Using Key Concepts in a Translation Model for Retrieval

Jae Hyun Park and W. Bruce Croft
Center for Intelligent Information Retrieval
University of Massachusetts Amherst, Amherst, MA 01003
{ jhpark, croft } @cs.umass.edu

ABSTRACT

Many queries, especially those in the form of longer questions, contain a subset of terms representing *key concepts* that describe the most important part of the user's information need. Detecting the key concepts in a query can be used as the basis for more effective weighting of query terms, but in this paper, we focus on a method of using the key concepts in a translation model for query expansion and retrieval. Translation models have been used previously in community-based question answering (CQA) systems in order to bridge the semantic gap between questions and the corresponding answer documents. Our method uses the key concepts of a question as the translation context and selectively applies the translation model to the secondary (non-key) parts of the question. We evaluate the proposed method using a CQA collection and show that selectively translating key and secondary concepts can significantly improve the retrieval performance compared to a baseline that applies the translation model without considering key concepts.

Categories and Subject Descriptors

H.3.3 [**Information System**]: Information Search and Retrieval

Keywords

translation model, query term classification, query expansion, answer passage retrieval

1. INTRODUCTION

Statistical translation models have been used for query term expansion in a number of previous studies (e.g., [2, 10]) and have been shown to be particularly effective for retrieving answer passages in collaborative question answering (CQA) systems [12, 11]. One of the challenging issues in using translation models for expansion is incorporating more of the query context into the translation process. For example, Figure 1 shows two example questions that include the word "*grow*". The first question is about growing hair while the other question is about growing flowers. The expansion words generated from the translation for "*grow*" should be different in these two contexts.

The hypothesis that we study in this paper is that the most important, *key concepts* in a query should be treated differently in the translation model compared to secondary concepts. The key concepts of queries represent the main topics of users' information needs [1]. For example, "*hair*" and "*Columbine flower*" in Figure 1 are the key concepts. We propose a translation model for expansion that estimates the translation probabilities of query terms given the key concept of a query as the context. This will result in the translation results (and query expansions) for "*grow*" being different depending on whether "*hair*" or "*Columbine flower*" is the key concept.

In addition to the key concepts of queries, we also identify the secondary (non-key) parts of queries. In our approach, these concepts are the focus of translation (expansion) since translating the key concept may significantly change the meaning of the query. Lee et al. [10] describe an approach to identifying important terms in documents for training the translation model. In contrast, we focus on using query analysis to affect the translation results. We use the secondary concepts of queries as the focus for translation and the key concepts as context for the translation.

The rest of the paper is organized as follows. In Section 2, we describe the key concept-based translation model. Section 3 describes a method for classifying question terms for the proposed translation model. In Section 4, we show experimental results.

Q:	How do you get your *hair* to **grow** faster?
A:	Supposedly this works but never tried it. prenantal vitamins. they're just vitimans so they're not going to make u grow ...
Q:	How to **grow** *Columbine flowers*?
A:	Plant outside in sun or light shade, they will grow in both places. Scratch or loosen the soil lightly with a garden claw or rake. Sprinkle your seeds on and cover with the loose soil. You just cover with enough ...

Figure 1: Example questions containing "*grow*" in different contexts.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766XXX.XXXXXXX.

2. A KEY CONCEPT-BASED TRANSLATION MODEL

Translation models have been used as a query term expansion method for both document and answer passage retrieval [7, 11, 12]. The models estimate probabilities that terms in documents are translated into terms in queries. We employ Xue et al's translation model [12] for CQA retrieval as follows:

$$\begin{aligned} P(q_i|D) = &(1-\beta_1-\beta_2)\cdot P(q_i|D)+ \\ &\beta_1\cdot\sum_{t_j\in D, t_j\neq q_i} P(q_i|t_j)P(t_j|D)+ \\ &\beta_2\cdot\sum_{t_j\in D, t_j\neq q_i} P(t_j|q_i)P(t_j|D), \end{aligned} \quad (1)$$

where t_j is a term in a document D, $P(t_j|D)$ represents the probability that a term t_j is generated by a document D and $P(t_j|q_i)$ and $P(q_i|t_j)$ represent the translation probability that a document term t_j is translated into a query term q_i and vice versa. Note that, since we are focusing on CQA, we use "question" instead of "query" in the rest of the paper.

In this translation model, Xue et al investigated the issues of self-translation and bi-directional translation. In a translation model for a single language, every word has some probability to translate into itself. In order to prevent low self-translation probabilities from assigning low weights to matching terms and high self-translation probabilities from nullifying the effect of the translation approach, Xue et al separate self-translation from the general translation model and use the parameter β to control the impact of self-translation. They use different parameter values β_1 and β_2 according to the directions of translation $P(t_j|q_i)$ and $P(q_i|t_j)$.

In order to estimate the translation probabilities $P(t_j|q_i)$ and $P(q_i|t_j)$, IBM model 1 was used. Higher-order versions of the IBM model were proposed to take into account other features such as the order of words, the length of aligned words in source and target expressions, etc. Because the number of sentences in questions and answer passages (or documents) vary widely and the goal of the translation is to generate expansion terms, these higher-order IBM models are not appropriate for this study.

Instead of using the higher-order models, some previous studies [8, 11] have used a simpler approach in which a sentence is converted to a sequence of term pairs. For example, a sentence "*A helicopter gets its power from rotors or blades*" is converted to *(helicopter-gets) (gets-power) (power-rotors) (rotors-blades)* to estimate the translation probabilities between adjacent term pairs. Without formulating translation models for specific types of linguistic features, Surdeanu et al. model translation between different types of text representations such as bags of words, n-grams, syntactically dependent pairs of terms and predicate-argument pairs of semantic labels. We use this approach to generate a key concept-based translation table. For a given question, we assume one of the terms is the key concept of the question. We generate a set of term-pair sequences in which we treat each term in a question as the key concept of the question. For example, for a question "*A **helicopter** gets its **power** from **rotors** or **blades***", we generate four pair sequences in which each one of the question terms is used as the key concept of the question as follows:

grow	**hair**−*grow*	**PLANT**−*grow*
make	hair	plant
plant	growth	soil
healthy	biotin	cut
take	long	make
month	take	water
hair	healthy	garden
help	supplement	fruit
biotin	take	start
fast	grow	good
water	microgram	concrete

Table 1: The translation results of "*grow*" with different contexts. PLANT represents a WordNet category.

- **Q1**: *helicopter*-helicopter *helicopter*-power *helicopter*-rotors *helicopter*-blades
- **Q2**: *power*-helicopter *power*-power *power*-rotors *power*-blades
- **Q3**: *rotors*-helicopter *rotors*-power *rotors*-rotors *rotors*-blades
- **Q4**: *blades*-helicopter *blades*-power *blades*-rotors *blades*-blades

In these pair sequences, each term is concatenated with a key concept (shown in italics). From a question with n terms, we produce n pairs. Using these pair-sequence representations of questions and the original answers, we estimate the key concept-based translation table. Table 1 shows the top 10 translation terms for *grow* with different contexts. The terms in the first column, where *grow* is translated without considering context, are a mixture of terms related to various topics. The second column uses *hair* as the key concept context and contains more specific terms.

We use WordNet senses and named-entity labels to substitute for words to generate additional data to address the data sparseness problem [11, 13]. The idea is to replace terms by their categories so that we can generate more samples per category and thus obtain better estimations. For example, *Columbine* is used once in our test collection, but the WordNet sense *PLANT* is used more frequently. We can then estimate translation probabilities for *PLANT-grow* instead of *Columbine-grow*. The third column of Table 1 is the translation terms for *grow* using *PLANT* as the key concept.

We use the key concept-based translation table for the translation probability $p(t_j|q_i)$ in Eq. 1 as follows:

$$\begin{aligned} P_{mx}(q_i|D) = &(1-\beta_1-\beta_2)\cdot P(q_i|D)+ \\ &\beta_1\cdot\varphi(q_i)\cdot\sum_{t_j\in D, t_j\neq q_i} P((\kappa_Q,q_i)|t_j)P(t_j|D)+ \\ &\beta_2\cdot\varphi(q_i)\cdot\sum_{t_j\in D, t_j\neq q_i} P(t_j|(\kappa_Q,q_i))P(t_j|D), \end{aligned} \quad (2)$$

in which κ_Q represents the key concept of a query Q and $\varphi(q_i)$ is the binary function that is 1 when q_i is a secondary concept and 0 otherwise. The translation probability with key concepts $P((\kappa_Q,q_i)|t_j)$ can be interpreted in two ways. We translate a question term q_i given the key concept of a question κ_Q. Or, we repeatedly interpret the key concept of a question κ_Q for each term in the question. The binary

function for secondary concepts $\varphi(q_i)$ is used to selectively apply the translation model to question terms. In the next section, we will explain how to predict κ_Q and $\varphi(q_i)$ in detail.

3. IDENTIFYING KEY AND SECONDARY CONCEPTS

In our approach, we classify question terms as key concepts and secondary concepts. Key concepts are used as the context of translation and secondary concepts are used to selectively apply the translation model for query term expansion. For this purpose, we use a machine learning method to classify question terms [6]. For the classifier, training data consists of triplets as follows:

$$(q_1, k_1, s_1), (q_2, k_2, s_2), \ldots, (q_n, k_n, s_n),$$

in which n is the number of question terms. k_i and s_i are the labels of a question term q_i for key concepts and secondary concepts, respectively.

The definition of key concepts can differ according to their intended use. Bendersky and Croft [1] annotated the key concepts of queries to assign higher weight to the most important terms in queries. Lee et al. [10] used the TextRank algorithm in which the importance of terms is measured by the PageRank scores of terms. Lee et al [9] proposed a method to empirically select important query terms that maximizes the mean average precision of retrieval results. In our case, for the training labels of key concepts, we select k_i values according to the effectiveness of the translation model when we use q_i as the key concept of a question κ_Q. The training label of secondary concepts s_i is selected according to the improvement in retrieval effectiveness when we apply the translation model to q_i. We select only one key concept per question, although there can be more than one term which can improve the effectiveness of the key concept-based translation model.

We use three types of features for identifying key concepts and secondary concepts: lexical features, syntactic features and semantic features. The aim is to estimate how likely a given term is to be a key concept or a secondary concept given these syntactic and semantic characteristics.

Lexical Features. Lexical features are used to take account of the characteristics of an individual term.

- **Is Capitalized**: This feature is a Boolean indicator that is set to TRUE iff the first character is capitalized.
- **All Capitalized**: This feature is a Boolean indicator that is set to TRUE iff all characters are capitalized.
- **Clarity score**: This feature is the relative entropy between a term in the query language model and the collection language model [5], which indicates how important the term is in describing the topic of the query.
- **OddRatio**: The odds ratio between a given term being used in a question and the term being used in an answer. This feature is motivated by the observation that some terms in questions such as commonly-used verbs do not occur in answers.
- **Unseen**: This feature is a Boolean indicator that is set to TRUE iff a term is not observed in the top 15 retrieved documents.

Syntactic Features. Syntactic features are used to consider the role of a given term in a question.

	Prec@1	**Recall@5**	**MRR**
Baseline	0.476	0.774	0.612
TM	0.492(3.4%)	0.794(2.6%)	0.628(2.6%)
Secondary	$0.515^{\dagger}_{\ddagger}$(8.2%)	$0.818^{\dagger}_{\ddagger}$(5.7%)	$0.650^{\dagger}_{\ddagger}$(6.2%)
Key+ Secondary	$0.537^{\dagger}_{\ddagger}$(12.8%)	$0.837^{\dagger}_{\ddagger}$(8.1%)	$0.669^{\dagger}_{\ddagger}$(9.3%)

Significant differences relative to *Baseline* and *TM* are marked by † and ‡, respectively (using the two-tailed Wilcoxon test with $p < 0.05$).

Table 2: The experimental results of answer retrieval using the CQA collection. MRR represents Mean Reciprocal Rank.

- **Phrase Label**: This feature is the types of phrase that contains a given term.
- **Part-of speech (POS) tags**: Four Boolean features represent whether a question term is a noun, verb, adjective, or proper noun.
- **Depth in a parse tree**: The distance from root node to a given term in the parse tree of a question.

Semantic Features. We use WordNet sense classes and the named-entity classes of terms as semantic features [3].

4. EXPERIMENTS AND ANALYSIS

4.1 Experimental Setup

We evaluate the key concept-based translation model using the Yahoo! Answers Comprehensive Questions and Answers test collection (version 1.0[1]). We are interested in questions containing multiple concepts for which we can classify the role of these concepts for the proposed translation model. Therefore, we follow the same refinement process as [11]. Briefly, we select how-to questions containing more than four content words. Among selected 148,102 question-answer pairs, we used 30,761 question answer pairs for which the baseline system retrieved answers within the top 15 retrieval results.

We used 60% of these question-answer pairs for training, 20% for development, and 20% for testing. The training data is used to estimate the translation probabilities and to train the key concept classification method. The development data is used to select optimal parameter settings of β_1, β_2 in Eq 2. We indexed answers as documents. Then, we retrieved answers by submitting the questions as queries, using the Galago toolkit [4] for indexing and retrieval, and the sequential dependence model as the baseline. The SuperSenseTag software [3] was used to annotate WordNet categories for question terms.

4.2 Experimental Results

Our experiments evaluated the effectiveness of using key concepts and secondary concepts in a translation model for query expansion in answer retrieval. Table 2 shows the experimental results for precision at rank 1, recall at rank 5 and the Mean Reciprocal Rank (MRR) of answers in retrieval results. The baseline is the sequential dependence model without query expansion. Among 8,715 question-answer pairs in

[1]http://webscope.sandbox.yahoo.com/

	Unchanged	Improved	Decreased
Secondary	6,759	1,521	435
Key+ Secondary	5,307	2,646	762

Table 3: The number of question-answer pairs for which retrieval results are unchanged, improved, and decreased by the translation-based model with key and secondary concepts.

the test data, the baseline system retrieved the answers at the first rank for 4,150 (47.6%) of questions. As another baseline, we use the translation-based model from Eq. 1 (*TM*) that does not use key concepts. We do not have a pseudo-relevance feedback baseline (such as the RM3 model provided in Galago) because the Xue et al model has already been shown to be superior to pseudo-relevance feedback for this type of data [12].

Secondary is the experimental results of the translation-based model that applies the translation only for the secondary concepts of question terms. By translating only secondary keywords in questions, we potentially reduce the number of non-relevant translation results. *Key+ Secondary* shows the results when we translate the secondary concepts of questions using the key concepts as context.

As we can see, considering key concepts as the context for translation significantly improves the performance of the system. To analyze this further, we compare the experimental results of the baseline system and the translation-based model for individual questions. Table 3 shows the number of question-answer pairs for which retrieval results are unchanged, improved and decreased by using key concepts and secondary concepts for the translation model. The translation-based model without the key concepts affects fewer retrieval results. This model introduces expansion terms related to a range of possible contexts, which consequently has less effect on ranking. Using the predicted key concepts as context, the translation model generates more precise translation results. However, the ratio of questions for which results were decreased by the translation-based model with key and secondary concepts is also higher than when only using the secondary concepts. This shows that, if the selection of key concepts is inaccurate, using them as context can have a negative impact on effectiveness.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed the key concept-based translation model for query expansion. Key concepts represent the most important part of the users' information needs. We use the key concepts of questions as the context of translation. In addition, we also classify question terms as secondary concepts to selectively apply the translation model. Key concepts can improve the effectiveness of the query expansion by constraining the translations of question terms within the contexts of questions. By translating only secondary concepts, we can also reduce the non-relevant translation results. The key concept-based translation model significantly improved the effectiveness of translation-based query expansion for finding answers in a CQA collection.

For future work, we plan to apply the proposed method to passage retrieval in documents. Because of the lack of training data, previous work on using a translation model for retrieval has used pseudo data such as synthesized queries for given documents [7].

6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF Grant IIS-1419693. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *Proceedings of the ACM SIGIR conference*, pages 491–498. ACM, 2008.

[2] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–199. ACM, 2000.

[3] M. Ciaramita and Y. Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of EMNLP*, pages 594–602, 2006.

[4] W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*. 2010.

[5] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. In *Proceedings of HLT*, pages 104–109, 2002.

[6] T. Joachims. *Learning to classify text using support vector machines: methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

[7] M. Karimzadehgan and C. Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceedings of the ACM SIGIR conference*, pages 323–330. ACM, 2010.

[8] K. Kishida and E. Ishita. Translation disambiguation for cross-language information retrieval using context-based translation probability. *Journal of Information Science*, 2009.

[9] C.-J. Lee, R.-C. Chen, S.-H. Kao, and P.-J. Cheng. A term dependency-based approach for query terms ranking. In *Proceedings of the CIKM*, pages 1267–1276. ACM, 2009.

[10] J.-T. Lee, S.-B. Kim, Y.-I. Song, and H.-C. Rim. Bridging lexical gaps between queries and questions on large online q&a collections with compact translation models. In *Proceedings of EMNLP*, pages 410–418. Association for Computational Linguistics, 2008.

[11] M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics*, 37(2):351–383, 2011.

[12] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In *Proceedings of the ACM SIGIR conference*, pages 475–482. ACM, 2008.

[13] G. Zhou, Y. Liu, F. Liu, D. Zeng, and J. Zhao. Improving question retrieval in community question answering using world knowledge. In *Proceedings of the IJCAI*, pages 2239–2245. AAAI Press, 2013.