Query Rewriting Using Monolingual Statistical Machine Translation

Stefan Riezler*
Google

Yi Liu**
Google

Long queries often suffer from low recall in Web search due to conjunctive term matching. The chances of matching words in relevant documents can be increased by rewriting query terms into new terms with similar statistical properties. We present a comparison of approaches that deploy user query logs to learn rewrites of query terms into terms from the document space. We show that the best results are achieved by adopting the perspective of bridging the "lexical chasm" between queries and documents by translating from a source language of user queries into a target language of Web documents. We train a state-of-the-art statistical machine translation model on query-snippet pairs from user query logs, and extract expansion terms from the query rewrites produced by the monolingual translation system. We show in an extrinsic evaluation in a real-world Web search task that the combination of a query-to-snippet translation model with a query language model achieves improved contextual query expansion compared to a state-of-the-art query expansion model that is trained on the same query log data.

1. Introduction

Information Retrieval (IR) applications have been notoriously resistant to improvement attempts by Natural Language Processing (NLP). With a few exceptions for specialized tasks,[1] the contribution of part-of-speech taggers, syntactic parsers, or ontologies of nouns or verbs has been inconclusive. In this article, instead of deploying NLP tools or ontologies, we apply NLP ideas to IR problems. In particular, we take a viewpoint that looks at the problem of the word mismatch between queries and documents in Web search as a problem of translating from a source language of user queries into a target language of Web documents. We concentrate on the task of query expansion by query rewriting. This task consists of adding expansion terms with similar statistical properties to the original query in order to increase the chances of matching words in relevant documents, and also to decrease the ambiguity of the query that is inherent in natural language. We focus on a comparison of models that learn to generate query

* Brandschenkestrasse 110, 8002 Zürich, Switzerland. E-mail: riezler@gmail.com.
** 1600 Amphitheatre Parkway, Mountain View, CA. E-mail: yliu@google.com.
1 See for example Sable, McKeown, and Church (2002), who report improvements in text categorization by using tagging and parsing for the task of categorizing captioned images.

Submission received: 19 June 2009; revised submission received: 4 March 2010; accepted for publication: 12 May 2010.

rewrites from large amounts of user query logs, and use query expansion in Web search for an extrinsic evaluation of the produced rewrites. The experimental query expansion setup used in this article is simple and direct: For a given set of randomly selected queries, *n*-best rewrites are produced. From the changes introduced by the rewrites, expansion terms are extracted and added as alternative terms to the query, leaving the ranking function untouched.

Figure 1 shows expansions of the queries *herbs for chronic constipation* and *herbs for mexican cooking* using AND and OR operators. Conjunctive matching of all query terms is the default, and indicated by the AND operator. Expansion terms are added using the OR operator. The example in Figure 1 illustrates the key requirements to successful query expansion, namely, to find appropriate expansions in the context of the query. While *remedies*, *medicine*, or *supplement* are appropriate expansions in the context of the first query, they would cause a severe query drift if used in the second query. In the context of the second query, *spices* is an appropriate expansion for *herbs*, whereas this expansion would again not work for the first query.

The central idea behind our approach is to combine the orthogonal information sources of the translation model and the language model to expand query terms *in context*. The translation model proposes expansion candidates, and the query language model performs a selection in the context of the surrounding query terms. Thus, in combination, the incessant problems of term ambiguity and query drift can be solved. One of the goals of this article is to show that existing SMT technology is readily applicable to this task. We apply SMT to large parallel data of queries on the source side, and snippets of clicked search results on the target side. Snippets are short text fragments that represent the parts of the result pages that are most relevant to the queries, for example, in terms of query term matches. Although the use of snippets instead of the full documents makes our approach efficient, it introduces noise because text fragments are used instead of full sentences. However, we show that state-of-the-art statistical machine translation (SMT) technology is in fact robust and flexible enough to capture the peculiarities of the language pair of user queries and result snippets. We evaluate our system in a comparative, extrinsic evaluation in a real-world Web search task. We compare our approach to the expansion system of Cui et al. (2002) that is trained on the same user logs data and has been shown to produce significant improvements over the local feedback technique of Xu and Croft (1996) in a standard evaluation on TREC data. Our extrinsic evaluation is done by embedding the expansion systems into a real-world search engine, and comparing the two systems based on the search results that are triggered by the respective query expansions. Our results show that the combination of translation and language model of a state-of-the-art SMT model produces high-quality rewrites and outperforms the expansion model of Cui et al. (2002).

In the following, we will discuss related work (Section 2) and quickly sketch Cui et al. (2002)'s approach (Section 3). Then we will recapitulate the essentials of

```
(AND (OR herbs herb remedies medicine supplements)
 for chronic constipation)
(AND (OR herbs spices) for mexican (OR cooking food))
```

Figure 1
Search queries *herbs for chronic constipation* and *herbs for mexican cooking* integrating expansion terms into OR-nodes in conjunctive matching.

state-of-the-art SMT and describe how to adapt an SMT system to the query expansion task (Section 4). Results of the extrinsic experimental evaluation are presented in Section 5. The presented results are based on earlier results presented in Riezler, Liu, and Vasserman (2008), and extended by deeper analyses and further experiments.

2. Related Work

Standard query expansion techniques such as local feedback, or pseudo-relevance feedback, extract expansion terms from the topmost documents retrieved in an initial retrieval round (Xu and Croft 1996). The local feedback approach is costly and can lead to query drift caused by irrelevant results in the initial retrieval round. Most importantly, though, local feedback models do not learn from data, in contrast to the approaches described in this article.

Recent research in the IR community has increasingly focused on deploying user query logs for query reformulations (Huang, Chien, and Oyang 2003; Fonseca et al. 2005; Jones et al. 2006), query clustering (Beeferman and Berger 2000; Wen, Nie, and Zhang 2002; Baeza-Yates and Tiberi 2007), or query similarity (Raghavan and Sever 1995; Fitzpatrick and Dent 1997; Sahami and Heilman 2006). The advantage of these approaches is that user feedback is readily available in user query logs and can efficiently be precomputed. Similarly to this recent work, our approach uses data from user query logs, but as input to a monolingual SMT model for learning query rewrites.

The SMT viewpoint has been introduced to the field of IR by Berger and Lafferty (1999) and Berger et al. (2000), who proposed to bridge the "lexical chasm" by a retrieval model based on IBM Model 1 (Brown et al. 1993). Since then, ranking models based on monolingual SMT have seen various applications, especially in areas like Question Answering where a large lexical gap between questions and answers has to be bridged (Berger et al. 2000; Echihabi and Marcu 2003; Soricut and Brill 2006; Riezler et al. 2007; Surdeanu, Ciaramita, and Zaragoza 2008; Xue, Jeon, and Croft 2008). Whereas most applications of SMT ideas to IR problems used translation system scores for (re)ranking purposes, only a few approaches use SMT to generate actual query rewrites (Riezler, Liu, and Vasserman 2008). Similarly to Riezler, Liu, and Vasserman (2008), we use SMT to produce actual rewrites rather than for (re)ranking, and evaluate the rewrites in a query expansion task that leaves the ranking model of the search engine untouched.

Lastly, monolingual SMT has been established in the NLP community as a useful expedient for paraphrasing, that is, the task of reformulating phrases or sentences into semantically similar strings (Quirk, Brockett, and Dolan 2004; Bannard and Callison-Burch 2005). Although the use of the SMT in paraphrasing goes beyond pure ranking to actual rewriting, SMT-based paraphrasing has to our knowledge not yet been applied to IR tasks.

3. Query Expansion by Query–Document Term Correlations

The query expansion model of Cui et al. (2002) is based on the principle that if queries containing one term often lead to the selection of documents containing another term, then a strong relationship between the two terms can be assumed. Query terms and document terms are linked via sessions in which users click on documents in the retrieval result for the query. Cui et al. define a session as follows:

session := <query text>[clicked document]*

According to this definition, a link is established if at least one user clicks on a document in the retrieval results for a query. Because query logs contain sessions from different users, an aggregation of clicks over sessions will reflect the preferences of multiple users. Cui et al. (2002) compute the following probability distribution of document words w^d given query words w^q from counts over clicked documents D aggregated over sessions:

$$P(w^d|w^q) = \sum_D P(w^d|D)P(D|w^q) \tag{1}$$

The first term in Equation (1) is a normalized *tfidf* weight of the document term in the clicked document, and the second term is the relative cooccurrence of the clicked document and query term.

Because Equation (1) calculates expansion probabilities for each term separately, Cui et al. (2002) introduce the following cohesion formula that respects the whole query Q by aggregating the expansion probabilities for each query term:

$$CoWeight_Q(w^d) = \ln\Big(\prod_{w^q \in Q} P(w^d|w^q) + 1\Big) \tag{2}$$

In contrast to local feedback techniques (Xu and Croft 1996), Cui et al. (2002)'s algorithm allows us to precompute term correlations off-line by collecting counts from query logs. This reliance on pure frequency counting is both a blessing and a curse: On the one hand it allows for efficient non-iterative estimation, but on the other hand it makes the implicit assumption that data sparsity will be overcome by counting from huge data sets. The only attempt at smoothing that is made in this approach is shifting the burden to words in the query context, using Equation (2), when Equation (1) assigns zero probability to unseen pairs. Nonetheless, Cui et al. (2002) show significant improvements over the local feedback technique of Xu and Croft (1996) in an evaluation on TREC data.

4. Query Expansion Using Monolingual SMT

4.1 Linear Models for SMT

The job of a translation system is defined in Och and Ney (2004) as finding the English string $\hat{\mathbf{e}}$ that is a translation of a foreign string $\mathbf{f}$ using a linear combination of feature functions $h_m(\mathbf{e}, \mathbf{f})$ and weights λ_m as follows:

$$\hat{\mathbf{e}} = \arg\max_e \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f})$$

As is now standard in SMT, several complex features such as lexical translation models and phrase translation models, trained in source-target and target-source directions, are combined with language models and simple features such as phrase and word counts. In the linear model formulation, SMT can be thought of as a general tool for computing string similarities or for string rewriting.

4.2 Word Alignment

The relationship of translation model and alignment model for source language string $\mathbf{f} = f_1^J$ and target string $\mathbf{e} = e_1^I$ is via a hidden variable describing an alignment mapping from source position j to target position a_j:

$$P(f_1^J|e_1^I) = \sum_{a_1^J} P(f_1^J, a_1^J|e_1^I)$$

The alignment a_1^J contains so-called null-word alignments $a_j = 0$ that align source words to the empty word.

In our approach, "sentence aligned" parallel training data are prepared by pairing user queries with snippets of search results clicked for the respective queries. The translation models used are based on a sequence of word alignment models, whereas in our case three Model-1 iterations and three HMM iterations were performed. Another important adjustment in our approach is the setting of the null-word alignment probability to 0.9 in order to account for the difference in sentence length between queries and snippets. This setting improves alignment precision by filtering out noisy alignments and instead concentrating on alignments with high support in the training data.

4.3 Phrase Extraction

Statistical estimation of alignment models is done by maximum-likelihood estimation of sentence-aligned strings $\{(\mathbf{f}_s, \mathbf{e}_s) : s = 1, \ldots, S\}$. Because each sentence pair is linked by a hidden alignment variable $\mathbf{a} = a_1^J$, the optimal $\hat{\theta}$ is found using unlabeled-data log-likelihood estimation techniques such as the EM algorithm:

$$\hat{\theta} = \arg\max_{\theta} \prod_{s=1}^{S} \sum_{\mathbf{a}} p_\theta(\mathbf{f}_s, \mathbf{a}|\mathbf{e}_s)$$

The (Viterbi-)alignment $\hat{a}_1^J$ that has the highest probability under a model is defined as follows:

$$\hat{a}_1^J = \arg\max_{a_1^J} p_{\hat{\theta}}(f_1^J, a_1^J|e_1^I)$$

Because a source–target alignment does not allow a source word to be aligned with two or more target words, source–target and target–source alignments can be combined via various heuristics to improve both recall and precision of alignments.

In our application, it is crucial to remove noise in the alignments of queries to snippets. In order to achieve this, we symmetrize Viterbi alignments for source–target and target–source directions by intersection only. That is, given two Viterbi alignments $A_1 = \{(a_j, j)|\ a_j > 0\}$ and $A_2 = \{(i, b_i)|\ b_i > 0\}$, the alignments in the intersection are defined as $A = A_1 \cap A_2$. Phrases are extracted as larger blocks of aligned words from the alignments in the intersection, as described in Och and Ney (2004).

4.4 Language Modeling

Language modeling in our approach deploys an n-gram language model that assigns the following probability to a string w_1^L of words:

$$P(w_1^L) = \prod_{i=1}^{L} P(w_i|w_1^{i-1})$$
$$\approx \prod_{i=1}^{L} P(w_i|w_{i-n+1}^{i-1})$$

Estimation of n-gram probabilities is done by counting relative frequencies of n-grams in a corpus of user queries. Remedies for sparse data problems are achieved by various smoothing techniques, as described in Brants et al. (2007).

The most important departure of our approach from standard SMT is the use of a language model trained on queries. Although this approach may seem counterintuitive from the standpoint of the noisy-channel model for SMT (Brown et al. 1993), it fits perfectly into the linear model. Whereas in the first view a query language model would be interpreted as a language model on the source language, in the linear model directionality of translation is not essential. Furthermore, the ultimate task of a query language model in our approach is to select appropriate phrase translations in the context of the original query for query expansion. This is achieved perfectly by an SMT model that assigns the identity translation as most probable translation to each phrase. Descending the n-best list of translations, in effect the language model picks alternative non-identity translations for a phrase in context of identity-translations of the other phrases.

Another advantage of using identity translations and word reordering in our approach is the fact that, by preferring identity translations or word reorderings over non-identity translations of source phrases, the SMT model can effectively abstain from generating any expansion terms. This will happen if none of the candidate phrase translations fits with high enough probability in the context of the whole query, as assessed by the language model.

5. Evaluating Query Expansion in a Web Search Task

5.1 Data

The training data for the translation model and the correlation-based model consist of pairs of queries and snippets for clicked results taken from query logs. Representing documents by snippets makes it possible to create a parallel corpus that contains data of roughly the same "sentence" length. Furthermore, this makes iterative training feasible. Queries and snippets are linked via clicks on result pages, where a parallel sentence pair is introduced for each query and each snippet of its clicked results. This yields a data set of 3 billion query–snippet pairs from which a phrase-table of 700 million query–snippet phrase translations is extracted. A collection of data statistics for the training data is shown in Table 1. The language model used in our experiment is a trigram language model trained on English queries in user logs. n-grams were cut off at a minimum frequency of 4. Data statistics for resulting unique n-grams are shown in Table 2.

Table 1
Statistics of query–snippet training data.

	query–snippet pairs	query words	snippet words
tokens	3 billion	8 billion	25 billion
avg. length	-	2.6	8.3

Table 2
Statistics of unique *n*-grams in language model.

1-grams	2-grams	3-grams
9 million	1.5 billion	5 billion

5.2 Query Expansion Setup

The setup for our extrinsic evaluation deploys a real-world search engine, google.com, for a comparison of expansions from the SMT-based system, the correlation-based system, and the correlation-based system using the language model as additional filter. All expansion systems are trained on the same set of parallel training data. SMT modules such as the language model and the translation models in source–target and target–source directions are combined in a uniform manner in order to give the SMT and correlation-based models the same initial conditions.

The expansion terms used in our experiments were extracted as follows: Firstly, a set of 150,000 randomly extracted 3+ word queries was rewritten by each of the systems. For each system, expansion terms were extracted from the 5-best rewrites, and stored in a table that maps source phrases to target phrases in the context of the full queries. For example, Table 3 shows unique 5-best translations of the SMT system for the queries *herbs for chronic constipation* and *herbs for mexican cooking*. Phrases that are newly introduced in the translations are highlighted in **boldface**. These phrases are extracted for expansion and stored in a table that maps source phrases to target phrases in the context of the query from which they were extracted. When applying the expansion

Table 3
Unique 5-best phrase-level translations of queries *herbs for chronic constipation* and *herbs for mexican cooking*. Terms extracted for expansion are highlighted in **boldface**.

(herbs , herbs) (for , for) (chronic , chronic) (constipation , constipation)
(herbs , **herb**) (for , for) (chronic , chronic) (constipation , constipation)
(herbs , **remedies**) (for , for) (chronic , chronic) (constipation , constipation)
(herbs , **medicine**) (for , for) (chronic , chronic) (constipation , constipation)
(herbs , **supplements**) (for , for) (chronic , chronic) (constipation , constipation)

(herbs , herbs) (for , for) (mexican , mexican) (cooking , cooking)
(herbs , herbs) (for , for) (cooking , cooking) (mexican , mexican)
(herbs , herbs) (for , for) (mexican , mexican) (cooking , **food**)
(mexican , mexican) (herbs , herbs) (for , for) (cooking , cooking)
(herbs , **spices**) (for , for) (mexican , mexican) (cooking , cooking)

table to the same 150,000 queries that were input to the translation, expansion phrases are included in the search query via an OR-operation. An example search query that uses the SMT-based expansions from Table 3 is shown in Figure 1.

In order to evaluate Cui et al. (2002)'s correlation-based system in this setup, we required the system to assign expansion terms to particular query terms. The best results were achieved by using a linear interpolation of scores in Equation (2) and Equation (1). Equation (1) thus introduces a preference for a particular query term to the whole-query score calculated by Equation (2). Our reimplementation uses unigram and bigram phrases in queries and expansions. Furthermore, we use *Okapi BM25* instead of *tfidf* in the calculation of Equation (1) (see Robertson, Walker, and Hancock-Beaulieu 1998).

In addition to SMT and correlation-based expansion, we evaluate a system that uses the query language model to rescore the rewrites produced by the correlation-based model. The intended effect is to filter correlation-based expansions by a more effective context model than the cohesion model proposed by Cui et al. (2002).

Because expansions from all experimental systems are done on top of the same underlying search engine, we can abstract away from interactions with the underlying system. Rewrite scores or translation probabilities were only used to create n-best lists for the respective systems; the ranking function of the underlying search engine was left untouched.

5.3 Experimental Evaluation

The evaluation was performed by three independent raters. The raters were presented with queries and 10-best search results from two systems, anonymized, and presented randomly on left or right sides. The raters' task was to evaluate the results on a 7-point Likert scale, defined as:

−1.5: much worse
−1.0: worse
−0.5: slightly worse
0: about the same
0.5: slightly better
1.0: better
1.5: much better

Table 4 shows evaluation results for all pairings of the three expansion systems. For each pairwise comparison, a set of 200 queries that has non-empty, different result lists for both systems is randomly selected from the basic set of 150,000 queries. The mean item score (averaged over queries and raters) for the experiment that compares the correlation-based model with language model filtering (corr+lm) against the correlation-based model (corr) shows a clear win for the experimental system.

Table 4
Comparison of query expansion systems on the Web search task with respect to a 7-point Likert scale.

experiment baseline	corr+lm corr	SMT corr	SMT corr+lm
mean item score	0.264 ± 0.095	0.254 ± 0.09125	0.093 ± 0.0850

An experiment that compares SMT-based expansion (SMT) against correlation-based expansions (corr) results in a clear preference for the SMT model. An experiment that compares the SMT-based expansions (SMT) against the correlation-based expansions filtered by the language model (corr+lm) shows a smaller, but still statistically significant, preference for the SMT model. Statistical significance of result differences has been computed with a paired t-test (Cohen 1995), yielding statistical significance at the 95% level for the first two columns in Table 4, and statistical significance at the 90% level for the last column in Table 4.

Examples for SMT-based and correlation-based expansions are given in Table 5. The first five examples show the five biggest wins in terms of mean item score for the SMT system over the correlation-based system. The second set of examples shows the five biggest losses of the SMT system compared to the correlation-based system. On inspection of the first set, we see that SMT-based expansions such as *henry viii restaurant portland, maine*, or *ladybug birthday ideas*, or *top ten restaurants, vancouver*, achieve a change in retrieval results that does not result in a query drift, but rather in improved retrieval results. The first and fifth result are wins for the SMT system because of nonsensical expansions by the baseline correlation-based system. A closer inspection of the second set of examples shows that the SMT-based expansion terms are all clearly related to the source terms, but not synonymous. In the first example, *shutdown* is replaced by *reboot* or *restart* which causes a demotion of the top result that matches the query exactly. In the second example, *passport* is replaced by the related term *visa* in the SMT-based expansion. The third example is a loss for SMT-based expansion because of a replacement of the specific term *debian* by the more general term *linux*. The correlation-based expansions *how many tv 30 rock* in the fourth example, and *lampasas county sheriff*

Table 5
5-best and 5-worst expansions from SMT system and corr system with mean item score.

query	SMT expansions	corr expansions	score
broyhill conference center boone	-	broyhill - welcome; boone - welcome	1.5
Henry VIII Menu Portland, Maine	menu - restaurant, restaurants	portland - six; menu - england	1.3
ladybug birthday parties	parties - ideas, party	ladybug - kids	1.3
top ten dining, vancouver	dining - restaurants	dining - 10	1.3
international communication in veterinary medicine	communication - communications, skills	international communication - college	1.3
SCRIPT TO SHUTDOWN NT 4.0	SHUTDOWN - shutdown, reboot, restart	-	−1.0
applying U.S. passport	passport - visa	applying - home	−1.0
configure debian to use dhcp	debian - linux; configure - install	configure - configuring	−1.0
how many episodes of 30 rock?	episodes - season, series	episodes - tv; many episodes - wikipedia	−0.83
lampasas county sheriff department	department - office	department - home	−0.83

home in the fifth example directly hit the title of relevant Web pages, while the SMT-based expansion terms do not improve retrieval results. However, even from these negative examples it becomes apparent that the SMT-based expansion terms are clearly related to the query terms, and for a majority of cases this has a positive effect. In contrast, the terms introduced by the correlation-based system are either only vaguely related or noise.

Similar results are shown in Table 6 where the five best and five worst examples for the comparison of the SMT model with the corr+lm model are listed. The wins for the SMT system are achieved by synonymous or closely related terms (*make - build, create; layouts - backgrounds; contractor - contractors*) or terms that properly disambiguate ambiguous query terms: For example, the term *vet* in the query *dr. tim hammond, vet* is expanded by the appropriate term *veterinarian* in the SMT-based expansion, whereas the correlation-based expansion to *vets* does not match the query context. The losses of the SMT-based system are due to terms that are only marginally related. Furthermore, the expansions of the correlation-based model are greatly improved by language model filtering. This can be seen more clearly in Table 7, which shows the five best and worst results from the comparison of correlation-based models with and without language model filtering. Here the wins by the filtered model are due to filtering non-sensical expansions or too general expansions by the unfiltered correlation-based model rather than promoting new useful expansions.

We attribute the experimental result of a significant preference for SMT-based expansions over correlation-based expansions to the fruitful combination of translation model and language model provided by the SMT system. The SMT approach can be viewed as a combined system that proposes already reasonable candidate expansions via the translation model, and filters them by the language model. We may find a certain amount of non-sensical expansion candidates at the phrase translation level of the SMT system. However, a comparison with unfiltered correlation-based expansions shows that the candidate pool of phrase translations of the SMT model is of higher quality, yielding overall better results after language model filtering. This can be seen

Table 6
5-best and 5-worst expansions from SMT system and corr+lm system with mean item score.

query	SMT expansions	corr+lm expansions	score
how to make bombs	make - build, create	make - book	1.5
dominion power va	-	dominion - virginia	1.3
purple myspace layouts	layouts - backgrounds	purple - free myspace - free	1.167
dr. tim hammond, vet	vet - veterinarian, veterinary, hospital	vet - vets	1.167
tci general contractor	contractor - contractors	-	1.167
health effects of drinking too much tea	tea - coffee	-	−1.5
tomahawk wis bike rally	-	wis - wisconsin	−1.0
apprentice tv show	-	tv - com	−1.0
super nes roms	roms - emulator	nes - nintendo	−1.0
family guy clips hitler	family - genealogy	clips - video	−1.0

Table 7
5-best and 5-worst expansions from corr system and corr+lm system with mean item score.

query	corr+lm expansions	corr expansions	score
outer cape health services	-	cape - home; health - home; services - home	1.5
Henry VII Menu Portland, Maine	-	menu - england; portland - six	1.5
easing to relieve gallbladder pain	gallbladder - gallstone	gallbladder - disease, gallstones, gallstone	1.333
guardian angel picture	-	picture - lyrics	1.333
view full episodes of naruto	episodes - watch	naruto - tv	1.333
iditarod 2007 schedule	iditarod 2007 - race	-	−1.5
40 inches plus	inches plus - review	inches - calculator	−1.333
Lovell sisters review	lovell sisters - website	-	−1.333
smartparts ion Review	smartparts ion - reviews	review - pbreview	−1.167
canon eos rebel xt slr + epinion	epinion - com	-	−1.167

from inspecting Table 8 which shows the most probable phrase translations that are applicable to the queries *herbs for chronic constipation* and *herbs for mexican cooking*. The phrase tables include identity translations and closely related terms as most probable translations for nearly every phrase. However, they also clearly include noisy and non-related terms. Thus an extraction of expansion terms from the phrase table alone would not allow the choice of the appropriate term for the given query context. This can be attained by combining the phrase translations with a language model: As shown in Table 3, the 5-best translations of the full queries attain a proper disambiguation of the senses of *herbs* by replacing the term with *remedies*, *medicine*, and *supplements* for the first

Table 8
Phrase translations for source strings *herbs for chronic constipation* and *herbs for mexican cooking*.

herbs	herbs, herbal, medicinal, spices, supplements, remedies
herbs for	herbs for, herbs, herbs and, with herbs
herbs for chronic	herbs for chronic, and herbs for chronic, herbs for
for chronic	for chronic, chronic, of chronic
for chronic constipation	for chronic constipation, chronic constipation, for constipation
chronic	chronic, acute, patients, treatment
chronic constipation	chronic constipation, of chronic constipation, with chronic constipation
constipation	constipation, bowel, common, symptoms
for mexican	for mexican, mexican, the mexican, of mexican
for mexican cooking	mexican food, mexican food and, mexican glossary
mexican	mexican, mexico, the mexican
mexican cooking	mexican cooking, mexican food, mexican, cooking
cooking	cooking, culinary, recipes, cook, food, recipe

Table 9
Correlation-based expansions for queries *herbs for chronic constipation* and *herbs for mexican cooking*.

query terms	*n*-best expansions		
herbs	com	treatment	encyclopedia
chronic	interpret	treating	com
constipation	**interpret**	treating	com
herbs for	medicinal	support	women
for chronic	com	gold	encyclopedia
chronic constipation	interpret	treating	
herbs	**cooks**	recipes	com
mexican	recipes	com	cooks
cooking	cooks	recipes	com
herbs for	medicinal	women	support
for mexican	cooks	com	allrecipes

query, and with *spices* for the second query. Table 9 shows the top three correlation-based expansion terms assigned to unigrams and bigrams in the queries *herbs for chronic constipation* and *herbs for mexican cooking*. Expansion terms are chosen by overall highest weight and shown in **boldface**. Relevant expansion terms such as *treatment* or *recipes* that would disambiguate the meaning of *herbs* are in fact in the candidate list; however, the cohesion score promotes general terms such as *interpret* or *com* as best whole-query expansions. Although language model filtering greatly improves the quality of correlation-based expansions, overall the combination of phrase translations and language model produces better results than the combination of correlation-based expansions and language model. This is confirmed by the pairwise comparison of the SMT and corr+lm systems shown in Table 4.

6. Conclusion

We presented a view of the term mismatch problem between queries and Web documents as a problem of translating from a source language of user queries to a target language of Web documents. We showed that a state-of-the-art SMT model can be applied to parallel data of user queries and snippets for clicked Web documents, and showed improvements over state-of-the-art probabilistic query expansion. Our experimental evaluation showed firstly that state-of-the-art SMT is robust and flexible enough to capture the peculiarities of query–snippet translation, thus questioning the need for special-purpose models to control noisy translations as suggested by Lee et al. (2008). Furthermore, we showed that the combination of translation model and language model significantly outperforms the combination of correlation-based model and language model. We chose to take advantage of the access the google.com search engine to evaluate the query rewrite systems by query expansion embedded in a real-word search task. Although this conforms with recent appeals for more extrinsic evaluations (Belz 2009), it decreases the reproducability of the evaluation experiment.

In future work, we hope to apply SMT-based rewriting to other rewriting tasks such as query suggestions. Also, we hope that our successful application of SMT to query expansion might serve as an example and perhaps open the doors for new applications and extrinsic evaluations of related NLP approaches such as paraphrasing.

References

Baeza-Yates, Ricardo and Alessandro Tiberi. 2007. Extracting semantic relations from query logs. In *Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'07)*, San Jose, CA, pages 76–85.

Bannard, Colin and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, MI, pages 597–604.

Beeferman, Doug and Adam Berger. 2000. Agglomerative clustering of a search engine query log. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00)*, Boston, MA, pages 407-416.

Belz, Anja. 2009. That's nice ... what can you do with it? *Computational Linguistics*, 35(1):111–118.

Berger, Adam and John Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, Berkeley, CA, pages 222-229.

Berger, Adam L., Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00)*, Athens, Greece, 192–199.

Brants, Thorsten, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'07)*, Prague Czech Republic, pages 858–867.

Brown, Peter F., 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.

Cohen, Paul R. 1995. *Empirical Methods for Artificial Intelligence*. The MIT Press, Cambridge, MA.

Cui, Hang, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. 2002. Probabilistic query expansion using query logs. In *Proceedings of the 11th International World Wide Web conference (WWW'02)*, Honolulu, HI, pages 325–332.

Echihabi, Abdessamad and Daniel Marcu. 2003. A noisy-channel approach to question answering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, Sapporo, Japan, pages 16–23.

Fitzpatrick, Larry and Mei Dent. 1997. Automatic feedback using past queries: Social searching? In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, Philadelphia, PA, pages 306–313.

Fonseca, Bruno M., Paulo Golgher, Bruno Possas, Berthier Ribeiro-Neto, and Nivio Ziviani. 2005. Concept-based interactive query expansion. In *Proceedings of the 14th Conference on Information and Knowledge Management (CIKM'05)*, Bremen, Germany, pages 696–703.

Huang, Chien-Kang, Lee-Feng Chien, and Yen-Jen Oyang. 2003. Relevant term suggestion in interactive Web search based on contextual information in query session logs. *Journal of the American Society for Information Science and Technology*, 54(7):638–649.

Jones, Rosie, Benjamin Rey, Omid Madani, and Wiley Greiner. 2006. Generating query substitutions. In *Proceedings of the 15th International World Wide Web conference (WWW'06)*, Edinburgh, Scotland, pages 387–396.

Lee, Jung-Tae, Sang-Bum Kim, Young-In Song, and Hae-Chang Rim. 2008. Bridging lexical gaps between queries and questions on large online QA collections with compact translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, Honolulu, HI, pages 410–418.

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

Quirk, Chris, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, Barcelona, Spain, pages 142–149.

Raghavan, Vijay V. and Hayri Sever. 1995. On the reuse of past optimal queries. In *Proceedings of the 18th Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95), Seattle, WA, pages 344–350.

Riezler, Stefan, Yi Liu, and Alexander Vasserman. 2008. Translating queries into snippets for improved query expansion. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08)*, Manchester, England, pages 737–744.

Riezler, Stefan, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague Czech Republic, Vol. 1, pages 464–471.

Robertson, Stephen E., Steve Walker, and Micheline Hancock-Beaulieu. 1998. Okapi at TREC-7. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, Gaithersburg, MD, pages 253–264.

Sable, Carl, Kathleen McKeown, and Kenneth W. Church. 2002. NLP found helpful (at least for one text categorization task). In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, Philadelphia, PA, pages 172–179.

Sahami, Mehran and Timothy D. Heilman. 2006. A Web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International World Wide Web conference (WWW'06)*, Edinburgh, Scotland, pages 377-386.

Soricut, Radu 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.

Surdeanu, M., M. Ciaramita, and H. Zaragoza. 2008. Learning to rank answers on large online QA collections. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, Columbus, OH, pages 719–727.

Wen, Ji-Rong, Jian-Yun Nie, and Hong-Jiang Zhang. 2002. Query clustering using user logs. *ACM Transactions on Information Systems*, 20(1):59–81.

Xu, Jinxi and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*, Zurich, Switzerland, pages 4–11.

Xue, Xiaobing, Jiwoon Jeon, and Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, Singapore, pages 475–482.