probability parameter in word alignment to 0.9. This allowed us to concentrate the word alignments to a small number of key words. Furthermore, extraction of phrases was based on the intersection of alignments from both translation directions, thus favoring precision over recall also in phrase alignment.

Table 2 shows unique translations of the query "how to live with cat allergies" on the phrase-level, with corresponding source and target phrases shown in brackets. Expansion terms are taken from phrase terms that have not been seen in the original query, and are highlighted in bold face.

4.2 SMT-Based Paraphrasing

Our SMT-based paraphrasing system is based on the approach presented in Bannard and Callison-Burch (2005). The central idea in this approach is to identify paraphrases or synonyms at the phrase level by pivoting on another language. For example, given a table of Chinese-to-English phrase translations, phrasal synonyms in the target language are defined as those English phrases that are aligned to the same Chinese source phrases. Translation probabilities for extracted para-phrases can be inferred from bilingual translation probabilities as follows: Given an English para-phrase pair (trg, syn), the probability $p(syn|trg)$ that trg translates into syn is defined as the joint probability that the English phrase trg translates into the foreign phrase src, and that the foreign phrase src translates into the English phrase syn. Under an independence assumption of those two events, this probability and the reverse translation direction $p(trg|syn)$ can be defined as follows:

$$\begin{aligned} p(syn|trg) &= \max_{src} p(src|trg)p(syn|src) \quad (3) \\ p(trg|syn) &= \max_{src} p(src|syn)p(trg|src) \end{aligned}$$

Since the same para-phrase pair can be obtained by pivoting on multiple foreign language phrases, a summation or maximization over foreign language phrases is necessary. In order not to put too much probability mass onto para-phrase translations that can be obtained from multiple foreign language phrases, we maximize instead of summing over src.

In our experiments, we employed equation (3) to infer for each para-phrase pair translation model probabilities $p_\phi(syn|trg)$ and $p_{\phi'}(trg|syn)$ from relative frequencies of phrases in bilingual tables. In contrast to Bannard and Callison-Burch (2005), we applied the same inference step to infer also lexical translation probabilities $p_w(syn|trg)$ and $p_{w'}(trg|syn)$ as defined in Koehn et al. (2003) for para-phrases. Furthermore, we deployed features for the number of words l_w, number of phrases c_ϕ, a reordering score p_d, and a score for a 6-gram language model p_{LM} trained on English web data. The final model combines these features in a log-linear model that defines the probability of paraphrasing a full sentence, consisting of a sequence of I phrases

qa-translation	(how, how) (to, to) (live, live) (with, with) (cat, **pet**) (allergies, allergies) (how, how) (to, to) (live, live) (with, with) (cat, cat) (allergies, **allergy**) (how, how) (to, to) (live, live) (with, with) (cat, cat) (allergies, **food**) (how, how) (to, to) (live, live) (with, with) (cat, **cats**) (allergies, allergies)
paraphrasing	(how, how) (to live, to live) (with cat, with cat) (allergies, **allergy**) (how, **ways**) (to live, to live) (with cat, with cat) (allergies, allergies) (how, how) (to live with, to live with) (cat, **feline**) (allergies, allergies) (how to, how to) (live, **living**) (with cat, with cat) (allergies, allergies) (how to, how to) (live, **life**) (with cat, with cat) (allergies, allergies) (how, **way**) (to live, to live) (with cat, with cat) (allergies, allergies) (how, how) (to live, to live) (with cat, with cat) (allergies, **allergens**) (how, how) (to live, to live) (with cat, with cat) (allergies, **allergen**)

Table 2: Unique n-best phrase-level translations of query "how to live with cat allergies".

as follows:

$$\begin{aligned} p(syn_1^I|trg_1^I) &= (\prod_{i=1}^{I} p_\phi(syn_i|trg_i)^{\lambda_\phi} \qquad (4) \\ &\times \; p_{\phi'}(trg_i|syn_i)^{\lambda_{\phi'}} \\ &\times \; p_w(syn_i|trg_i)^{\lambda_w} \\ &\times \; p_{w'}(trg_i|syn_i)^{\lambda_{w'}} \\ &\times \; p_d(syn_i, trg_i)^{\lambda_d}) \\ &\times \; l_w(syn_1^I)^{\lambda_l} \\ &\times \; c_\phi(syn_1^I)^{\lambda_c} \\ &\times \; p_{LM}(syn_1^I)^{\lambda_{LM}} \end{aligned}$$

For estimation of the feature weights $\vec{\lambda}$ defined in equation (4) we employed minimum error rate (MER) training under the BLEU measure (Och, 2003). Training data for MER training were taken from multiple manual English translations of Chinese sources from the NIST 2006 evaluation data. The first of four reference translations for each Chinese sentence was taken as source paraphrase, the rest as reference paraphrases. Discriminative training was conducted on 1,820 sentences; final evaluation on 2,390 sentences. A baseline paraphrase table consisting of 33 million English para-phrase pairs was extracted from 1 billion phrase pairs from three different languages, at a cutoff of para-phrase probabilities of 0.0025.

Query expansion is done by adding terms introduced in n-best paraphrases of the query. Table 2 shows example paraphrases for the query "how to live with cat allergies" with newly introduced terms highlighted in bold face.

5 Experimental Evaluation

Our baseline answer retrieval system is modeled after the *tfidf* retrieval model of Jijkoun and de Rijke (2005). Their model calculates a linear combination of vector similarity scores between the user query and several fields in the question-answer pair. We used the cosine similarity metric with logarithmically weighted term and document frequency weights in order to reproduce the Lucene[3] model used in Jijkoun and de Rijke (2005). For indexing of fields, we adopted the settings that were reported to be optimal in Jijkoun and de Rijke (2005). These settings comprise the use of 8 question-answer pair fields, and a weight vector $\langle 0.0, 1.0, 0.0, 0.0, 0.5, 0.5, 0.2, 0.3 \rangle$ for fields ordered as follows: (1) full FAQ document text, (2) question text, (3) answer text, (4) title text, (5)-(8) each of the above without stopwords. The second field thus takes takes wh-words, which would typically be filtered out, into account. All other fields are matched without stopwords, with higher weight assigned to document and question than to answer and title fields. We did not use phrase-matching or stemming in our experiments, similar to Jijkoun and de Rijke (2005), who could not find positive effects for these features in their experiments.

Expansion terms are taken from those terms in the n-best translations of the query that have not been seen in the original query string. For paraphrasing-based query expansion, a 50-best list of paraphrases of the original query was used. For the noisier question-answer translation, expansion terms and phrases were extracted from a 10-

[3]http://lucene.apache.org

	S_2@10	S_2@20	$S_{1,2}$@10	$S_{1,2}$@20
baseline *tfidf*	27	35	58	65
local expansion	30 (+ 11.1)	40 (+ 14.2)	57 (- 1)	63 (- 3)
SMT-based expansion	38 (+ 40.7)	43 (+ 22.8)	58	65

Table 3: Success rate at 10 or 20 results for retrieval of adequate (2) or material (1) answers; relative change in brackets.

best list of query translations. Terms taken from query paraphrases were matched with the same field weight vector $\langle 0.0, 1.0, 0.0, 0.0, 0.5, 0.5, 0.2, 0.3\rangle$ as above. Terms taken from question-answer translation were matched with the weight vector $\langle 0.0, 1.0, 0.0, 0.0, 0.5, 0.2, 0.5, 0.3\rangle$, preferring answer fields over question fields. After stopword removal, the average number of expansion terms produced was 7.8 for paraphrasing, and 3.1 for question-answer translation.

The local expansion technique used in our experiments follows Xu and Croft (1996) in taking expansion terms from the top n answers that were retrieved by the baseline *tfidf* system, and by incorporating cooccurrence information with query terms. This is done by calculating term frequencies for expansion terms by summing up the *tfidf* weights of the answers in which they occur, thus giving higher weight to terms that occur in answers that receive a higher similarity score to the original query. In our experiments, expansion terms are ranked according to this modified *tfidf* calculation over the top 20 answers retrieved by the baseline retrieval run, and matched a second time with the field weight vector $\langle 0.0, 1.0, 0.0, 0.0, 0.5, 0.2, 0.5, 0.3\rangle$ that prefers answer fields over question fields. After stopword removal, the average number of expansion terms produced by the local expansion technique was 9.25.

The test queries we used for retrieval are taken from query logs of the MetaCrawler search engine[4] and were provided to us by Valentin Jijkoun. In order to maximize recall for the comparative evaluation of systems, we selected 60 queries that were well-formed natural language questions without metacharacters and spelling errors. However, for one third of these well-formed queries none of the five compared systems could retrieve an answer. Examples are *"how do you make a cornhusk doll"*, *"what is the idea of materialization"*, or *"what does 8x certified mean"*, pointing to a severe recall problem of the question-answer database.

Evaluation was performed by manual labeling of top 20 answers retrieved for each of 60 queries for each system by two independent judges. For the sake of consistency, we chose not to use the assessments provided by Jijkoun and de Rijke. Instead, the judges were asked to find agreement on the examples on which they disagreed after each evaluation round. The ratings together with the question-answer pair id were stored and merged into the retrieval results for the next system evaluation. In this way consistency across system evaluations could be ensured, and the effort of manual labeling could be substantially reduced. The quality of retrieval results was assessed according to Jijkoun and de Rijke's (2005) three point scale:

- adequate (2): answer is contained
- material (1): no exact answer, but important information given
- unsatisfactory (0): user's information need is not addressed

The evaluation measure used in Jijkoun and de Rijke (2005) is the success rate at 10 or 20 answers, i.e., $S_2@n$ is the percentage of queries with at least one adequate answer in the top n retrieved question-answer pairs, and $S_{1,2}@n$ is the percentage of queries with at least one adequate or material answer in the top n results. This evaluation measure accounts for improvements in coverage, i.e., it rewards cases where answers are found for queries that did not have an adequate or material answer before. In contrast, the mean reciprocal rank (MRR) measure standardly used in QA can have the effect of preferring systems that find answers only for a small set of queries, but rank them higher than systems with

[4]http://www.metacrawler.com

(1)	query:	how to live with cat allergies
	local expansion (-):	allergens allergic infections filter plasmacluster rhinitis introduction effective replacement
	qa-translation (+):	allergy cats pet food
	paraphrasing (+):	way allergens life allergy feline ways living allergen
(2)	query:	how to design model rockets
	local expansion (-):	models represented orientation drawings analysis element environment different structure
	qa-translation (+):	models rocket
	paraphrasing (+):	missiles missile rocket grenades arrow designing prototype models ways paradigm
(3)	query:	what is dna hybridization
	local expansion (-):	instructions individual blueprint characteristics chromosomes deoxyribonucleic information biological genetic molecule
	qa-translation (+):	slides clone cdna sitting sequences
	paraphrasing (+):	hibridization hybrids hybridation anything hibridacion hybridising adn hybridisation nothing
(4)	query:	how to enhance competitiveness of indian industries
	local expansion (+):	resources production quality processing established investment development facilities institutional
	qa-translation (+):	increase industry
	paraphrasing (+):	promote raise improve increase industry strengthen
(5)	query:	how to induce labour
	local expansion (-):	experience induction practice imagination concentration information consciousness different meditation relaxation
	qa-translation (-):	birth industrial induced induces
	paraphrasing (-):	way workers inducing employment ways labor working child work job action unions

Table 4: Examples for queries and expansion terms yielding improved (+), decreased (-), or unchanged (0) retrieval performance compared to retrieval without expansion.

higher coverage. This makes MRR less adequate for the low-recall setup of FAQ retrieval.

Table 3 shows success rates at 10 and 20 retrieved question-answer pairs for five different systems. The results for the *baseline tfidf* system, following Jijkoun and de Rijke (2005), are shown in row 2. Row 3 presents results for our variant of *local expansion* by pseudo-relevance feedback (Xu and Croft, 1996). Results for *SMT-based expansion* are given in row 4. A comparison of success rates for retrieving at least one adequate answer in the top 10 results shows relative improvements over the baseline of 11.1% for local query expansion, and of 40.7% for combined SMT-based expansion. Success rates at top 20 results show similar relative improvements of 14.2% for local query expansion, and of 22.8% for combined SMT-based expansion. On the easier task of retrieving a material or adequate answer, success rates drop by a small amount for local expansion, and stay unchanged for SMT-based expansion.

These results can be explained by inspecting a few sample query expansions. Examples (1)-(3) in Table 4 illustrate cases where SMT-based query expansion improves results over baseline performance, but local expansion decreases performance by introducing irrelevant terms. In (4) retrieval performance is improved over the baseline for both expansion techniques. In (5) both local and SMT-based expansion introduce terms that decrease retrieval performance compared to retrieval without expansion.

6 Conclusion

We presented two techniques for query expansion in answer retrieval that are based on SMT technology. Our method for question-answer translation uses a large corpus of question-answer pairs extracted from FAQ pages to learn a translation model from questions to answers. SMT-based paraphrasing utilizes large amounts of bilingual data as a new information source to extract phrase-level synonyms. Both SMT-based techniques take the entire query context into account when adding new terms to the original query. In an experimental comparison with a baseline *tfidf* approach and a local query expansion technique on the task of answer retrieval from FAQ pages, we showed a significant improvement of both SMT-based query expansion over both baselines.

Despite the small-scale nature of our current experimental results, we hope to apply the presented techniques to general web retrieval in future work. Another task for future work is to scale up the extraction of question-answer pair data in order to provide an improved resource for question-answer translation.

References

Eugene Agichtein, Steve Lawrence, and Luis Gravano. 2004. Learning to find answers to questions on the web. *ACM Transactions on Internet Technology*, 4(2):129–162.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of (ACL'05)*, Ann Arbor, MI.

Adam L. Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of SIGIR'00*, Athens, Greece.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Robin B. Burke, Kristian J. Hammond, and Vladimir A. Kulyukin. 1997. Question answering from frequently-asked question files: Experiences with the FAQ finder system. *AI Magazine*, 18(2):57–66.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yo-ram Singer. 2006. Online passive-agressive algorithms. *Machine Learning*, 7:551–585.

Pablo Ariel Duboue and Jennifer Chu-Carroll. 2006. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *Proceedings of (HLT-NAACL'06)*, New York, NY.

Abdessamad Echihabi and Daniel Marcu. 2003. A noisy-channel approach to question answering. In *Proceedings of (ACL'03)*, Sapporo, Japan.

Sanda Harabagiu and Finley Lacatusu. 2004. Strategies for advanced question answering. In *Proceedings of the HLT-NAACL'04 Workshop on Pragmatics of Question Answering*, Boston, MA.

Sanda Harabagiu, Dan Moldovan, Marius Paşca, Rada Mihalcea, Mihai Surdeanu, Răzvan Bunescu, Roxana Gîrju, Vasile Rus, and Paul Morărescu. 2001. The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of (ACL'01)*, Toulouse, France.

Ulf Hermjakob, Abdessamad Echihabi, and Daniel Marcu. 2002. Natural language based reformulation resource and web exploitation for question answering. In *Proceedings of TREC-11*, Gaithersburg, MD.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, and Chin-Yew Lin. 2000. Question answering in webclopedia. In *Proceedings of TREC 9*, Gaithersburg, MD.

Abraham Ittycheriah, Martin Franz, and Salim Roukos. 2001. IBM's statistical question answering system. In *Proceedings of TREC 10*, Gaithersburg, MD.

Valentin Jijkoun and Maarten de Rijke. 2005. Retrieving answers from frequently asked questions pages on the web. In *Proceedings of the Tenth ACM Conference on Information and Knowledge Management (CIKM'05)*, Bremen, Germany.

Thorsten Joachims. 2003. Transductive learning via spectral graph partitioning. In *Proceedings of ICML'03*, Washington, DC.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of (HLT-NAACL'03)*, Edmonton, Cananda.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Journal of Natural Language Engineering*, 7(3):343–360.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of (HLT-NAACL'03)*, Edmonton, Cananda.

John Prager, Jennifer Chu-Carroll, and Krysztof Czuba. 2001. Use of wordnet hypernyms for answering what-is questions. In *Proceedings of TREC 10*, Gaithersburg, MD.

Yonggang Qiu and H. P. Frei. 1993. Concept based query expansion. In *Proceedings of SIGIR'93*, Pittsburgh, PA.

Dragomir R. Radev, Hong Qi, Zhiping Zheng, Sasha Blair-Goldensohn, Zhu Zhang, Weigo Fan, and John Prager. 2001. Mining the web for answers to natural language questions. In *Proceedings of (CIKM'01)*, Atlanta, GA.

Radu Soricut and Eric Brill. 2006. Automatic question answering using the web: Beyond the factoid. *Journal of Information Retrieval - Special Issue on Web Information Retrieval*, 9:191–206.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of SIGIR'94*, Dublin, Ireland.

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of SIGIR'96*, Zurich, Switzerland.