Combining Domain Adaptation Approaches for Medical Text Translation

Longyue Wang, Yi Lu, Derek F. Wong, Lidia S. Chao, Yiming Wang, Francisco Oliveira
Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory,
Department of Computer and Information Science,
University of Macau, Macau, China
vincentwang0229@hotmail.com,
{mb25435, derekfw, lidiasc, mb25433, olifran}@umac.mo

Abstract

This paper explores a number of simple and effective techniques to adapt statistical machine translation (SMT) systems in the medical domain. Comparative experiments are conducted on large corpora for six language pairs. We not only compare each adapted system with the baseline, but also combine them to further improve the domain-specific systems. Finally, we attend the WMT2014 medical summary sentence translation constrained task and our systems achieve the best BLEU scores for Czech-English, English-German, French-English language pairs and the second best BLEU scores for reminding pairs.

1. Introduction

This paper presents the experiments conducted by the NLP^{2}CT Laboratory at the University of Macau for WMT2014 medical sentence translation task on six language pairs: Czech-English (cs-en), French-English (fr-en), German-English (de-en) and the reverse direction pairs, i.e., en-cs, en-fr and en-de.

By comparing the medical text with common text, we discovered some interesting phenomena in medical genre. We apply domain-specific techniques in data pre-processing, language model adaptation, translation model adaptation, numeric and hyphenated words translation. Compared to the baseline systems (detailed in Section 2 & 3), the results of each method show reasonable gains. We combine individual approach to further improve the performance of our systems. To validate the robustness and language-independency of individual and combined systems, we conduct experiments on the official training data (detailed in Section 3) in all six language pairs. We anticipate the numeric comparison (BLEU scores) on these individual and combined domain adaptation approaches that could be valuable for others on building a real-life domain-specific system.

The reminder of this paper is organized as follows. In Section 2, we detail the configurations of our experiments as well as the baseline systems. Section 3 presents the specific preprocessing for medical data. In Section 4 and 5, we describe the language model (LM) and translation model (TM) adaptation, respectively. Besides, the techniques for numeric and hyphenated words translation are reported in Section 6 and 7. Finally, the performance of design systems and the official results are reported in Section 8.

2. Experimental Setup

All available training data from both WMT2014 standard translation task[1] (general-domain data) and medical translation task[2] (in-domain data) are used in this study. The official medical summary development sets (dev) are used for tuning and evaluating all the comparative systems. The official medical summary test sets (test) are only used in our final submitted systems.

The experiments were carried out with the Moses 1.0[3] (Koehn et al., 2007). The translation and the re-ordering model utilizes the "*grow-diag-final*" symmetrized word-to-word alignments created with MGIZA++[4] (Och and Ney,

[1] http://www.statmt.org/wmt14/translation-task.html.
[2] http://www.statmt.org/wmt14/medical-task/.
[3] http://www.statmt.org/moses/.
[4] http://www.kyloo.net/software/doku.php/mgiza:overview.

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 254–259,
Baltimore, Maryland USA, June 26–27, 2014.

2003; Gao and Vogel, 2008) and the training scripts from Moses. A 5-gram LM was trained using the SRILM toolkit[5] (Stolcke et al., 2002), exploiting improved modified Kneser-Ney smoothing, and quantizing both probabilities and back-off weights. For the log-linear model training, we take the minimum-error-rate training (MERT) method as described in (Och, 2003).

3. Task Oriented Pre-processing

A careful pre-processing on training data is significant for building a real-life SMT system. In addition to the general data preparing steps used for constructing the baseline system, we introduce some extra steps to pre-process the training data.

The first step is to remove the duplicate sentences. In data-driven methods, the more frequent a term occurs, the higher probability it biases. Duplicate data may lead to unpredicted behavior during the decoding. Therefore, we keep only the distinct sentences in monolingual corpus. By taking into account multiple translations in parallel corpus, we remove the duplicate sentence pairs. The second concern in preprocessing is symbol normalization. Due to the nature of medical genre, symbols such as numbers and punctuations are commonly-used to present chemical formula, measuring unit, terminology and expression. Fig. 1 shows the examples of this case. These symbols are more frequent in medical article than that in the common texts. Besides, the punctuations of *apostrophe* and *single quotation* are interchangeably used in French text, e.g. "*l'effet de l'inhibition*". We unify it by replacing with the *apostrophe*. In addition, we observe that some monolingual training subsets (e.g., Gene Regulation Event Corpus) contain sentences of more than 3,000 words in length. To avoid the long sentences from harming the truecase model, we split them into sentences with a sentence splitter[6] (Rune et al., 2007) that is optimized for biomedical texts. On the other hand, we consider the target system is intended for summary translation, the sentences tend to be short in length. For instance, the average sentence lengths in development sets of cs, fr, de and en are around 15, 21, 17 and 18, respectively. We remove sentence pairs which are more than 80 words at length. In order to that our experiments are reproducible, we give the detailed statistics of task oriented pre-processed training data in Table 2.

1,25-OH
47 to **80%**
10-20 ml/kg
A**&**E department
Infective endocarditis **(**IE**)**

Figure 1. Examples of the segments with symbols in medical texts.

To validate the effectiveness of the preprocessing, we compare the SMT systems trained on original data[7] (*Baseline*1) and task-oriented-processed data (*Baseline*2), respectively. Table 1 shows the results of the baseline systems. We found all the *Baseline*2 systems outperform the *Baseline*1 models, showing that the systems can benefit from using the processed data. For cs-en and en-cs pairs, the BLEU scores improve quite a lot. For other language pairs, the translation quality improves slightly.

By analyzing the *Baseline*2 results (in Table 1) and the statistics of training corpora (in Table 2), we can further elaborate and explain the results. The en-cs system performs poorly, because of the short average length of training sentences, as well as the limited size of in-domain parallel and monolingual corpora. On the other hand, the fr-en system achieves the best translation score, as we have sufficient training data. The translation quality of cs-en, en-fr, fr-en and de-en pairs is much higher than those in the other pairs. Hence, *Baseline*2 will be used in the subsequent comparisons with the proposed systems described in Section 4, 5, 6 and 7.

Lang. Pair	*Baseline*1	*Baseline*2	Diff.
en-cs	12.92	**17.57**	+4.65
cs-en	20.85	**31.29**	+10.44
en-fr	38.31	**38.36**	+0.05
fr-en	44.27	**44.36**	+0.09
en-de	17.81	**18.01**	+0.20
de-en	32.34	**32.50**	+0.16

Table 1: BLEU scores of two baseline systems trained on original and processed corpora for different language pairs.

4. Language Model Adaptation

The use of LMs (trained on large data) during decoding is aided by more efficient storage and inference (Heafield, 2011). Therefore, we not

[5] http://www.speech.sri.com/projects/srilm/.
[6] http://www.nactem.ac.uk/y-matsu/geniass/.
[7] Data are processed according to Moses baseline tutorial: http://www.statmt.org/moses/?n=Moses.Baseline.

Data Set	Lang.	Sent.	Words	Vocab.	Ave. Len.
In-domain Parallel Data	cs/en	**1,770,421**	9,373,482/ 10,605,222	134,998/ 156,402	**5.29/ 5.99**
	de/en	3,894,099	52,211,730/ 58,544,608	1,146,262/ 487,850	13.41/ 15.03
	fr/en	**4,579,533**	77,866,237/ 68,429,649	495,856/ 556,587	17.00/ 14.94
General-domain Parallel Data	cs/en	12,426,374	180,349,215/ 183,841,805	1,614,023/ 1,661,830	14.51/ 14.79
	de/en	4,421,961	106,001,775/ 112,294,414	1,912,953/ 919,046	23.97/ 25.39
	fr/en	36,342,530	1,131,027,766/ 953,644,980	3,149,336/ 3,324,481	31.12/ 26.24
In-domain Mono. Data	cs	**106,548**	1,779,677	150,672	16.70
	fr	1,424,539	53,839,928	644,484	37.79
	de	2,222,502	53,840,304	1,415,202	24.23
	en	**7,802,610**	199430649	1,709,594	25.56
General-domain Mono. Data	cs	33,408,340	567,174,266	3,431,946	16.98
	fr	30,850,165	780,965,861	2,142,470	25.31
	de	84,633,641	1,548,187,668	10,726,992	18.29
	en	85,254,788	2,033,096,800	4,488,816	23.85

Table 2: Statistics summary of corpora after pre-processing.

only use the in-domain training data, but also the selected pseudo in-domain data[8] from general-domain corpus to enhance the LMs (Toral, 2013; Rubino et al., 2013; Duh et al., 2013). Firstly, each sentence *s* in general-domain monolingual corpus is scored using the cross-entropy difference method in (Moore and Lewis, 2010), which is calculated as follows:

$$score(s) = H_I(s) - H_G(s) \quad (1)$$

where $H(s)$ is the length-normalized cross-entropy. *I* and *G* are the in-domain and general-domain corpora, respectively. *G* is a random subset (same size as the *I*) of the general-domain corpus. Then top *N* percentages of ranked data sentences are selected as a pseudo in-domain subset to train an additional LM. Finally, we linearly interpolate the additional LM with in-domain LM.

We use the top *N*% of ranked results, where *N*={0, 25, 50, 75, 100} percentages of sentences out of the general corpus. Table 3 shows the absolute BLEU points for *Baseline*2 (*N*=0), while the LM adapted systems are listed with values relative to the *Baseline*2. The results indicate that LM adaptation can gain a reasonable improvement if the LMs are trained on more relevant data for each pair, instead of using the whole training data. For different systems, their BLEU scores peak at different values of *N*. It gives the best results for cs-en, en-fr and de-en pairs when *N*=25, en-cs and en-de pairs when *N*=50, and fr-en pair when *N*=75. Among them, en-cs and en-fr achieve the highest BLEU scores. The reason is that their original monolingual (in-domain) data for training the LMs are not sufficient. When introducing the extra pseudo in-domain data, the systems improve the translation quality by around 2 BLEU points. While for cs-en, fr-en and de-en pairs, the gains are small. However, it can still achieve a significant improvement of 0.60 up to 1.12 BLEU points.

Lang.	*N*=0	*N*=25	*N*=50	*N*=75	*N*=100
en-cs	17.57	+1.66	**+2.08**	+1.72	+2.04
cs-en	31.29	**+0.94**	+0.60	+0.66	+0.47
en-fr	38.36	**+1.82**	+1.66	+1.60	+0.08
fr-en	44.36	+0.91	+1.09	**+1.12**	+0.92
en-de	18.01	+0.57	**+1.02**	-4.48	-4.54
de-en	32.50	**+0.60**	+0.50	+0.56	+0.38

Table 3: BLEU scores of LM adapted systems.

5. Translation Model Adaptation

As shown in Table 2, general-domain parallel corpora are around 1 to 7 times larger than the in-domain ones. We suspect if general-domain corpus is broad enough to cover some in-domain sentences. To observe the domain-specificity of general-domain corpus, we firstly evaluate systems trained on general-domain corpora. In Ta-

[8] Axelrod et al. (2011) names the selected data as *pseudo in-domain data*. We adopt both terminologies in this paper.

ble 4, we show the BLEU scores of general-domain systems[9] on translating the medical sentences. The BLEU scores of the compared systems are relative to the *Baseline*2 and the size of the used general-domain corpus is relative to the corresponding in-domain one. For en-cs, cs-en, en-fr and fr-en pairs, the general-domain parallel corpora we used are 6 times larger than the original ones and we obtain the improved BLEU scores by 1.72 up to 3.96 points. While for en-de and de-en pairs, the performance drops sharply due to the limited training corpus we used. Hence we can draw a conclusion: the general-domain corpus is able to aid the domain-specific translation task if the general-domain data is large and broad enough in content.

Lang. Pair	BLEU	Diff.	Corpus
en-cs	21.53	+3.96	+601.89%
cs-en	33.01	+1.72	
en-fr	41.57	+3.21	+693.59%
fr-en	47.33	+2.97	
en-de	16.54	-1.47	+13.63%
de-en	27.35	-5.15	

Table 4: The BLEU scores of systems trained on general-domain corpora.

Taking into account the performance of general-domain system, we explore various data selection methods to derive the pseudo in-domain sentence pairs from general-domain parallel corpus for enhancing the TMs (Wang et al., 2013; Wang et al., 2014). Firstly, sentence pair in corresponding general-domain corpora is scored by the modified Moore-Lewis (Axelrod et al., 2011), which is calculated as follows:

$$score(s) = \left[H_{I-src}(s) - H_{G-src}(s)\right] + \left[H_{I-tgt}(s) - H_{G-tgt}(s)\right] \quad (2)$$

which is similar to Eq. (1) and the only difference is that it considers the both the source (*src*) and target (*tgt*) sides of parallel corpora. Then top *N* percentage of ranked sentence pairs are selected as a pseudo in-domain subset to train an individual translation model. The additional model is log-linearly interpolated with the in-domain model (*Baseline*2) using the multi-decoding method described in (Koehn and Schroeder, 2007).

Similar to LM adaptation, we use the top *N*% of ranked results, where *N*={0, 25, 50, 75, 100} percentages of sentences out of the general corpus. Table 5 shows the absolute BLEU points for *Baseline*2 (*N*=0), while for the TM adapted systems we show the values relative to the *Baseline*2. For different systems, their BLEU peak at different *N*. For en-fr and en-de pairs, it gives the best translation results at *N*=25. Regarding cs-en and fr-en pairs, the optimal performance is peaked at *N*=50. While the best results for de-en and en-cs pairs are *N*=75 and *N*=100 respectively. Besides, performance of TM adapted system heavily depends on the size and (domain) broadness of the general-domain data. For example, the improvements of en-de and de-en systems are slight due to the small general-domain corpora. While the quality of other systems improve about 3 BLEU points, because of their large and broad general-domain corpora.

Lang.	*N*=0	*N*=25	*N*=50	*N*=75	*N*=100
en-cs	17.57	+0.84	+1.53	+1.74	**+2.55**
cs-en	31.29	+2.03	**+3.12**	+3.12	+2.24
en-fr	38.36	**+3.87**	+3.66	+3.53	+2.88
fr-en	44.36	+1.29	**+3.36**	+1.84	+1.65
en-de	18.01	**+0.02**	-0.13	-0.07	0
de-en	32.50	-0.12	+0.06	**+0.31**	+0.24

Table 5: BLEU scores of TM adapted systems.

6. Numeric Adaptation

As stated in Section 3, *numeric* occurs frequently in medical texts. However, numeric expression in dates, time, measuring unit, chemical formula are often sparse, which may lead to OOV problems in phrasal translation and reordering. Replacing the sparse numbers with placeholders may produce more reliable statistics for the MT models.

Moses has support using placeholders in training and decoding. Firstly, we replace all the numbers in monolingual and parallel training corpus with a common symbol (a sample phrase is illustrated in Fig. 2). Models are then trained on these processed data. We use the XML markup translation method for decoding.

Original:	Vitamin D 1,25-OH
Replaced:	Vitamin D @num@, @num@-OH

Figure 2. Examples of placeholders.

Table 6 shows the results on this number adaptation approach as well as the improvements compared to the *Baseline*2. The method improves the *Baseline*2 systems by 0.23 to 0.40 BLEU scores. Although the scores increase slightly, we still believe this adaptation method is significant for medical domain. The WMT2014 medical task only focuses on the summary of

[9] General-domain systems are trained only on general-domain training corpora (i.e., parallel, monolingual).

medical text, which may contain fewer chemical expression in compared with the full article. As the used of numerical instances increases, placeholder may play a more important role in domain adaptation.

Lang. Pair	BLEU (Dev)	Diff.
en-cs	17.80	+0.23
cs-en	31.52	+0.23
en-fr	38.72	+0.36
fr-en	44.69	+0.33
en-de	18.41	+0.40
de-en	32.88	+0.38

Table 6: BLEU scores of numeric adapted systems.

7. Hyphenated Word Adaptation

Medical texts prefer a kind of compound words, hyphenated words, which is composed of more than one word. For instance, "*slow-growing*" and "*easy-to-use*" are composed of words and linked with hyphens. These hyphenated words occur quite frequently in medical texts. We analyze the development sets of cs, fr, en and de respectively, and observe that there are approximately 3.2%, 11.6%, 12.4% and 19.2% of sentences that contain one or more hyphenated words. The high ratio of such compound words results in Out-Of-Vocabulary words (OOV) [10], and harms the phrasal translation and reordering. However, a number of those hyphenated words still have chance to be translated, although it is not precisely, when they are tokenized into individual words.

Algorithm: Alternative-translation Method

Input:

1. A sentence, *s*, with *M* hyphenated words
2. Translation lexicon

Run:

1. **For** $i = 1, 2, \ldots, M$
2. Split the *i*th hyphenated word (C_i) into P_i
3. Translate P_i into T_i
4. **If** (T_i are not OOVs):
5. Put alternative translation T_i in XML
6. **Else**: keep C_i unchanged

Output:

Sentence, *s'*, embedded with alternative translations for all T_i.

End

Table 7: Alternative-translation algorithm.

[10] Default tokenizer does not handle the hyphenated words.

To resolve this problem, we present an *alternative-translation* method in decoding. Table 7 shows the proposed algorithm.

In the implementation, we apply XML markup to record the translation (terminology) for each compound word. During the decoding, a hyphenated word delimited with markup will be replaced with its corresponding translation. Table 8 shows the BLEU scores of adapted systems applied to hyphenated translation. This method is effective for most language pairs. While the translation systems for en-cs and cs-en do not benefit from this adaptation, because the hyphenated words ratio in the en and cs dev are asymmetric. Thus, we only apply this method for en-fr, fr-en, de-en and en-de pairs.

Lang. Pair	BLEU (Dev)	Diff.
en-cs	16.84	-0.73
cs-en	31.23	-0.06
en-fr	39.12	+0.76
fr-en	45.02	+0.66
en-de	18.64	+0.63
de-en	33.01	+0.51

Table 8: BLEU scores of hyphenated word adapted systems.

3. Final Results and Conclusions

According to the performance of each individual domain adaptation approach, we combined the corresponding models for each language pair. In Table 10, we show the BLEU scores and its increments (compared to the *Baseline*2) of combined systems in the second column. The official test set is converted into the *recased* and *detokenized* SGML format. The official results of our submissions are given in the last column of Table 9.

Lang. Pair	BLEU of Combined systems	Official BLEU
en-cs	23.66 (+6.09)	22.60
cs-en	38.05 (+6.76)	37.60
en-fr	42.30 (+3.94)	41.20
fr-en	48.25 (+3.89)	47.10
en-de	21.14 (+3.13)	20.90
de-en	36.03 (+3.53)	35.70

Table 9: BLEU scores of the submitted systems for the medical translation task.

This paper presents a set of experiments conducted on all available training data for six language pairs. We explored various domain adaptation approaches for adapting medical transla-

tion systems. Compared with other methods, language model adaptation and translation model adaptation are more effective. Other adapted techniques are still necessary and important for building a real-life system. Although all individual methods are not fully additive, combining them together can further boost the performance of the overall domain-specific system. We believe these empirical approaches could be valuable for SMT development.

Acknowledgments

The authors are grateful to the Science and Technology Development Fund of Macau and the Research Committee of the University of Macau for the funding support for their research, under the Reference nos. MYRG076 (Y1-L2)-FST13-WF and MYRG070 (Y1-L2)-FST12-CS. The authors also wish to thank the colleagues in CNGL, Dublin City University (DCU) for their helpful suggestion and guidance on related work.

Reference

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP*, pages 355-362.

K. Duh, G. Neubig, K. Sudoh, H. Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages, 678–683.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49-57.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187-197.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the 2nd ACL Workshop on Statistical Machine Translation*, pages 224-227.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran et al. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL*, pages 177-180.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of ACL*: Short Papers, pages 220-224.

Sætre Rune, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi and Tomoko Ohta. 2007. AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 209-212.

Raphael Rubino, Antonio Toral, Santiago Cortés Va Ílo, Jun Xie, Xiaofeng Wu, Stephen Doherty, and Qun Liu. 2013. The CNGL-DCU-Prompsit translation systems for WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 213-218.

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

Antonio Toral. 2013. Hybrid selection of language model training data using linguistic information and perplexity. In *ACL Workshop on Hybrid Machine Approaches to Translation*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19-51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160-167.

Longyue Wang, Derek F. Wong, Lidia S. Chao, Yi Lu, and Junwen Xing. 2014 "A Systematic Comparison of Data Selection Criteria for SMT Domain Adaptation," *The Scientific World Journal*, vol. 2014, Article ID 745485, 10 pages.

Longyue Wang, Derek F. Wong, Lidia S. Chao, Yi Lu, Junwen Xing. 2013. *i*CPE: A Hybrid Data Selection Model for SMT Domain Adaptation. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer Berlin Heidelberg. pages, 280-290.