Postech's System Description for Medical Text Translation Task

Jianri Li **Se-Jong Kim** **Hwidong Na** **Jong-Hyeok Lee**
Department of Computer Science and Engineering
Pohang University of Science and Technology, Pohang, Republic of Korea
{skywalker, sejong, leona, jhlee}@postech.ac.kr

Abstract

This short paper presents a system description for intrinsic evaluation of the WMT 14's medical text translation task. Our systems consist of phrase-based statistical machine translation system and query translation system between German-English language pairs. Our work focuses on the query translation task and we achieved the highest BLEU score among the all submitted systems for the English-German intrinsic query translation evaluation.

1 Overview

The goal of WMT14's medical text translation task is investigation of capability of machine translation (MT) technologies when it is applied to translating texts and query terms in medical domain. In our work, we focus on its application on cross-lingual information retrieval (CLIR) and evaluation of query translation task.

CLIR techniques aim to increase the accessibility of web documents written by foreign language. One of the key techniques of cross-lingual IR is query translation, which aims to translate the input query into relevant terms in target language.

One way to translate queries is dictionary-based query translation. However, an input query usually consists of multiple terms, which cause low coverage of bilingual dictionary. Alternative way is translating queries using statistical machine translation (SMT) system. However, translation model could contain some noise that is meaningless translation. The goal of our method is to overcome the shortcomings of these approaches by a heuristic hybrid approach.

As a baseline, we use phrase-based statistical machine translation (PBSMT) (Koehn, Och, & Marcu, 2003) techniques to handle queries that consist of multiple terms. To identify multiple terms in a query, we analyze three cases of the formation of queries and generate query translation candidates using term-to-term dictionaries and PBSMT system, and then score these candidates using *co-occurrence word frequency measure* to select the best candidate.

We have done experiment on two language pairs

- English-German
- German-English

The rest of parts in this paper are organized as following: section 2 describes the techniques and system settings used in our experiment, section 3 presents used corpus and experiment result, and section 4 shows a brief conclusion of our work.

2 Method

2.1 Phrase-based machine translation system

The phrase-based statistical machine translation system is implemented using MOSE'S toolkits (Koehn et al., 2007). Bidirectional word alignments were built by MGIZA [1], a multi-thread version of GIZA++ (Och & Ney, 2003), run on a 24 threads machine. The alignment symmetrization method is *grow-diag-final-and* (Koehn et al., 2003), and lexicalized-reordering method is *msd-bidirectional-fe* (Koehn et al., 2007).

For each monolingual corpus, we used a five-gram language model, which was built by IRSTLM toolkit[2] (Federico, Bertoldi, & Cettolo, 2008) with improved Kneser Ney smoothing (Chen & Goodman, 1996; Kneser & Ney, 1995). The language model was integrated as a log-linear feature to decoder.

All the sentences in the training, development and test corpus were tokenized by inserting spaces between words and punctuations, and then converted to most probable cases by truecaseing. Both tokenization and truecasing were done by embedded tools in the MOSE'S toolkits. Finally, all the sentences in the train corpus were cleaned with maximum length 80.

[1] http://www.kyloo.net/software
[2] http://sourceforge.net/projects/irstlm

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 229–232,
Baltimore, Maryland USA, June 26–27, 2014.

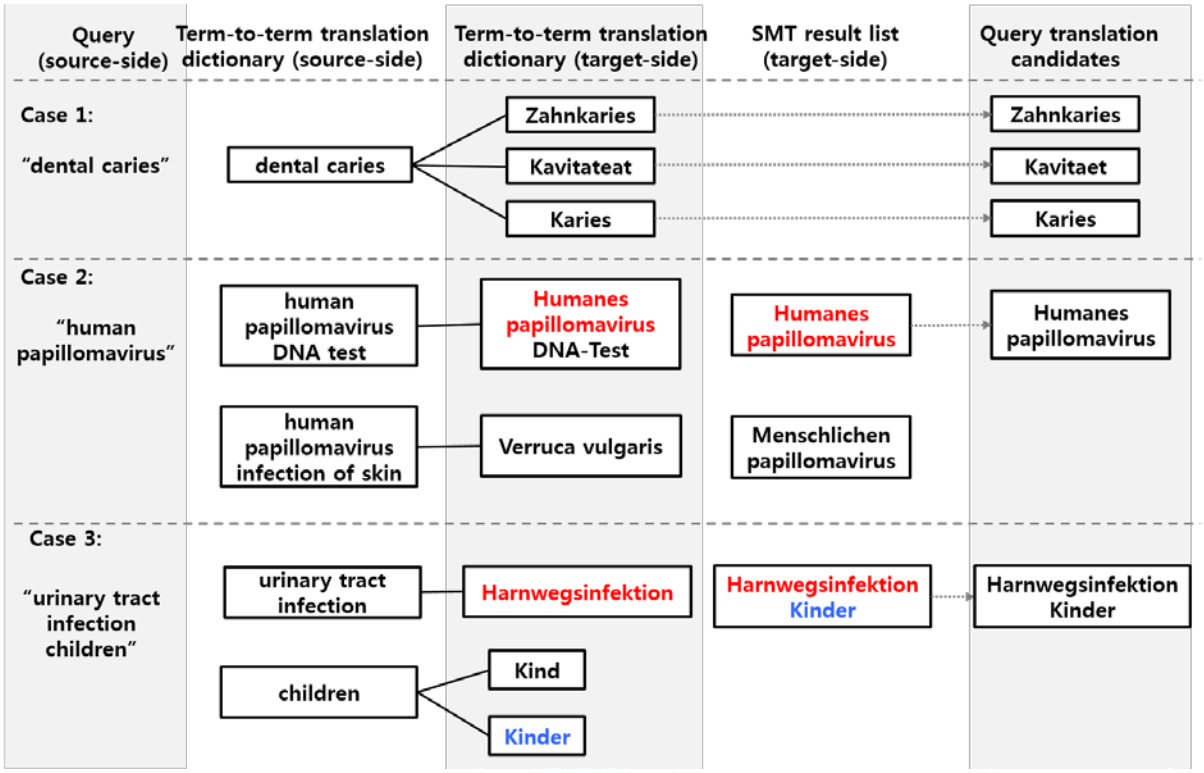

Figure 1. Flow from queries to query translation candidates for each case.

2.2 Query translation system

In general, an input query is not a full sentence. Instead, most of queries contain one or more phrases that consist of several keywords. Furthermore, in the medical domain, many keywords are unfamiliar terminologies for general users. Therefore, term-to-term translation dictionaries in medical domain could be useful resources to translate the queries. In our experiment, we used the parallel terms from Unified Medical Language System (UMLS) and titles of Wikipedia in medical domain, as the term-to-term translation dictionary.

First of all, if a given query is a combination of two or more phrases that concatenated by terms like comma, coordinate conjunction, then the given query is divided into several single phrases, and each of them is translated by our SMT system as a new single query. If the new query satisfies one of cases shown in Figure 1, then its query translation candidates are selected according to the corresponding case, and select the best one of them using proposed measures. Otherwise, if the new query does not satisfy any case, the top 1 result by our PBSMT system is selected as the best query translation candidate. Our method combines the translation results of single queries by following rules: 1) if the original query consists of multiple phrases concatenated by functional words like coordinate conjunctions, then the translation results are combined by translated functional words, 2) if the original query is concatenated by punctuation, then the results are combined by the original punctuation. Finally, the final result is selected by comparing the result from QT system and PBSMT system using the co-occurrence word frequency measure (see Section 2.2.4). The following three subsections describe how we select translation candidate case by case.

2.2.1 Case 1: Full matching

If a single query exactly matches one instance in the dictionary, query translation candidates are the target-side entries in the translation dictionary (Case 1 in Figure 1). If a query translation candidate *qt* is a sequence of words (w_1 to w_n), it is ranked by the co-occurrence word frequency measure (*CF*) using the provided articles of Wikipedia in the medical domain:

$$CF(qt) = \frac{freq(w_1)}{N_{uni}} \prod_{i=2}^{n} \frac{\frac{freq(w_i, w_{i-1})}{N_{bi}}}{\frac{freq(w_{i-1})}{N_{uni}}}, \quad (1)$$

where $freq(w_1)$ is the frequency of a unigram w_1 in the articles; $freq(w_i, w_{i-1})$ is the frequency of a

bigram "$w_i\ w_{i-1}$" in the articles; and N_{uni} and N_{bi} is the sum of frequency of all unigram and bigram, respectively.

2.2.2 Case 2: Full inclusion

If a source-side entry of the term-to-term translation dictionary exactly includes a query, its query translation candidate is its SMT result whose all words appear in the target-side entry of the translation dictionary (Case 2 in Figure 1). Among the top 10 results by our PBSMT system, we select the results satisfying this case, and rank them using *CF* and our PBSMT result score ($Score_{SMT}$):

$$Score_{QT}(qt) = \lambda \frac{CF(qt)}{\sum_{qt' \in QT} CF(qt')} + (1-\lambda) \frac{Score_{SMT}(qt)}{\sum_{qt' \in QT} Score_{SMT}(qt')} \quad , \quad (2)$$

where λ is the weight by the provided development set; and *QT* is the set of query translation candidates for a query.

2.2.3 Case 1: Full matching

If the left phrase t_{left} or right phrase t_{right} of a query exactly matches one instance in the dictionary, its query translation candidate is its SMT result that includes all words in the target-side entry of the translation dictionary (Case 3 in Figure 1). To rank our SMT results satisfying this case, if the total number of words in t_{left} and t_{right} is same or larger than that in a query, $Score_{QT}$ is used, and the other case uses the weighted $Score_{QT}$ ($WScore_{QT}$):

$$WScore_{QT}(qt) = \frac{N(t_{left}) + N(t_{right})}{N(q)} Score_{QT}(qt) \quad , \quad (3)$$

where $N(t_{left})$ is the number of words in t_{left}; and q is a given query.

2.2.4 Select final result

If a query satisfies any case above, and the candidate with highest score is selected, then we compare the candidate with translation of original query directly obtained from PBSMT system using equation (1). The final result would be the result with higher score between them.

3 Experiment

3.1 Corpus

We only use constrained data provided by WMT 2014 medical translation task.

To train PBSMT system, we use parallel corpora

- EMEA
- MuchMore
- Wikipedia-titles
- Patent-abstract, claim, title
- UMLS

We simply mixed up all available parallel corpora to train a unique translation model.

And for English-German language pair we use monolingual corpora

- Wikipedia-articles
- Patent-descriptions
- UMLS descriptions

And for German-English language pair we use monolingual corpora

- Wikipedia-articles
- Patent-descriptions
- UMLS descriptions
- AACT
- GENIA
- GREC
- FMA
- PIL

We also use target side of parallel corpora as additional monolingual resource to train language model. We separately train a 5-gram language model for each monolingual corpus and integrate them as features to log-linear model in the PBSMT system.

For the query translation (QT) system, we use parallel corpus *Wikipedia-titles* and *UMLS dictionary*, and use monolingual corpus *Wikipedia-articles*.

3.2 Experiment Setting

For the tuning of PBSMT system, we use development set provided by WMT 14 medical task (*khresmoi-summary-dev*). And we use query translation development set (*khresmoi-query-dev*) for the tuning of QT system.

We test our systems on two test set provided by WMT 14 medical task.

- khresmoi-summary-test (for PBSMT)
- khresmoi-query-test (for QT)

For comparison with result from QT system, we translate the test set of query translation task (*khresmoi-query-test*) using PBSMT system without any post-processing.

In our experiment, the performance of translation system is measured by BLEU (%) and translation error rate - TER (%). All these results are evaluated from the evaluation website[3].

3.3 Experiment Result

Table 1 shows the results for the task of translation of sentences from summaries of medical articles.

Table 2 shows the results for the task of translation of queries entered by users of medical information search engines. The performance of QT system is relatively higher than PBSMT system. Especially, the BLEU score of QT system on English-German language pair is the highest score among the all submitted systems.

Language Pair	BLEU	TER
English-German	15.8	0.746
German-English	26.9	0.618

Table 1: BLEU scores of result from PBSMT system for summary translation task.

Language Pair		BLEU	TER
PBSMT	English-German	15.1	0.748
	German-English	22.1	0.638
QT	English-German	**15.3**	**0.746**
	German-English	**24.5**	**0.586**

Table 2: BLEU scores of result for query translation task.

4 Conclusion

We describe the PBSMT system and QT system that are developed for summary translation and query translation of WMT 14 medical translation task. We focus on intrinsic query translation evaluation and propose a hybrid approach by combining dictionary-based approach and SMT based approach using heuristics. The result of query translation experiment shows that our method obtained higher translation accuracy than the baseline (PBSMT) system.

Acknowledgments

This work was supported in part by the National Korea Science and Engineering Foundation (KOSEF) (NRF-2010-0012662), in part by the Brain Korea 21+ Project, and in part by the Korea Ministry of Knowledge Economy (MKE) under Grant No.10041807.

References

Chen, S. F., & Goodman, J. (1996). *An empirical study of smoothing techniques for language modeling.* Paper presented at the Proceedings of the 34th annual meeting on Association for Computational Linguistics.

Federico, M., Bertoldi, N., & Cettolo, M. (2008). *IRSTLM: an open source toolkit for handling large scale language models.* Paper presented at the Interspeech.

Kneser, R., & Ney, H. (1995). *Improved backing-off for m-gram language modeling.* Paper presented at the Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., . . . Herbst, E. (2007). *Moses: open source toolkit for statistical machine translation.* Paper presented at the Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Prague, Czech Republic.

Koehn, P., Och, F. J., & Marcu, D. (2003). *Statistical phrase-based translation.* Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist., 29*(1), 19-51. doi: 10.1162/089120103321337421

[3] http://matrix.statmt.org