# Domain Adaptation for Medical Text Translation Using Web Resources

**Yi Lu, Longyue Wang, 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
takamachi660@gmail.com, vincentwang0229@hotmail.com,
derekfw@umac.mo, lidiasc@umac.mo, wang2008499@gmail.com,
olifran@umac.mo

## Abstract

This paper describes adapting statistical machine translation (SMT) systems to medical domain using in-domain and general-domain data as well as web-crawled in-domain resources. In order to complement the limited in-domain corpora, we apply domain focused web-crawling approaches to acquire in-domain monolingual data and bilingual lexicon from the Internet. The collected data is used for adapting the language model and translation model to boost the overall translation quality. Besides, we propose an alternative filtering approach to clean the crawled data and to further optimize the domain-specific SMT system. We attend the medical summary sentence unconstrained translation task of the Ninth Workshop on Statistical Machine Translation (WMT2014). Our systems achieve the second best BLEU scores for Czech-English, fourth for French-English, English-French language pairs and the third best results for reminding pairs.

## 1 Introduction

In this paper, we report the experiments carried out by the NLP$^2$CT Laboratory at 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).

As data in specific domain are usually relatively scarce, the use of web resources to complement the training resources provides an effective way to enhance the SMT systems (Resnik and smith, 2003; Espl à-Gomis and Forcada, 2010; Pecina et al., 2011; Pecina et al., 2012; Pecina et al., 2014). In our experiments, we not only use all available training data provided by the WMT2014 standard translation task [1] (general-domain data) and medical translation task [2] (in-domain data), but also acquire addition in-domain bilingual translations (i.e. dictionary) and monolingual data from online sources.

First of all, we collect the medical terminologies from the web. This tiny but significant parallel data are helpful to reduce the out-of-vocabulary words (OOVs) in translation models. In addition, the use of larger language models during decoding is aided by more efficient storage and inference (Heafield, 2011). Thus, we crawl more in-domain monolingual data from the Internet based on domain focused web-crawling approach. In order to detect and remove out-domain data from the crawled data, we not only explore text-to-topic classifier, but also propose an alternative filtering approach combined the existing one (text-to-topic classifier) with perplexity. After carefully pre-processing all the available training data, we apply language model adaptation and translation model adaptation using various kinds of training corpora. Experimental results show that the presented approaches are helpful to further boost the baseline system.

The reminder of this paper is organized as follows. In Section 2, we detail the workflow of web resources acquisition. Section 3 describes the pre-processing steps for the corpora. Section 5 presents the baseline system. Section 6 reports the experimental results and discussions. Finally,

[1] http://www.statmt.org/wmt14/translation-task.html.
[2] http://www.statmt.org/wmt14/medical-task/.



*Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 233–238,
Baltimore, Maryland USA, June 26–27, 2014. 

the submitted systems and the official results are reported in Section 7.

## 2 Domain Focused Web-Crawling

In this section, we introduce our domain focused web-crawling approaches on acquisition of in-domain translation terminologies and monolingual sentences.

### 2.1 Bilingual Dictionary

Terminology is a system of words used to name things in a particular discipline. The in-domain vocabulary size directly affects the performance of domain-specific SMT systems. Small size of in-domain vocabulary may result in serious OOVs problem in a translation system. Therefore, we crawl medical terminologies from some online sources such as dict.cc[3], where the vocabularies are divided into different subjects. We obtain the related bilingual entries in medicine subject by using Scala build-in XML parser and XPath. After cleaning, we collected 28,600, 37,407, and 37,600 entries in total for cs-en, de-en, and fr-en respectively.

### 2.2 Monolingual Data

The workflow for acquiring in-domain resources consists of a number of steps such as domain identification, text normalization, language identification, noise filtering, and post-processing as well as parallel sentence identification.

Firstly we use an open-source crawler, Combine[4], to crawl webpages from the Internet. In order to classify these webpages as relevant to the medical domain, we use a list of triplets <*term, relevance weight, topic class*> as the basic entries to define the topic. *Term* is a word or phrase. We select terms for each language from the following sources:

- The Wikipedia title corpus, a WMT2014 official data set consisting of titles of medical articles.
- The dict.cc dictionary, as is described in Section 2.1.
- The DrugBank corpus, which is a WMT2014 official data set on bioinformatics and cheminformatics.

For the parallel data, i.e. Wikipedia and dict.cc dictionary, we separate the source and target text into individual text and use either side of them for constructing the term list for different languages. Regarding the DrugBank corpus, we directly extract the terms from the "*name*" field. The vocabulary size of collected text for each language is shown in Table 1.

| | EN | CS | DE | FR |
|---|---|---|---|---|
| Wikipedia Titles | 12,684 | 3,404 | 10,396 | 8,436 |
| dict.cc | 29,294 | 16,564 | 29,963 | 22,513 |
| DrugBank | 2,788 | | | |
| Total | 44,766 | 19,968 | 40,359 | 30,949 |

Table 1: Size of terms used for topic definition.

*Relevance weight* is the score for each occurrence of the term, which is assigned by its length, i.e., number of tokens. The *topic class* indicates the topics. In this study, we are interested in medical domain, the topic class is always marked with "MED" in our topic definition.

The topic relevance of each document is calculated[5] as follows:

$$s = \sum_{i=1}^{N} \sum_{j=1}^{4} n_{ij} w_i^t w_j^l \qquad (1)$$

where $N$ is the amount of terms in the topic definition; $w_i^t$ is the weight of term $i$; $w_j^l$ is the weight of term at location $j$. $n_{ij}$ is the number of occurrences of term $i$ at $j$ position. In implementation, we use the default values for setting and parameters. Another input required by the crawler is a list of seed URLs, which are web sites that related to medical topic. We limit the crawler from getting the pages within the http domain guided by the seed links. We acquired the list from the Open Directory Project[6], which is a repository maintained by volunteer editors. Totally, we collected 12,849 URLs from the medicine category.

Text normalization is to convert the text of each HTML page into UTF-8 encoding according to the content_charset of the header. In addition, HTML pages often consist of a number of irrelevant contents such as the navigation links, advertisements disclaimers, etc., which may negatively affect the performance of SMT system. Therefore, we use the Boilerpipe tool (Kohlschütter et al., 2010) to filter these noisy data and preserve the useful content that is marked by the tag, <canonicalDocument>. The resulting text is saved in an XML file, which will be further processed by the subsequent tasks. For language identification, we use the language-detection[7] toolkit to determine the possible lan-

[3] http://www.dict.cc/.
[4] http://combine.it.lth.se/.
[5] http://combine.it.lth.se/documentation/DocMain/node6.html.
[6] http://www.dmoz.org/Health/Medicine/.
[7] https://code.google.com/p/language-detection/.

guage of the text, and discard the articles which are in the right language we are interested.

### 2.3 Data Filtering

The web-crawled documents (described in Section 2.2) may consist a number of out-domain data, which would harm the domain-specific language and translation models. We explore and propose two filtering approaches for this task. The first one is to filter the documents based on their relative score, Eq. (1). We rank all the documents according to their relative scores and select top *K* percentage of entire collection for further processing.

Second, we use a combination method, which takes both the perplexity and relative score into account for the selection. Perplexity-based data selection has shown to be a powerful mean on SMT domain adaptation (Wang et al., 2013; Wang et al., 2014; Toral, 2013; Rubino et al., 2013; Duh et al., 2013). The combination method is carried out as follows: we first retrieve the documents based on their relative scores. The documents are then split into sentences, and ranked according to their perplexity using Eq. (2) (Stolcke et al., 2002). The used language model is trained on the official in-domain data. Finally, top *N* percentage of ranked sentences are considered as additional relevant in-domain data.

$$ppl1(s) = 10^{-\log\frac{P(T)}{Word}} \qquad (2)$$

where $s$ is a input sentence or document, $P(T)$ is the probability of $n$-gram segments estimated from the training set. $Word$ is the number of tokens of an input string.

## 3 Pre-processing

Both official training data and web-crawled resources are processed using the Moses scripts[8], this includes the text tokenization, truecasing and length cleaning. For trusecasing, we use both the target side of parallel corpora and monolingual data to train the trucase models. We consider the target system is intended for summary translation, the sentences tend to be short in length. We remove sentence pairs which are more than 80 words at length in either sides of the parallel text.

In addition to these general data filtering steps, 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. We also use a biomedical sentence splitter[9] (Rune et al., 2007) to split sentences in monolingual corpora. The statistics of the data are provided in Table 2.

## 4 Baseline System

We built our baseline system on an optimized level. It is trained on all official in-domain training corpora and a portion of general-domain data. We apply the Moore-Lewis method (Moore and Lewis, 2010) and modified Moore-Lewis method (Axelrod et al., 2011) for selecting in-domain data from the general-domain monolingual and parallel corpora, respectively. The top *M* percentages of ranked sentences are selected as a pseudo in-domain data to train an additional LM and TM. For LM, we linearly interpolate the additional LM with in-domain LM. For TM, the additional model is log-linearly interpolated with the in-domain model using the multi-decoding method described in (Koehn and Schroeder, 2007). Finally, LM adaptation and TM adaptation are combined to further improve the translation quality of baseline system.

## 5 Experiments and Results

The official medical summary development sets (dev) are used for tuning and evaluating 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[10] (Koehn et al., 2007). The translation and the re-ordering model utilizes the "*grow-diag-final*" symmetrized word-to-word alignments created with MGIZA++[11] (Och and Ney, 2003; Gao and Vogel, 2008) and the training scripts from Moses. A 5-gram LM was trained using the SRILM toolkit[12] (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).

[8] http://www.statmt.org/moses/?n=Moses.Baseline.

[9] http://www.nactem.ac.uk/y-matsu/geniass/.

[10] http://www.statmt.org/moses/.

[11] http://www.kyloo.net/software/doku.php/mgiza:overview.

[12] http://www.speech.sri.com/projects/srilm/.

| Data Set | Lang. | Sent. | Words | Vocab. | Ave. Len. | Sites | Docs |
|---|---|---|---|---|---|---|---|
| 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 | | |
| Web-crawled In-domain Mono. Data | en | 8,448,566 | 280,211,580 | 3,047,758 | 33.16 | 26 | 1,601 |
| | cs | 44,198 | 1,280,326 | 137,179 | 28.96 | 4 | 388 |
| | de | 473,171 | 14,087,687 | 728,652 | 29.77 | 17 | 968 |
| | fr | 852,036 | 35,339,445 | 718,141 | 41.47 | 10 | 683 |

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

In the following sub-sections, we describe the results of **baseline systems**, which are trained on the official corpora. We also present the **enhanced systems** that make use of the web-crawled bilingual dictionary and monolingual data as the additional training resources. Two variants of enhanced system are constructed based on different filtering criteria.

### 5.1 Baseline System

The baseline systems is constructed based on the combination of TM adaptation and LM adaptation, where the corresponding selection thresholds ($M$) are manually tuned. Table 3 shows the BLEU scores of baseline systems as well as the threshold values of $M$ for general-domain monolingual corpora and parallel corpora selection, respectively.

By looking into the results, we find that en-cs system performs poorly, because of the limited in-domain parallel and monolingual corpora (shown in Table 2). While the fr-en and en-fr systems achieve the best scores, due the availability of the high volume training data. We experiment with different values of $M$={0, 25, 50, 75, 100} that indicates the percentages of sentences out of the general corpus used for constructing the LM adaptation and TM adaptation. After tuning the parameter $M$, we find that BLEU scores of different systems peak at different values of $M$. LM adaptation can achieve the best translation results for cs-en, en-fr and de-en pairs when $M$=25, en-cs and en-de pairs when $M$=50, and fr-en pair when $M$=75. While TM adaptation yields the best scores for en-fr and en-de pairs at $M$=25 and cs-en and fr-en pairs at $M$=50, de-en pair when $M$=75 and en-cs pair at $M$=100.

| Lang. Pair | BLEU | Mono. ($M$%) | Parallel ($M$%) |
|---|---|---|---|
| en-cs | 17.57 | 50% | 100% |
| cs-en | 31.29 | 25% | 50% |
| en-fr | 38.36 | 25% | 25% |
| fr-en | 44.36 | 75% | 50% |
| en-de | 18.01 | 50% | 25% |
| de-en | 32.50 | 25% | 75% |

Table 3: BLEU scores of baseline systems for different language pairs.

### 5.2 Based on Relevance Score Filtering

As described in Section 2.3, we use the relevance score to filter out the non-in-domain documents. Once again, we evaluate different values of

$K$={0, 25, 50, 75, 100} that represents the percentages of crawled documents we used for training the LMs. In Table 4, we show the absolute BLEU scores of the evaluated systems, listed with the optimized thresholds, and the relative improvements ( $\Delta$ %) in compared to the baseline system. The size of additional training data (for LM) is displayed at the last column.

| Lang. Pair | Docs ($K$%) | BLEU | Δ (%) | Sent. |
|---|---|---|---|---|
| en-cs | 50 | 17.59 | 0.11 | 31,065 |
| en-de | 75 | 18.52 | 2.83 | 435,547 |
| en-fr | 50 | 39.08 | 1.88 | 743,735 |
| cs-en | 75 | 32.22 | 2.97 | 7,943,931 |
| de-en | 25 | 33.50 | 3.08 | 4,951,189 |
| fr-en | 100 | 45.45 | 2.46 | 8,448,566 |

Table 4: Evaluation results for systems that trained on relevance-score-filtered documents.

The relevance score filtering approach yields an improvement of 3.08% of BLEU score for de-en pair that is the best result among the language pairs. On the other hand, en-cs pair obtains a marginal gain. The reason is very obvious that the training data is very insufficient. Empirical results of all language pairs expect fr-en indicate that data filtering is the necessity to improve the system performance.

### 5.3 Based on Moore-Lewis Filtering

In this approach, we need to determine the values of two parameters, top $K$ documents and top $N$ sentences, where $K$={100, 75, 50} and $N$={75, 50, 25}, $N < K$. When $K$=100, it is a conventional perplexity-based data selection method, i.e. no document will be filtered. Table 5 shows the combination of different $K$ and $N$ that gives the best translation score for each language pair. We provide the absolute BLEU for each system, together with relative improvements (Δ%) that compared to the baseline system.

| Lang. Pair | Docs ($K$%) | Target Size ($N$%) | BLEU | Δ (%) |
|---|---|---|---|---|
| en-cs | 50 | 25 | 17.69 | 0.68 |
| en-de | 100 | 50 | 18.03 | 0.11 |
| en-fr | 100 | 50 | 38.73 | 0.96 |
| cs-en | 100 | 25 | 32.20 | 2.91 |
| de-en | 100 | 25 | 33.10 | 1.85 |
| fr-en | 100 | 25 | 45.22 | 1.94 |

Table 5: Evaluation results for systems that trained on combination filtering approach.

In this shared task, we have a quality and quantity in-domain monolingual training data for English. All the systems that take English as the target translation always outperform the other reverse pairs. Besides, we found the systems based on the perplexity data selection method tend to achieve a better scores in BLEU.

## 6 Official Results and Conclusions

We described our study on developing unconstrained systems in the medical translation task of 2014 Workshop on Statistical Machine Translation. In this work, we adopt the web crawling strategy for acquiring the in-domain monolingual data. In detection the domain data, we exploited Moore-Lewis data selection method to filter the collected data in addition to the build-in scoring model provided by the crawler toolkit. However, after investigation, we found that the two methods are very competitive to each other.

The systems we submitted to the shared task were built using the language models and translation models that yield the best results in the individual testing. The official test set is converted into the *recased* and *detokenized* SGML format. Table 9 presents the official results of our submissions for every language pair.

| Lang. Pair | BLEU of Combined systems | Official BLEU |
|---|---|---|
| en-cs | 23.16 (+5.59) | 22.10 |
| cs-en | 36.8 (+5.51) | 37.40 |
| en-fr | 40.34 (+1.98) | 40.80 |
| fr-en | 45.79 (+1.43) | 43.80 |
| en-de | 19.36 (+1.35) | 18.80 |
| de-en | 34.17 (+1.67) | 32.70 |

Table 6: BLEU scores of the submitted systems for the medical translation task in six language pairs.

## 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.

## References

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.

M. Espl à-Gomis and M. L. Forcada. 2010. Combining Content-Based and URL-Based Heuristics toHarvest Aligned Bitexts from Multilingual Sites with Bitextor. *The Prague Bulletin of Mathemathical Lingustics*, 93:77–86.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pp. 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.

Papineni, Kishore, Salim Roukos, ToddWard, and-Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, USA.

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.

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.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 441-450.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. *Proceedings of ACL*, pp. 160-167.

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

P. Pecina, A. Toral, A. Way, V. Papavassiliou, P. Prokopidis, and M. Giagkou. 2011. Towards Using WebCrawled Data for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 297-304.

P. Pecina, A. Toral, V. Papavassiliou, P. Prokopidis, J. van Genabith, and R. I. C. Athena. 2012. Domain adaptation of statistical machine translation using web-crawled resources: a case study. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pp. 145-152.

P. Pecina, O. Dušek, L. Goeuriot, J. Hajič, J. Hlaváčová, G. J. Jones, and Z. Urešová. 2014. Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial intelligence in medicine*, pages 1-25.

Philip Resnik and Noah A. Smith. 2003. The Web as a parallel corpus. Computational Linguistics, 29:349–380

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.

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*, pp. 209-212.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. *Proceedings of the International Conference on Spoken Language Processing*, pp. 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*.

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