The DCU Terminology Translation System for the Medical Query Subtask at WMT14

Tsuyoshi Okita, Ali Hosseinzadeh Vahid, Andy Way, Qun Liu
Dublin City University, School of Computing
Glasnevin, Dublin 9
Ireland
{tokita,avahid,away,qliu}@computing.dcu.ie

Abstract

This paper describes the Dublin City University terminology translation system used for our participation in the query translation subtask in the medical translation task in the Workshop on Statistical Machine Translation (WMT14). We deployed six different kinds of terminology extraction methods, and participated in three different tasks: FR–EN and EN–FR query tasks, and the CLIR task. We obtained 36.2 BLEU points absolute for FR–EN and 28.8 BLEU points absolute for EN–FR tasks where we obtained the first place in both tasks. We obtained 51.8 BLEU points absolute for the CLIR task.

1 Introduction

This paper describes the terminology translation system developed at Dublin City University for our participation in the query translation subtask at the Workshop on Statistical Machine Translation (WMT14). We developed six kinds of terminology extraction methods for the problem of medical terminology translation, especially where rare and new words are considered. We have several motivations which we address before providing a description of the actual algorithms undeprinning our work.

First, terminology translation cannot be seen just as a simple extension of the translation process if we use an analogy from human translation. Terminology translation can be considered as more important and a quite different task than translation *per se*, so we need a considerably different way of solving this particular problem. Bilingual terminology selection has been claimed to be the touchstone in human translation, especially where scientific and legal translation are concerned. Terminology selection is often the hardest and most time-consuming process in the translation workflow. Depending on the particular requirements of the use-case (Way, 2013), users may not object to disfluent translations, but will invariably be very sensitive to the wrong selection of terminology, even if the meaning of the chosen terms is correct. This is especially true if this selected terminology does not match with that preferred by the users themselves, in which case users are likely to express some kind of complaint; it may even be that the entire translation is rejected as sub-standard or inappropriate on such grounds.

Second, we look at how to handle new and rare words. If we inspect the process of human translation more closely, it is easy to identify several differences compared to the methods used in statistical MT (SMT). Unless stipulated by the client, the selection of bilingual terminology can be a highly subjective process. Accordingly, it is not necessarily the bilingual term-pair with the highest probability that is chosen by the human translator. It is often the case that statistical methods often forget about or delete less frequent n-grams, but rely on more frequent n-grams using maximum likelihood or Maximum A Priori (MAP) methods. If some terminology is highly suitable, a human translator can use it quite freely. Furthermore, there are a lot of new words in reality for which new target equivalents have to be created by the translators themselves, so the question arises as to how human translators actually select appropriate new terminology. Transliteration, which is often supported by many Asian languages including Hindi, Japanese, and Chinese, is perhaps the easiest things to do under such circumstances. Slight modifications of alphabets/accented characters can sometimes successfully create a valid new term, even for European languages.

The remainder of this paper is organized as follows. Section 2 describes our algorithms. Our decoding strategy in Section 3. Our experimental settings and results are presented in Section 4, and we conclude in Section 5.

2 Our Methods

Apart from the conventional statistical approach to extract bilingual terminology, this medical query task reminds us of two frequently occurring problems which are often ignored: (i) “Can we forget about terminology which occurs only once in a corpus?”, and (ii) “What can we do if the terminology does not occur in a corpus?” These two problems require computationally quite different approaches than what is usually done in the standard statistical approach. Furthermore, the medical query task in WMT14 provides a wide range of corpora: parallel and monolingual corpora, as well as dictionaries. These two interesting aspects motivate our extraction methods which we present in this section, including one relatively new Machine Learning algorithm of zero-shot learning arising from recent developments in the neural network community (Bengio et al., 2000; Mikolov et al., 2013b).

2.1 Translation Model

Word alignment (Brown et al., 1993) and phrase extraction (Koehn et al., 2003) can capture bilingual word- and phrase-pairs with a good deal of accuracy. We omit further details of these standard methods which are freely available elsewhere in the SMT literature (e.g. (Koehn, 2010)).

2.2 Extraction from Parallel Corpora

(Okita et al., 2010) addressed the problem of capturing bilingual term-pairs from parallel data which might otherwise not be detected by the translation model. Hence, the requirement in Okita et al. is not to use SMT/GIZA++ (Och and Ney, 2003) to extract term-pairs, which are the common focus in this medical query translation task.

The classical algorithm of (Kupiec, 1993) used in (Okita et al., 2010) counts the statistics of terminology $c(e_{term_i}, f_{term_j}|s_t)$ on the source and the target sides which jointly occur in a sentence s_t after detecting candidate terms via POS tagging, which are then summed up over the entire corpus $\sum_{t=1}^{N} c(e_{term_i}, f_{term_j}|s_t)$. Then, the algorithm adjusts the length of e_{term_i} and f_{term_j}. It can be said that this algorithm captures term-pairs which occur rather frequently. However, this apparent strength can also be seen in disadvantageous terms since the search for terminology occurs densely in each of the sentences which increases the computational complexity of this algorithm, and causes the method to take a considerable time to run. Furthermore, if we suppose that most frequent term-pairs are to be extracted via a standard translation model (as described briefly in the previous section), our efforts to search among frequent pairs is not likely to bring about further gain.

It is possible to approach this in a reverse manner: “less frequent pairs can be outstanding term candidates”. Accordingly, if our aim changes to capture only those less frequent pairs, the situation changes dramatically. The number of terms we need to capture is considerably decreased. Many sentences do not include any terminology at all, and only a relatively small subset of sentences includes a few terms, such that term-pairs become sparse with regard to sentences. Term-pairs can be found rather easily if a candidate term-pair co-occurs on the source and the target sides *and* on the condition that the items in the term-pair actually correspond with one another.

This condition can be easily checked in various ways. One way is to translate the source side of the targeted pairs with the alignment option in the Moses decoder (Koehn et al., 2007), which we did in this evaluation campaign. Another way is to use asupervised aligner, such as the Berkeley aligner (Haghighi et al., 2009), to align the targeted pairs and check whether they are actually aligned or not.

We assume two predefined sets of terms at the outset, $E_{term} = \{e_{term_1}, \ldots, e_{term_n}\}$ and $F_{term} = \{f_{term_1}, \ldots, f_{term_n}\}$. We search for possible alignment links between the term-pair only when they co-occur in the same sentence. One obvious advantage of this approach is the computational complexity which is fairly low.

Note that the result of (Okita et al., 2010) shows that the frequency-based approach of (Kupiec, 1993) worked well for NTCIR patent terminology (Fujii et al., 2010), which otherwise would have been difficult to capture via the traditional SMT/GIZA++ method. In contrast, however, this did not work well on the Europarl corpus (Koehn, 2005).

2.3 Terminology Dictionaries

Terminology dictionaries themselves are obviously among the most important resources for bilingual term-pairs. In this medical query translation subtask, two corpora are provided for this purpose: (i) Unified Medical Language System corpus (UMLS corpus),[1] and (ii) Wiki entries.[2]

2.4 Extraction from Terminology Dictionaries: lower-order *n*-grams

Terminology dictionaries provide reliable higher-order *n*-gram pairs. However, they do not often provide the correspondences between the lower-order *n*-grams contained therein. For example, the UMLS corpus provides a term-pair of "abdominal compartment syndrome ||| *syndrome du compartiment abdominal*" (EN|||FR). However, such terminology dictionaries often do not explicitly provide the correspondent pairs "abdominal ||| *abdominal*" (EN|||FR) or "syndrome ||| *syndrome*" (EN|||FR). Clearly, these terminology dictionaries implicitly provide the correspondent pairs. Note that UMLS and Wiki entries provide terminology dictionaries. Hence, it is possible to obtain some suggestion by higher order *n*-gram models if we know their alignments between words on the source and target sides. Algorithm 1 shows the overall procedure.

Algorithm 1 Lower-order *n*-gram extraction algorithm

```
Perform monolingual word alignment for higher-order $n$-gram pairs.
Collect only the reliable alignment pairs (i.e. discard unreliable alignment pairs).
Extract the lower-order word pairs of our interest.
```

2.5 Extraction from Monolingual Corpora: Transliteration and Abbreviation

Monolingual corpora can be used in various ways, including:

1. *Transliteration*: Many languages support the fundamental mechanism of between European and Asian languages. Japanese even supports a special alphabet – katakana – for this purpose. Chinese and Hindi also permit transliteration using their own alphabets. However, even among European languages, this mechanism makes it possible to find possible translation counterparts for a given term. In this query task, we did this only for the French-to-English direction and only for words containing accented characters (by rule-based conversion).

2. *Abbreviation*: It is often the case that abbreviations should be resolved in the same language. If the translation includes some abbreviation, such as "C. difficile", this needs to be investigated exhaustively in the same language. However, in the specific domain of medical terminology, it is quite likely that possible phrase matches will be successfully identified.

2.6 Extraction from Monolingual Corpora: Zero-Shot Learning

Algorithm 2 Algorithm to connect two word embedding space

```
Prepare the monolingual source and target sentences.
Prepare the dictionary which consists of $U$ entries of source and target sentences among non-stop-words.
Train the neural network language model on the source side and obtain the continuous space real vectors of $X$ dimensions for each word.
Train the neural network language model on the target side and obtain the continuous space real vectors of $X$ dimensions for each word.
Using the real vectors obtained in the above steps, obtain the linear mapping between the dictionary in two continuous spaces using canonical component analysis (CCA).
```

Another interesting terminology extraction method requires neither parallel nor comparable corpora, but rather just monolingual corpora on both sides (possibly unrelated to each other) together with a small amount of dictionary entries which provide already known correspondences between words on the source and target sides (henceforth, we refer to this as the 'dictionary'). This method uses the recently developed zero-shot learning (Palatucci et al., 2009) using neural network language modelling (Bengio et al., 2000; Mikolov et al., 2013b). Then, we train both sides

[1] http://www.nlm.nih.gov/research/umls/.

[2] http://www.wikipedia.org.

with the neural network language model, and use a continuous space representation to project words to each other on the basis of a small amount of correspondences in the dictionary. If we assume that each continuous space is linear (Mikolov et al., 2013c), we can connect them via linear projection (Mikolov et al., 2013b). Algorithm 2 shows this situation.

In our experiments we use U the same as the entries of Wiki and X as 50. Algorithm 3 shows the algorithm to extract the counterpart of OOV words.

Algorithm 3 Algorithm to extract the counterpart of OOV words.

```
Prepare the projection by Algorithm 2.
Detect unknown words in the translation outputs.
Do the projection of it (the source word) into the target word using the trained linear mappings in the training step.
```

3 Decoding Strategy

We deploy six kinds of extraction methods: (1) translation model, (2) extraction from parallel corpora, (3) terminology dictionaries, (4) lower-order *n*-grams, (5) transliteration and abbreviation, and (6) zero-shot learning. Among these we deploy four of them – (2), (4), (5) and (6) – in a limited context, while the remaining two are used without any context, mainly owing to time constraints; only when we did not find the correspondent pairs via (1) and (3), did we complement this by the other methods.

The detected bilingual term-pairs using (1) and (3) can be combined using various methods. One way is to employ a method similar to (confusion network-based) system combination (Okita and van Genabith, 2011; Okita and van Genabith, 2012). First we make a lattice: if we regard one candidate of (1) and two candidates in (3) as translation outputs where the words of two candidates in (3) are connected using an underscore (i.e. one word), we can make a lattice. Then, we can deploy monotonic decoding over them. If we do this for the devset and then apply it to the test set, we can incorporate a possible preference learnt from the development set, i.e. whether the query translator prefers method (1) or UMLS/Wiki translation. MERT process and language model are applied in a similar manner with (confusion network-based) system combination (cf. (Okita and van Genabith, 2011)).

We note also that a lattice structure is useful for handling grammatical coordination. Since queries are formed by real users, reserved words for database query such as "AND" (or "*ET*" (FR)) and "OR" (or "*OU*" (FR)) are frequently observed in the test set. Furthermore, there is repeated use of "and" more than twice, for example "*douleur abnominal et Helicobacter pylori et cancer*", which makes it very difficult to detect the correct coordination boundaries. The lattice on the input side can express such ambiguity at the cost of splitting the source-side sentence in a different manner.

4 Experimental Results

The baseline is obtained in the following way. The GIZA++ implementation (Och and Ney, 2003) of IBM Model 4 is used as the baseline for word alignment: Model 4 is incrementally trained by performing 5 iterations of Model 1, 5 iterations of HMM, 3 iterations of Model 3, and 3 iterations of Model 4. For phrase extraction the grow-diag-final heuristics described in (Koehn et al., 2003) is used to derive the refined alignment from bidirectional alignments. We then perform MERT (Och, 2003) which optimizes parameter settings using the BLEU metric (Papineni et al., 2002), while a 5-gram language model is derived with Kneser-Ney smoothing (Kneser and Ney, 1995) trained using SRILM (Stolcke, 2002). We use the whole training corpora including the WMT14 translation task corpora as well as medical domain data. UMLS and Wikipedia are used just as training corpora for the baseline.

For the extraction from parallel corpora (cf. Section 2.2), we used Genia tagger (Tsuruoka and Tsujii, 2005) and the Berkeley parser (Petrov and Klein, 2007). For the zero-shot learning (cf. Section 2.6) we used scikit learn (Pedregosa et al., 2011), word2vec (Mikolov et al., 2013a), and a recurrent neural network (Mikolov, 2012). Other tools used are in-house software.

Table 2 shows the results for the FR–EN query task. We obtained 36.2 BLEU points absolute, which is an improvement of 6.3 BLEU point absolute (21.1% relative) over the baseline. Table 3 shows the results for the EN–FR query task. We obtained 28.8 BLEU points absolute, which is an improvement of 8.7 BLEU points absolute (43% relative) over the baseline. Our system was the best system for both of these tasks. These improvements over the baseline were statistically significant by a paired bootstrap test (Koehn, 2004).

	Query task FR–EN	
	Our method	baseline
BLEU	36.2	29.9
BLEU cased	30.9	26.5
TER	0.340	0.443

Table 1: Results for FR–EN query task.

extraction	LM	MERT	BLEU (cased)
(1) - (6)	all	Y	30.9
(1), (2), (3)	all	Y	30.3
(1), (3), (6)	all	Y	30.1
(1), (3), (4)	all	Y	29.1
(1), (3), (5)	all	Y	29.0
(1) and (3)	all	Y	29.0
(1) and (3)	medical	Y	27.5
(1) and (3)	WMT	Y	27.0
(1) and (3)	medical	N	25.1
(1) and (3)	WMT	N	24.3
(1)	medical	Y	25.9
(1)	WMT	Y	25.0

Table 2: Table shows the effects of extraction methods, language model and MERT process. All the measurements are by BLEU (cased). In this table, “medical” indicates a language model built on all the medical corpora while “WMT” indicates a language model built on all the non-medical corpora. Note that some sentence in testset can be considered as non-medical domain. Extraction methods (1) - (6) correspond to those described in Section 2.1 - 2.6.

Table 4 shows the results for CLIR task. We obtained 51.8 BLEU points absolute, which is an improvement of 9.4 BLEU point absolute (22.2% relative) over the baseline. Although CLIR task allowed 10-best lists, our submission included only 1-best list. This resulted in the score of P@5 of 0.348 and P@10 of 0.346 which correspond to the second place, despite a good result in terms of BLEU. This is since unlike BLEU score P@5 and P@10 measure whether the whole elements in reference and hypothesis are matched or not. We noticed that our submission included a lot of near miss sentences only in terms of capitalization: “abnominal pain and Helicobacter pylori and cancer” (reference) and “abnominal pain and helicobacter pylori and cancer” (submission). These are counted as incorrect in terms of P@5 and P@10.[3] Noted that after submission we obtained the revised score of P@5 of 0.560 and P@10 of 0.560 with the same method but with 2-best lists which handles the capitalization varieties.

	Query task EN–FR	
	Our method	baseline
BLEU	28.8	20.1
BLEU cased	27.7	18.7
TER	0.483	0.582

Table 3: Results for EN–FR query task.

	CLIR task FR–EN	
	Our method	baseline
BLEU	51.8	42.2
BLEU cased	46.0	38.3
TER	0.364	0.398
P@5	0.348 (0.560*)	–
P@10	0.346 (0.560*)	–
NDCG@5	0.306	–
NDCG@10	0.307	–
MAP	0.2252	–
Rprec	0.2358	–
bpref	0.3659	–
relRet	1524	–

Table 4: Results for CLIR task.

5 Conclusion

This paper provides a description of the Dublin City University terminology translation system for our participation in the query translation subtask in the medical translation task in the Workshop on Statistical Machine Translation (WMT14). We deployed six different kinds of terminology extraction methods. We obtained 36.2 BLEU points absolute for FR–EN, and 28.8 BLEU points absolute for EN–FR tasks, obtaining first place on both tasks. We obtained 51.8 BLEU points absolute for the CLIR task.

[3] The method which incorporates variation in capitalization in its n-best lists outperforms the best result in terms of P@5 and P@10.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of CNGL at Dublin City University.

References

Yoshua Bengio, Rejean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *In Proceedings of Neural Information Systems*, pages 1137–1155.

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

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizenya, and Sayori Shimohata. 2010. Overview of the patent translation task at the NTCIR-8 workshop. *In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 293–302.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised itg models. In *In Proceedings of the Conference of Association for Computational Linguistics*, pages 923–931.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for n-gram language modeling. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computationa Linguistics (HLT / NAACL 2003)*, pages 115–124.

Philipp Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for Statistical Machine Translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. *In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 388–395.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *In Proceedings of the Machine Translation Summit*, pages 79–86.

Philipp Koehn. 2010. Statistical machine translation. *Cambridge University Press.*

Julian. Kupiec. 1993. An algorithm for finding Noun phrase correspondences in bilingual corpora. *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 17–22.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *In Proceedings of Workshop at International Conference on Learning Representations.*

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *ArXiv:1309.4168.*

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics / Human Language Technology (NAACL/HLT 2005)*, pages 746–751.

Tomas Mikolov. 2012. Statistical language models based on neural networks. *PhD thesis at Brno University of Technology.*

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Tsuyoshi Okita and Josef van Genabith. 2011. DCU Confusion Network-based System Combination for ML4HMT. *Shared Task on Applying Machine Learning techniques to optimising the division of labour in Hybrid MT (ML4HMT-2011, collocated with LIHMT-2011)*, pages 93–98.

Tsuyoshi Okita and Josef van Genabith. 2012. Minimum Bayes Risk Decoding with Enlarged Hypothesis Space in System Combination. *In Proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*, pages 40–51.

Tsuyoshi Okita, Alfredo Maldonado Guerra, Yvette Graham, and Andy Way. 2010. Multi-word expression-sensitive word alignment. *In Proceedings of the Fourth International Workshop On Cross Ling ual Information Access (CLIA2010, collocated with COLING2010)*, pages 26–34.

Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom Mitchell. 2009. Zero-shot learning with semantic output codes. In *Neural Information Processing Systems (NIPS)*, pages 1410–1418.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: A Method For Automatic Evaluation of Machine Translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02).*

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Slav Petrov and Dan Klein. 2007. Learning and inference for hierarchically split PCFGs. In *Proceedings of AAAI (Nectar Track)*, pages 1663–1666.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. *In Proceedings of the International Conference on Spoken Language Processing*, pages 901–904.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. *In Proceedings of the Conference on Human Language Technology / Empirical Methods on Natural Language Processing (HLT/EMNLP 2005)*, pages 467–474.

Andy Way. 2013. Traditional and emerging use-cases for machine translation. *In Proceedings of Translating and the Computer 35*.