THUTR: A Translation Retrieval System

Chunyang Liu*, Qi Liu*, Yang Liu, and Maosong Sun
Department of Computer Science and Technology
State Key Lab on Intelligent Technology and Systems
National Lab for Information Science and Technology
Tsinghua University, Beijing 100084, China
{liuchunyang2012,flaminglq,liuyang.china,sunmaosong}@gmail.com

ABSTRACT

We introduce a translation retrieval system *THUTR*, which casts translation as a retrieval problem. Translation retrieval aims at retrieving a list of target-language translation candidates that may be helpful to human translators in translating a given source-language input. While conventional translation retrieval methods mainly rely on parallel corpus that is difficult and expensive to collect, we propose to retrieve translation candidates directly from target-language documents. Given a source-language query, we first translate it into target-language queries and then retrieve translation candidates from target language documents. Experiments on Chinese-English data show that the proposed translation retrieval system achieves 95.32% and 92.00% in terms of P@10 at sentence level and phrase level tasks, respectively. Our system also outperforms a retrieval system that uses parallel corpus significantly.

TITLE AND ABSTRACT IN CHINESE

THUTR：一个译文检索系统

我们介绍一个译文检索系统THUTR。该系统将翻译视作为一个检索问题。译文检索旨在为输入的源语言文本搜索一组候选翻译，为翻译人员提供帮助。传统的译文检索方法主要基于双语语料库，其面临的主要问题是构建双语语料库代价高昂。为此，我们提出使用单语语料库实现译文检索。给定源语言查询，我们首先将其翻译成目标语言查询，然后再从单语语料库中检索候选译文。在汉英数据上的实验表明，我们提出的译文检索系统分别在句子级和短语级取得95.32%和92.00%的P@10值。我们的系统也显著超过了使用平行语料库的译文检索系统。

KEYWORDS: Translation retrieval, monolingual corpus, statistical machine translation.

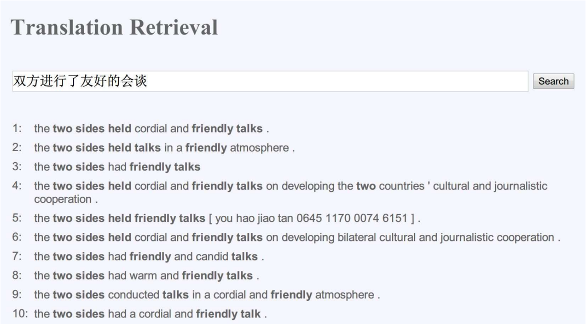
KEYWORDS IN CHINESE: 译文检索；单语语料库；统计机器翻译.

*Chunyang Liu and Qi Liu have equal contribution to this work

Proceedings of COLING 2012: Demonstration Papers, pages 321–328,
COLING 2012, Mumbai, December 2012.

1 Introduction

This demonstration introduces a **translation retrieval** system *THUTR*, which combines machine translation and information retrieval to provide useful information to translation users. Unlike machine translation, our system casts translation as a retrieval problem: given a source-language string, returns a list of ranked target-language strings that contain its (partial) translations from a large set of target-language documents.



双方进行了友好的会谈

1: the **two sides held** cordial and **friendly talks** .
2: the **two sides held talks** in a **friendly** atmosphere .
3: the **two sides** had **friendly talks**
4: the **two sides held** cordial and **friendly talks** on developing the **two** countries ' cultural and journalistic cooperation .
5: the **two sides held friendly talks** [you hao jiao tan 0645 1170 0074 6151] .
6: the **two sides held** cordial and **friendly talks** on developing bilateral cultural and journalistic cooperation .
7: the **two sides** had **friendly** and candid **talks** .
8: the **two sides** had warm and **friendly talks** .
9: the **two sides** conducted **talks** in a cordial and **friendly** atmosphere .
10: the **two sides** had a cordial and **friendly talk** .

Figure 1: A screenshot of the translation retrieval system.

For example, as shown in Figure 1, given a Chinese query, our system searches for its translations in a large set of English documents and returns a list of ranked sentences. [1] Although the retrieved documents might not be exact translations of the query, they often contain useful partial translations to help human translators produce high-quality translations. As the fluency of target-language documents are usually guaranteed, the primary goal of translation retrieval is to find documents that are *relevant* to the source-language query. By relevant, we mean that retrieved documents contain (partial) translations of queries.

Our system is divided into two modules:

1. *machine translation module*: given a source-language query, translates it into a list of translation candidates;

2. *information retrieval module*: takes the translation candidates as target-language queries and retrieves relevant documents.

Generally, the MT module ensures the fidelity of retrieved documents and the IR module ensures the fluency. Therefore, the relevance can be measured based on the coupling of translation and retrieval models.

We evaluate our system on Chinese-English data. Experiments show that our translation retrieval system achieves 95.32% and 92.00% in terms of P@10 at sentence level and phrase level tasks, respectively. Our system also significantly outperforms a retrieval system that uses parallel corpus.

[1] In our system, each document is a single sentence.

2 Related Work

Translation retrieval is firstly introduced in translation memory (TM) systems (Baldwin and Tanaka, 2000; Baldwin, 2001). Translation equivalents of maximally similar source language strings are chosen as the translation candidates. This is similar with example-based machine translation (Nagao, 1984) that translates by analogy based on parallel corpus. Unfortunately, these systems suffer from a major drawback: the amount and domain of parallel corpus are relatively limited. Hong et al. (2010) propose a method for mining parallel data from the Web. They focus on parallel data mining and report significant improvements on MT experiments.

Many researchers have explored the application of MT techniques to information retrieval tasks. Berger and Lafferty (1999) introduce a probabilistic approach to IR based on statistical machine translation models. Federico and Bertoldi (2002) divide CLIR into two translation models: a query-translation model and a query-document model. They show that offering more query translations improves retrieval performance. Murdock and Croft (2005) propose a method for sentence retrieval but in a monolingual scenario. They incorporate a machine translation in two steps: estimation and ranking. Sanchez-Martinez and Carrasco (2011) investigate how to retrieve documents that are plausible translations of a given source language document using statistical machine translation techniques.

3 System Description

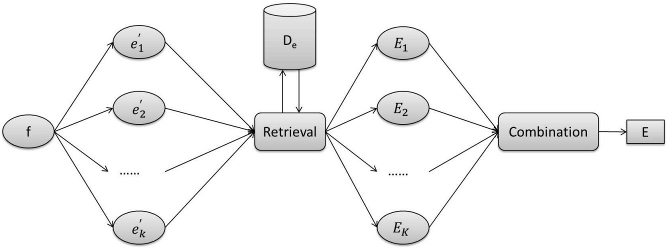


Figure 2: System architecture.

Given a source language query f, conventional translation retrieval systems (Baldwin and Tanaka, 2000; Baldwin, 2001) search for a source string $\hat{f}$ that has the maximal similarity with f in a parallel corpus $D_{f,e} = \{(f_1, e_1), \ldots, (f_N, e_N)\}$:

$$\hat{f} = \underset{f' \in D_{f,e}}{\arg\max} \Big\{ sim(f, f') \Big\} \tag{1}$$

where $sim(f, f')$ calculates the similarity between two source language strings f and f'. Then, retrieval systems return the target language string $\hat{e}$ corresponding to $\hat{f}$ in $D_{f,e}$. Therefore, conventional translation retrieval systems only rely on source language string matching and do not actually consider the translation probability between f and $\hat{e}$.

Alternatively, given a source language query f, our translation retrieval system searches for a target string $\hat{e}$ that has the maximal translation probability with f in a monolingual corpus $D_e = \{e_1, \ldots, e_M\}$:

$$\hat{e} = \underset{e \in D_e}{\arg\max} \Big\{ P(e|f) \Big\} \tag{2}$$

where $P(e|f)$ is the probability that e contains a translation of f.

This problem definition is similar with cross-lingual information retrieval (CLIR) (Ballesteros and Croft, 1997; Nie et al., 1999) except for the relevance judgement criterion. While CLIR requires the retrieved documents to be relevant to users' information need, translation retrieval expects to return documents containing the translations of queries. For example, given a Chinese query "奥运会" (i.e., "*Olympic Games*"), both "*Olympic Games*" and "*London*" are relevant in CLIR. However, in translation retrieval, only "*Olympic Games*" is relevant. Therefore, translation retrieval can be seen as a special case of CLIR.

The translation probability $P(e|f)$ can be further decomposed by introducing a target language query e' as a hidden variable:

$$P(e|f) = \sum_{e'} P(e, e'|f) = \sum_{e'} P(e'|f) \times P(e|e') \tag{3}$$

where $P(e'|f)$ is a **translation model** and $P(e|e')$ is a **retrieval model**. [2]

In this work, we use a phrase-based model as the translation model. Phrase-based models (Och and Ney, 2002; Marcu and Wong, 2002; Koehn et al., 2003) treat phrase as the basic translation unit. Based on a log-linear framework (Och and Ney, 2002), the phrase-based model used in our system is defined as

$$\begin{aligned} P(e'|f) &= \frac{score(e', f)}{\sum_{e'} score(e', f)} & (4) \\ score(e', f) &= P_{\phi}(e'|f)^{\lambda_1} \times P_{\phi}(f|e')^{\lambda_2} \times P_{lex}(e'|f)^{\lambda_3} \times P_{lex}(f|e')^{\lambda_4} \times \\ & \quad exp(1)^{\lambda_5} \times exp(|e'|)^{\lambda_6} \times P_{lm}(e')^{\lambda_7} \times P_d(f, e')^{\lambda_8} & (5) \end{aligned}$$

The λ's are feature weights that can be optimized using the minimum error rate training algorithm (Och, 2003).

We use the vector space model for calculating the cosine similarity of a target language query e' and a target language document e.

Therefore, the decision rule for translation retrieval is [3]

$$\begin{aligned} \hat{e} &= \operatorname*{argmax}_{e} \Big\{ \sum_{e'} P(e'|f) \times P(e|e') \Big\} & (6) \\ &\approx \operatorname*{argmax}_{e, e'} \Big\{ score(e', f) \times sim(e', e) \Big\} & (7) \end{aligned}$$

The pipeline of our translation retrieval system is shown in Figure 2. Given a source language query f, the system first obtains a K-best list of target language query candidates: $e'_1, e'_2, \ldots, e'_K$. For each target language query e'_k, the system returns a list of translation candidates E_k. These translation candidates are merged and sorted according to Eq. (7) to produce the final ranked list E.

[2] Such "query translation" framework has been widely used in CLIR (Nie et al., 1999; Federico and Bertoldi, 2002) In this work, e' is a translation of f produced by MT systems and e is a target language document that probably contains a translation of f. While e' is usually ungrammatical and erroneous, e is often written by native speakers.

[3] In practice, we add a parameter λ_9 to Eq (12) to achieve a balance between translation model and retrieval model: $score(e', f) \times sim(e', e)^{\lambda_9}$. We set $\lambda_9 = 2$ in our experiments.

4 Experiments

We evaluated our translation retrieval system on Chinese-English data that contains $2.2M$ sentence pairs. They are divided into three parts:

1. Training set ($220K$): training phrase-based translation model and feature weights.

2. Query set ($5K$): source language queries paired with their translations.

3. Document set ($1.99M$): target language sentences paired with their corresponding source language sentences.

The statistical machine translation system we used is Moses (Koehn et al., 2007) with its default setting except that we set the maximal phrase length to 4. For language model, we used SRI Language Modeling Toolkit (Stolcke, 2002) to train a 4-gram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998). Our retrieval module is based on Apache Lucene, the state-of-the-art open-source search software. [4]

4.1 Sentence Level

We first evaluated our system at sentence level: treating a source language sentence as query. Given the bilingual query set $\{(f_1^q, e_1^q), \ldots, (f_Q^q, e_Q^q)\}$ and the bilingual document set $\{(f_1^d, e_1^d), \ldots, (f_D^d, e_D^d)\}$, we treat $\{f_1^q, \ldots, f_Q^q\}$ as the query set and $\{e_1^q, \ldots, e_Q^q, e_1^d, \ldots, e_D^d\}$ as the document set. Therefore, relevance judgement can be done automatically because the gold standard documents (i.e., $\{e_1^q, \ldots, e_Q^q\}$) are included in the document set. The evaluation metric is $P@n$, $n = 1, 10, 20, 50$.

phrase length	BLEU	P@1	P@10	P@20	P@50
1	11.75	83.12	89.38	90.80	92.52
2	24.15	89.54	93.66	94.36	95.08
3	27.48	90.46	94.26	94.98	95.52
4	28.37	90.62	94.44	95.28	95.86
5	30.12	90.50	94.14	94.98	95.66
6	29.52	90.40	94.42	94.96	95.56
7	29.96	90.62	94.40	95.10	95.74

Table 1: Effect of maximal phrase length.

Table 1 shows the effect of maximal phrase length on retrieval performance. Beside retrieval precisions, Table 2 also lists the BLEU scores of query translation. Retrieval performance generally rises with the increase of translation accuracy. Surprisingly, our system achieves over 90% in terms of $P@1$ when the maximal phrase length is greater than 2. This happens because it is much easier for translation retrieval to judge relevance of retrieved documents than conventional IR systems. We find that the performance hardly increase when the maximal phrase length is greater than 3. This is because long phrases are less likely to be used to translate unseen text. As a result, extracting phrase pairs within 4 words is good enough to achieve reasonable retrieval performance.

Table 2 shows the effect of the K-best list translations output by the MT system. It can be treated as "query expansion" for translation retrieval, which proves to be effective in other IR

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

	P@1	P@10	P@20	P@50
1-best	90.62	94.44	95.28	95.86
10-best	91.88	95.32	96.16	96.84

Table 2: Effect of K-best list translations.

systems. We find that providing more query translations to the retrieval module does improve translation quality significantly.

system	P@1	P@10	P@20	P@50
parallel	34.48	58.62	63.98	70.88
monolingual	69.04	81.17	83.68	87.45

Table 3: Comparison to translation retrieval with parallel corpus.

We also compared our system with a retrieval system that relies on parallel corpus. Table 3 shows the results for retrieval systems using parallel and monolingual corpora, respectively. Surprisingly, our system significantly outperforms retrieval system using parallel corpus. We notice that Baldwin (2001) carefully chose data sets in which the language is controlled or highly constrained. In other words, a given word often has only one translation across all usages and syntactic constructions are limited. This is not the case for our datasets. Therefore, it might be problematic for conventional retrieval systems to calculate source language string similarities. This problem is probably alleviated in our system by providing multiple query translations.

4.2 Phrase Level

	P@1	P@10
1-best	64.00	88.00
10-best	72.00	92.00

Table 4: Results of phrase level retrieval.

Finally, we evaluated our system at phrase level: a source language phrase as query. We selected 50 phrases from the source language query set. The average length of a phrasal query is 2.76 words. On average, a query occurs in the source language query set for 3 times. Given a source query, our system returns a list of ranked target language documents. Relevance judgement was performed manually. Table 5 shows the results of phrase level evaluation. We compared between using 1-best translations and 10-best translations produced by MT systems. The $P@10$ for using 10-best translations reaches 92%.

Conclusion

We have presented a system *THUTR* that retrieves translations directly from monolingual corpus. Experiments on Chinese-English data show that our retrieval system achieves 95.32% and 92.00% in terms of $P@10$ for sentence level and phrase level queries, respectively. In the future, we would like to extend our approach to the web that provides enormous monolingual data. In addition, we plan to investigate more accurate retrieval models for translation retrieval.

Acknowledgments

This work is supported by NSFC project No. 60903138, 863 project No. 2012AA011102, and a Research Fund No. 20123000007 from Microsoft Research Asia.

References

Baldwin, T. (2001). Low-cost, high-performance translation retrieval: Dumber is better. In *ACL 2001*.

Baldwin, T. and Tanaka, H. (2000). The effects of word order and segmentation on translation retrieval performance. In *COLING 2000*.

Ballesteros, L. and Croft, B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *SIGIR 1997*.

Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. In *SIGIR 1999*.

Chen, S. F. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical report, Harvard University Center for Research in Computing Technology.

Federico, M. and Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *SIGIR 2002*.

Hong, G., Li, C.-H., Zhou, M., and Rim, H.-C. (2010). An empirical study on web mining of parallel data. In *COLING 2010*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL 2007 (the Demo and Poster Sessions)*.

Koehn, P., Och, F., and Marcu, D. (2003). Statistical phrase-based translation. In *NAACL 2003*.

Marcu, D. and Wong, D. (2002). A phrase-based, joint probability model for statistical machine translation. In *EMNLP 2002*.

Murdock, V. and Croft, B. (2005). A translation model for sentence retrieval. In *EMNLP 2005*.

Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. In Elithorn, A. and Banerji, R., editors, *Artificial and Human Intelligence*. North-Holland.

Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *SIGIR 1999*.

Och, F. (2003). Minimum error rate training in statistical machine translation. In *ACL 2003*.

Och, F. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *ACL 2002*.

Sanchez-Martinez, F. and Carrasco, R. (2011). Document translation retrieval based on statistical machine translation techniques. *Applied Artificial Intelligence*, 25(5).

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In *ICSLP 2002*.