Experiments in Medical Translation Shared Task at WMT 2014

Jian Zhang, Xiaofeng Wu,
Iacer Calixto, Ali Hosseinzadeh Vahid, Xiaojun Zhang,
Andy Way, Qun Liu

The CNGL Centre for Global Intelligent Content
School of Computing
Dublin City University, Ireland
{zhangj,xiaofengwu,
icalixto,avahid,xzhang,
away,qliu}@computing.dcu.ie

Abstract

This paper describes Dublin City University's (DCU) submission to the WMT 2014 Medical Summary task. We report our results on the test data set in the French to English translation direction. We also report statistics collected from the corpora used to train our translation system. We conducted our experiment on the Moses 1.0 phrase-based translation system framework. We performed a variety of experiments on translation models, reordering models, operation sequence model and language model. We also experimented with data selection and removal the length constraint for phrase-pair extraction.

1 System Description

1.1 Training Data Statistics and Preparation

The training corpora provided to the medical translation shared task can be divided into 3 categories:

Medical in-domain corpora: these corpora contain documents, patents, articles, terminology lists, and titles that are representative of the same medical domain as the development and test data sets (Table 1, second column).

Medical out-of-domain corpora: these corpora also contain medical documents, patents, articles, terminologies lists and titles, but describe a different domain from the development and test data sets (Table 1, third column).

General domain corpora: these corpora consist of general-domain text (WMT 2014 general translation subtask corpora), and encompass various domains. (We did not use these corpora in our system).

Corpus	In-domain parallel sentence number	Out-of-domain parallel sentence number
EMEA	1,092,568	0
COPPA	664,658	2,841,849
PatTR-title	408,502	2,096,270
PatTR-abstract	688,147	3,009,523
PatTR-claims	1,105,230	5,861,621
UMLS	85,705	0
Wikipedia	8,448	0
TOTAL	4,053,258	13,809,263

Table 1: WMT 2014 Medical Translation shared task parallel training data before preprocessing.

Within all the provided training corpora from WMT 2014, 70.72% of the medical in domain bilingual sentences, and 100% of the medical out-of-domain bilingual sentences were obtained from patent document collections. Motivated by these percentages, we view the WMT 2014 medical translation shared task as similar to training a patent-specific translation system. The monolingual corpora are taken from 9 different corpora collections, and there is no clear demarcation of the in/out-of-domain boundaries (except the PatTR collection). Our method of differentiating between the in/out-of-domain monolingual corpora is that only English sentences from the third column of Table 1, and the patent description documents from PatTR collection, are out-of-domain monolingual corpora. All other English

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 260–265,
Baltimore, Maryland USA, June 26–27, 2014.

sentences are treated as an in-domain monolingual resource.

A patent document usually comprises title, abstract, claims and description fields. The documents often use its unique formatting and contain linguistic idiosyncrasies, which distinguish patent-specific translation systems from general translation systems, in both training and translation phases (Ceauşu et al., 2011). We have also found that some common writing styles are constantly used, especially for long sentences. For example, a typical patent claim begins with

```
Method of [X], which comprising:
```

followed by a numbered list. The abstract field normally contains one paragraph only, but with multiple sentences. Those long sentences are necessarily filtered out to facilitate efficient word alignment, using a tool such as GIZA++ (Och, 2003) word aligner with the default parameter settings. However, because statistical machine translation depends on the training data to estimate translation probability, more high quality training data often leads a better translation result. One possible method of including long sentences into the training cycle is to change the word aligner's parameter settings to handle longer sentences; however, aligning long sentences is time consuming. Our solution is to capture the styled long sentences and attempt to split them on both source and target side simultaneously according to the numbered list or sentence boundary indications. If the sentence number after splitting are matching in both source and target sides, and each sentence pair is within the token length ratio of 3, we assume the split attempt is successful, otherwise the sentences are kept unchanged and will be filtered out eventually. We applied our splitting attempt approach on the patent documents at the data preparation step which consequently results in 19.35% and 7.1% increase in the number of sentence pairs compared with the original medical in-domain (from 4053258 to 4837382) and overall medical (from 17862521 to 19124142) datasets respectively.

Another finding from the training corpora is that the titles of the patent documents are often capitalized in the training corpora. Since we are training a true-cased translation system, and the translation inputs contain non-title sentences, capitalized training sentences will contribute biased weights to our true-case model. We addressed this issue by creating a lowercase version of the title corpora, then we trained our true-case model with the lowercased titles corpora and other non-title corpora. We also included the lowercased title corpora in the translation system training.

We tokenized the training corpora using the tokenizer script distributed in the Moses 1.0 framework with additional patent document non-breaking preferences observed during data preparation, such as Figs and FIGS etc., and a modified aggressive setting (split hyphen character in all cases). Other data preparation steps included character normalization, character/token based foreign language detection, HTML/XML tag removal, case insensitive duplication removal, longer sentence removal (2-80, length ratio 9), resulting in the preprocessed data shown in Table 2.

Corpus	**In-domain parallel sentence number**	**Out-of-domain parallel sentence number**
EMEA	273,532	0
COPPA	1,374,371	6,075,599
PatTR-title	63,856	3,457,164
PatTR-abstract	599,435	2,595,515
PatTR-claims	876,603	4,244,324
UMLS	85,683	0
Wikipedia	8,438	0
TOTAL	3,956,478	16,372,602

Table 2: WMT 2014 Medical Translation shared task parallel training data after preprocessing steps.

1.2 Training Data Selection

It is an open secret that high quality and large quantity of the parallel corpus are the two most important factors for a high-quality SMT system. These factors assist the word aligner in producing a precise alignment model, which in turn brings benefits to the other SMT training steps.

The quantity factor also helps the SMT system to cover more translation input variations. In order to efficiently use the training corpora listed in Table 2, we explored some data selection methodologies. We used the feature decay algorithm (Bicici et al., 2014) to select the training instances transductively, using the source side of the test set. We built systems with the pre-defined selection proportions in token number, 1/64, 1/32, 1/16, 1/8, 1/2, 3/4 and 1 of all the in-domain medical training data, then searched for the best performing

system using the test data set as our baseline (Table 3). For the purpose of making the potential baseline systems comparable, instance selection was employed after word alignment using word aligner MGIZA++ (Gao and Vogel, 2008) on all the available data. The transductive learning uses features extracted from the source data of the development set with the default feature decay algorithm weight settings. All of systems were trained using the default phrase-based training parameter settings of Moses 1.0 framework, with additional msd-bidirectional-fe reordering model (Koehn et al., 2005). We extract phrase pairs based on grow-diag-final-and (Koehn et al., 2003) heuristics. The language model was created with open source IRSTLM toolkit (Federico et al., 2008) using all the English in-domain data (monolingual and parallel). We used 5-gram with modied Kneser-Ney smoothing (Kneser and Ney, 1995). The tuning step used minimum error rate training (MERT) (Och, 2003). The performance was measured by the test data set in case insensitive BLEU score.

Proportions	**Test set case insensitive BLEU**
1/64	0.4374
1/32	0.4409
1/16	0.4370
1/8	**0.4419**
1/4	0.4390
1/2	0.4399
3/4	0.4397
1	0.4260

Table 3: Feature decay algorithm transductive learning selection on all in-domain data using extracted features from the source side of the test data set. We choose system uses 1/8 proportions of the in-domain data as our baseline system.

Our results show that the system trained with 1/8 proportion of the in-domain medical training data (398,098 sentence pairs) selected by FDA outperformed the others. We chose this system as our baseline system.

2 Experiments

2.1 Maximum Phrase Length

While extracting phrase pairs, collecting longer phrases is not guaranteed to produce a better quality phrase table than the shorter settings, even setting the maximum phrase length to three can achieve top performance (Koehn et al., 2003). We take this WMT 2014 opportunity to study the capability of long phrase lengths (>=10). We trained translation models with phrase length setting from 10 to 15, employed them to our baseline system and compared the performance with the default setting (length = 7).

Phrase Length	**Phrase Table Entries**	**Test set case insensitive BLEU**
7 (Baseline)	19.31	0.4419
10	29.67	0.4400
11	32.87	0.4416
12	35.95	0.4444*
13	38.91	**0.4448***
14	41.75	0.4444*
15	44.47	0.4362

Table 4: -max-phrase-length setting experiment, where phrase table entries is in millions. * indiactes statistically significant improvement at the p = 0.05 level.[1]

As stated in (Koehn et al., 2003) and expected, the size of the phrase table is linear with respect to the maximum phrase length restriction. Surprisingly, we also found the performance can still improve after the default length setting, until a peak point (Table 4).

It is also interesting to see the effect for each sentence in the test set when the default phrase length setting in Moses framework is changed. We first evaluated the sentence level BLEU scores for the systems listed in Table 4, then compared them with our baseline system sentence level BLEU scores and categorised the compared results into increased, decreased or unaffected groups (Figure 1). We found that system with -max-phrase-length set to 12 is influenced the least (158, 118 and 724 sentences have BLEU score increased, decreased and unaffected respectively) and with -max-phrase-length sets to 10 is influenced the most (261, 257 and 482 sentences have BLEU score increased, decreased and unaffected respectively).

We then looked into the decoding phase and tried to discover the actual phrase length that was used to generate the translation outputs. We exposed the translation segmentations by triggering the -report-segmentation decoding parameter

[1]The same notation is used for the rest of the tables in this paper

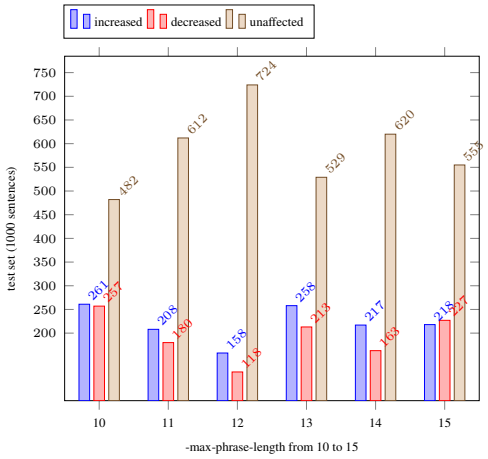

Figure 1: Sentence level BLEU score affects when enlarge -max-phrase-length

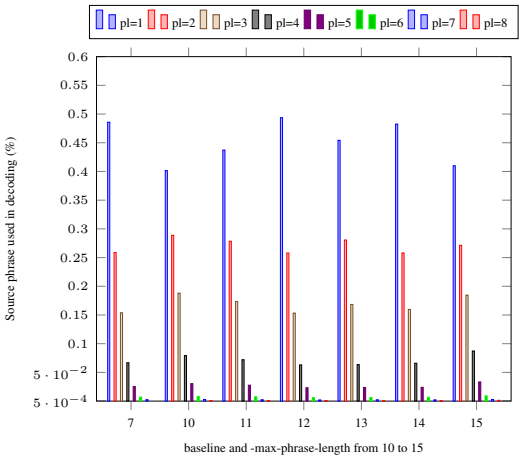

Figure 2: Phrase length (pl) distribution used in decoding

in the Moses framework and computed the percentage of different phrases used according to the phrase token number (Figure 2). The translation is mostly generated from short source phrases (length<4) in all the systems during decoding, which we think is the reason that setting phrase extraction to length 3 can achieve top performance.

We did not carry out more experiments in this case, as we think there is no absolute maximum phrase length setting which can fit into all experiments and such experiments depend on many factors, such as the similarity between the training corpus and then testing data. The choice to set -max-phrase-length to 13 is purely directed by the BLEU score shown in Table 4.

2.2 Reordering Models

Ceauşu et al. (2011) also found that long-range reordering is one of the characteristics of patent documents; however, long-range reordering increases the difficulty of SMT training and decoding. We experimented two approaches to address this challenge. Apart from the msd-bidirectional-fe lexical reordering model (Koehn et al., 2005) in our baseline system, the phrase-based orientation and hierarchical orientation reordering models (Galley and Manning, 2008) can capture long distance dependencies. The phrase-based orientation reordering model is similar to the lexical reordering approach, the only difference between these two models is the phrase-based reordering model performs reordering only on the phrase level, but the hierarchical reordering model does not have such constraint - it does not require phrases to be adjacent. OSM (Durrani, 2011) (Durrani, 2013b) is a sequence model integrating the N-gram-based translation model and reordering model. It defines three operations for reordering and considers all reordering possibilities within a fixed window while searching. We experimented with both reordering models, and found that the system defined with three reordering models performs better (Table 5) than OSM. We then tried to use both OSM and the reordering models together, which produced the best system at this point.

Systems	**Test set case insensitive BLEU**
Baseline + 13	0.4448
+ OSM	0.4472
+ pho-ho	0.4551*
+ pho-ho + OSM	**0.4561***

Table 5: Reordering Model or/and OSM results

2.3 Two Translation Models

The back-off model aims to produce translations for the unknown words or unknown phrases in the primary translation table by yielding the phrase table translation probability from primary translation table to the back-off table, as in (Koehn et al., 2012a)

$$p_{BO}(e|f) = \begin{cases} p_1(e|f) & if\ count_1(f) > 0 \\ p_2(e|f) & otherwise \end{cases}$$

Moreover, we look at using the back off model

as a domain adaptation approach, which is to constrain the translation options within the target domain unless no options can be found, in which case the translation will be selected from the back-off model.

Phrase table fill-up (Bisazza et al., 2011) is a very similar approach with back-off models, it collects and uses the phrase pairs from the out-of-domain phrase table only when the input is unavailable at the in-domain phrase table. It merges the in-domain and out-of-domain translation models into one, where the scores are taken from more reliable source. To distinguish the source of a phrase pair entry, fill-up assigns a binary value as an additional feature at the merged phrase table.

We trained our out-of-domain translation model separately using all of the out-of-domain medical data listed at Table 2 with the same parameter settings as our baseline system, then employed Moses's back-off model feature to pass the primary and back-off translation models to the decoder at tuning and translation time. The fill-up tool was sourced from (Bisazza et al., 2011) at Moses's distribution. Our experiment results (Table 6) show that the fill-up approach performed better than the back-off model approach.

Systems	**Test set case insensitive BLEU**
Baseline + 13 + pho-ho + OSM	0.4561
Back-off	0.4573
Fill-up	**0.4599***

Table 6: Back-off and fill-up experiment results

2.4 Language Model

Until now, we have reported our results using a language model trained with all in-domain medical data only. We also took the similar approach to (Koehn et al., 2007) and carried out language model experiments. We trained our out-of-domain language model with all the out-of-domain English sentences mentioned in section 1.1, then interpolated the in-domain and out-of-domain language model by optimizing the perplexity to the development data set. We received a similar picture to (Koehn et al., 2007), where the language model trained with only in-domain data performed the best (Table 7).

Our final submission for WMT 2014 Medical Translation shared task is the * system at Table 7.

Systems	**Test set case insensitive BLEU**
Baseline + 13 + pho-ho + OSM + Fill-up*	**0.4599**
out-of-domain LM	0.4461
interpolated LM	0.4592

Table 7: Language model experiment results

3 Conclusion

In this paper, we report our results on the WMT 2014 in the French to English translation direction. We shared our statistics for the bilingual corpora used to train our translation system. All systems were trained using the open source Moses 1.0 translation framework. Based on the feature set of Moses phrased-based translation system, we carried out our experiments on translation models, reordering models, operation sequence model and language model. We also experimented on data selection and releasing the length restriction while extracting phrase pairs.

4 Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. We would also like to acknowledge Ergun Bicici who gives suggestions at the data selection approach.

References

Alexandru Ceauşu, John Tinsley, Jian Zhang and Andy Way. 2011. *Experiments on domain adaptation for patent machine translation in the PLuTO project*, The 15th conference of the European Association for Machine Translation, Leuven, Belgium.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. *Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation.*, In International Workshop on Spoken Language Translation (IWSLT), San Francisco, CA.

Durrani, N., Schmid, H., and Fraser, A. 2011. *A Joint Sequence Translation Model with Integrated Reordering.*, The 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, USA.

Durrani, N., Fraser, A., Schmid, H., Hoang, H., and Koehn, P. 2013b. *Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT*, The 51th Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria.

Ergun Bicici and Deniz Yuret. 2014. *Optimizing Instance Selection for Statistical Machine Translation with Feature Decay Algorithms*, IEEE/ACM Transactions On Audio, Speech, and Language Processing (TASLP).

Franz J. Och and Hermann Ney. 2003. *A systematic comparison of various statistical alignment models*, Computational Linguistics, 29(1):1951.

Franz Josef Och. 2003. *Minimum error rate training in statistical machine translation*, The 41th Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. *IRSTLM: an open source toolkit for handling large scale language models*, Interspeech, Brisbane, Australia.

Michel Galley and Christopher D. Manning. 2008. *A simple and effective hierarchical phrase reordering model.* , The 2008 Conference on Empirical Methods in Natural Language Processing,pages 848856, Honolulu, Hawaii, October. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. *Parallel implementations of word alignment tool*, In Software Engineering, Testing, and Quality Assurance for Natural Language Processing, SETQA-NLP 2008, pages 49-57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne and David Talbot. 2005. *Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation*, International Workshop on Spoken Language Translation.

Philipp Koehn and Josh Schroeder. 2007. *Experiments in Domain Adaptation for Statistical Machine Translation*, The Second Workshop on Statistical Machine Translation, pages 224227, Prague.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne and David Talbot. 2003. *Statistical phrase-based translation*, 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 4854, Edmonton, Canada.

Philipp Koehn, and Barry Haddow. 2012. *Interpolated backoff for factored translation models.*, The 10th Conference of the Association for Machine Translation in the Americas (AMTA).

Reinhard Kneser and Hermann Ney 1995. *Improved backing-off for m-gram language modeling.*, IEEE International Conference on Acoustics, Speech and Signal Processing, pages 181184.