Uses of Monolingual In-Domain Corpora for Cross-Domain Adaptation with Hybrid MT Approaches

An-Chang Hsieh, Hen-Hsen Huang and Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan
{achsieh,hhhuang}@nlg.csie.ntu.edu.tw;hhchen@ntu.edu.tw

Abstract

Resource limitation is challenging for cross-domain adaption. This paper employs patterns identified from a monolingual in-domain corpus and patterns learned from the post-edited translation results, and translation model as well as language model learned from pseudo bilingual corpora produced by a baseline MT system. The adaptation from a government document domain to a medical record domain shows the rules mined from the monolingual in-domain corpus are useful, and the effect of using the selected pseudo bilingual corpus is significant.

1 Introduction

Bilingual dictionary and corpus are important resources for MT applications. They are used for lexical choice and model construction. However, not all resources are available in bilingual forms in each domain. For example, medical records are in English only in some countries. In such a case, only bilingual dictionary and monolingual corpus is available. Lack of bilingual corpus makes domain adaptation more challenging.

A number of adaptation approaches (Civera and Juan, 2007; Foster and Kuhn 2007; Foster et al., 2010, Matsoukas et al., 2009; Zhao et al., 2004) have been proposed. They address the reliability of a model in a new domain and count the domain similarities between a model and the in-domain development data. The domain relevance in different granularities including words, phrases, sentences, documents and corpora are considered. Ueffing et al. (2007) propose semi-supervised methods which use monolingual data in source language to improve translation performance. Schwenk (2008) present lightly-supervised training to generate additional training data from the translation results of monolingual data. To deal with the resource-poor issue, Bertoldi and Federico (2009) generate a pseudo bilingual corpus from the monolingual in-domain corpus, and then train a translation model from the pseudo bilingual corpus.

Besides counting similarities and generating pseudo bilingual in-domain corpus, text simplification (Zhu et al., 2010; Woodsend and Lapata, 2011; Wubben et al., 2012) is another direction. Simplifying a source language text makes the translation easier in a background MT system. Chen et al. (2012a) propose a method to simplify a sentence before MT and to restore the translation of the simplified part after MT. They focus on the treatments of input text only, but do not consider how to adapt the background MT to the specific domain. The translation performance depends on the coverage of the simplification rules and the quality of the background system.

This paper adopts the simplification-translation-restoration methodology (Chen et al., 2012a), but emphasizes on how to update bilingual translation rules, translation model and language model, which are two kernels of rule-based and statistics-based MT systems, respectively. This paper is organized as follows. Section 2 specifies the proposed hybrid MT approaches to resource-limited domains. The characteristics of available resources including their types, their linguality, their belonging domains, and their belonging languages are analyzed and their uses in translation rule mining and model construction are presented. Section 3 discusses how to adapt an MT system from a government document domain to a medical record domain. The experimental setups reflect various settings. Section 4 concludes the remarks.

Proceedings of the Second Workshop on Hybrid Approaches to Translation, pages 117–122,
Sofia, Bulgaria, August 8, 2013.

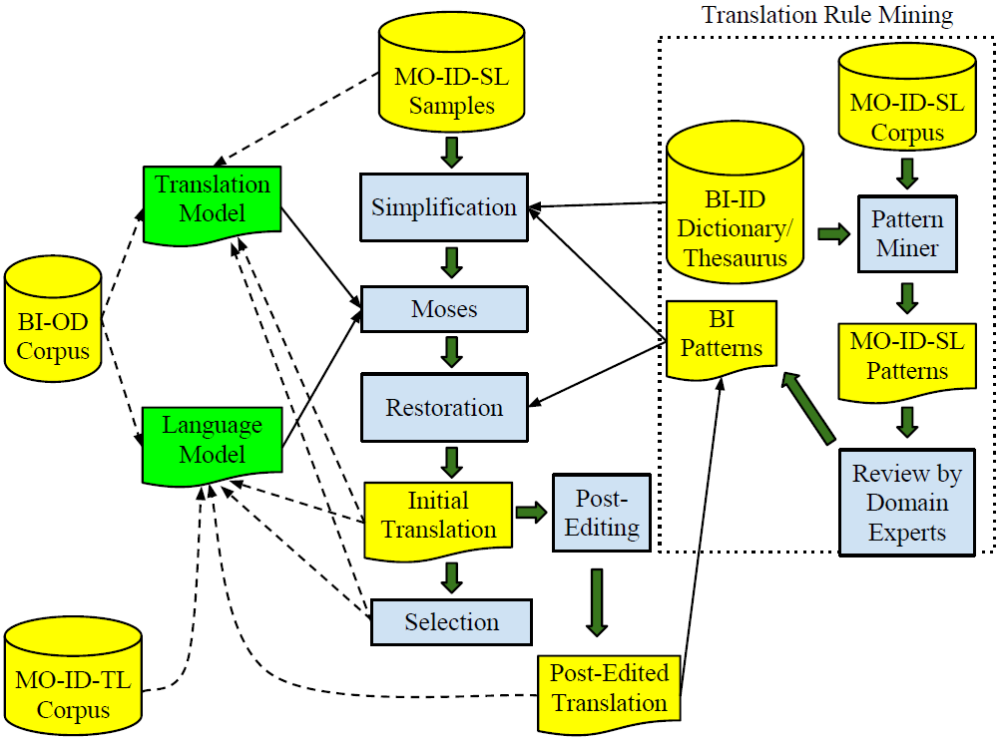

Figure 1: Hybrid MT Approaches

2 Hybrid MT Approaches

Figure 1 sketches the overall picture of our proposed hybrid MT approaches. A resource is represented in terms of its *linguality*, *domain*, *language*, and *type*, where MO/BI denotes monolingual/bilingual, ID/OD denotes in-domain/out-domain, and SL/TL denotes source language/target language. For example, an MO-ID-SL corpus and an MO-ID-TL corpus mean monolingual in-domain corpora in source and in target languages, respectively. Similarly, a BI-OD corpus and a BI-ID dictionary denote a bilingual out-domain corpus, and a bilingual in-domain dictionary, respectively.

Resources may be provided by some organizations such as LDC, or collected from heterogeneous resources. The MO-ID-SL/TL corpus, the BI-OD corpus, and the BI-ID dictionary belong to this type. Besides, some outputs generated by the baseline MT systems are regarded as other kinds of resources for enhancing the proposed methods incrementally. Initial translation results, selected translation results, and post-edited translation results, which form pseudo bilingual in-domain corpora, belong to this type.

The following subsections first describe the baseline systems with the original resources and then specify the advanced systems with the generated resources.

2.1 A baseline translation system

In an extreme case, only a bilingual out-domain corpus, a monolingual in-domain corpus in source/target language, a bilingual in-domain dictionary and a monolingual in-domain thesaurus in source language are available. The bilingual out-domain corpus is used to train translation and language models by Moses. They form a background out-domain translation system.

A pattern miner is used to capture the written styles in the monolingual in-domain corpus in source language. A monolingual in-domain thesaurus in source language is looked up to extract the class (sense) information of words. Monolingual patterns are mined by counting frequent word/class n-grams. Then, the bilingual in-domain dictionary is introduced to formulate translation rules based on the mined monolingual patterns. Here in-domain experts may be involved in reviewing the bilingual rules. The human cost will affect the number of translation rules formulated and thus its coverage.

The baseline translation system is composed of four major steps shown as follows. (1) and (2) are pre-processing steps before kernel MT, and (4) is a post-processing step after kernel MT.

(1) Identifying and translating in-domain segments from an input sentence by using translation rules.

(2) Simplifying the input sentence by replacing the in-domain segments as follows.
   (a) If an in-domain segment is a term in the bilingual in-domain dictionary, we find a related term (i.e., hypernym or synonym) in the in-domain thesaurus which has relatively more occurrences in the background SMT system to replace the term.
   (b) If an in-domain segment is a noun phrase, we keep its head only, and find a related term of the head as (a).
   (c) If an in-domain segment is a verb phrase composed of a verb and a noun phrase, we keep the verb and simplify the noun phrase as (b).
   (d) If an in-domain segment is a verb phrase composed of a verb and a prepositional phrase, we keep the verb and remove the prepositional phrase if it is optional. If the prepositional phrase is mandatory, it is kept and simplified as (e).
   (e) If an in-domain segment is a prepositional phrase, we keep the preposition and simplify the noun phrase as (b).
   (f) If an in-domain segment is a clause, we simplify its children recursively as (a)-(e).
(3) Translating the simplified source sentence by using the out-domain background MT system.
(4) Restoring the results of the bilingual in-domain segments translated in (1) back to the translation results generated in (3). The restoration is based on the internal alignment between the source and the target sentences.

2.2 Incremental learning

There are several alternatives to update the baseline translation system incrementally. The first consideration is the in-domain translation rules. They are formed semi-automatically by domain experts. The cost of domain experts results that only small portion of n-gram patterns along with the corresponding translation are generated. The post-editing results suggests more translation rules and they are fed back to revise the baseline translation system.

The second consideration is translation model and language model in the Moses. In an ideal case, the complete monolingual in-domain corpus in source language is translated by the baseline translation system, then the results are post-edited by domain experts, and finally the complete post-edited bilingual corpus is fed back to revise both translation model and language model. However, the post-editing cost by domain experts is high. Only some samples of the initial translation are edited by domain experts. On the one hand, the sampled post-edited in-domain corpus in target language is used to revise the language model. On the other hand, the in-domain bilingual translation result before post-editing is used to revise the translation model and the language model. Size and translation quality are two factors to be considered. We will explore the effect of different size of imperfect in-domain translation results on refining the baseline MT system. Moreover, a selection strategy, e.g., only those translation results completely in target language are considered, is introduced to sample "relatively more accurate" bilingual translation results.

In the above incremental learning, translation rules, translation model and language model are revised individually. The third consideration is to merge some refinements together and examine their effects on the translation performance.

3 Cross-Domain Adaptation

To evaluate the feasibility of the proposed hybrid MT approaches, we adapt an English-Chinese machine translation system from a government document domain to a medical record domain. The linguistic resources are described first and then the experimental results.

3.1 Resource description

Hong Kong parallel text (LDC2004T08), which contains official records, law codes, and press releases of the Legislative Council, the Department of Justice, and the Information Services Department of the HKSAR, respectively, and UN Chinese-English Parallel Text collection (LDC2004E12) is used to train the translation model. These two corpora contain total 6.8M sentences. The Chinese counterpart of the above parallel corpus and the Central News Agency part of the Tagged Chinese Gigaword (LDC2007T03) are used to train trigram language model. These two corpora contain total 18.8M sentences. The trained models are used in Step (3) of the baseline translation system.

Besides the out-domain corpora for the development of translation model and language model, we select 60,448 English medical records (1.8M sentences) from National Taiwan University

Hospital (NTUH) to learn the n-gram patterns. Metathesaurus of the Unified Medical Language System (UMLS) provides medical classes of in-domain words. A bilingual medical domain dictionary composed of 71,687 pairs is collected. Total 7.2M word/class 2-grams~5-grams are identified. After parsing, there remain 57.2K linguistic patterns. A higher order pattern may be composed of two lower order patterns. Keeping the covering patterns and ruling out the covered ones further reduce the size of the extracted patterns. The remaining 40.1K patterns are translated by dictionary look-up. Because of the high cost of medical record domain experts (i.e., physicians), only a small portion is verified. Finally, 981 translation rules are formulated. They are used in Step (1) of the baseline MT system. The detail rule mining and human correction process please refer to Chen et al. (2012b).

We further sample 2.1M and 1.1M sentences from NTUH medical record datasets, translate them by the baseline MT system, and get 2.1M- and 1.1M-pseudo bilingual in-domain corpora. We will experiment the effects of the corpus size. On the other hand, we apply the selection strategy to select 0.95M "good" translation from 2.1M-pseudo bilingual in-domain corpus. Furthermore, some other 1,004 sentences are post-edited by the domain experts. They are used to learn the advanced MT systems.

To evaluate the baseline and the advanced MT systems, we sample 1,000 sentences different from the above corpora as the test data, and translate them manually as the ground truth.

3.2 Results and discussion

Table 1 lists the methods along with the resources they used. B is the baseline MT system. Most patterns appearing in the 57.2K learned n-grams mentioned in Section 3.1 are not reviewed by physicians due to their cost. Part of these unreviewed patterns may occur in the post-edited data. They will be further introduced into M1. In the experiments, patterns appearing at least two times in the post-edited result are integrated into M1. Total 422 new patterns are identified. Translation model and language model in M1 is the same as those in baseline system.

In M2-M6, the translation rules are the same as those in baseline MT system, only translation model and/or language model are re-trained. In

	Translation Rules	Translation Model	Language Model	Tuning Data
B	981 bilingual translation rules	6.8M government domain bilingual sentences	18.8M government/news domain Chinese sentences	1000 government domain bilingual sentences
M1	981 bilingual translation rules + 422 mined rules from post-editing	6.8M government domain bilingual sentences	18.8M government/news domain Chinese sentences	200 post-edited medical domain sentences
M2	981 bilingual translation rules	6.8M government domain bilingual sentences	804 post-edited Chinese sentences	200 post-edited medical domain sentences
M3	981 bilingual translation rules	6.8M government domain bilingual sentences	30,000 Chinese sentences selected from medical literature	200 post-edited medical domain sentences
M4	981 bilingual translation rules	1.1M pseudo medical domain bilingual sentences generated by M1	1.1M pseudo medical domain Chinese sentences generated by M1	200 post-edited medical domain sentences
M5	981 bilingual translation rules	2.1M pseudo medical domain bilingual sentences generated by M1	2.1M pseudo medical domain Chinese sentences generated by M1	200 post-edited medical domain sentences
M6	981 bilingual translation rules	0.95M selected pseudo medical domain bilingual sentences generated by M1	0.95M selected pseudo medical domain Chinese sentences generated by M1	200 post-edited medical domain sentences
M12	981 bilingual translation rules + 422 mined rules from post-editing	6.8M government domain bilingual sentences	804 post-edited Chinese sentences	200 post-edited medical domain sentences
M13	981 bilingual translation rules + 422 mined rules from post-editing	6.8M government domain bilingual sentences	30,000 medical domain Chinese sentences	200 post-edited medical domain sentences
M14	981 bilingual translation rules + 422 mined rules from post-editing	1.1M pseudo medical domain bilingual sentences generated by M1	1.1M pseudo medical domain Chinese sentences generated by M1	200 post-edited medical domain sentences
M15	981 bilingual translation rules + 422 mined rules from post-editing	2.1M pseudo medical domain bilingual sentences generated by M1	2.1M pseudo medical domain Chinese sentences generated by M1	200 post-edited medical domain sentences
M16	981 bilingual translation rules + 422 mined rules from post-editing	0.95M selected pseudo medical domain bilingual sentences generated by M1	0.95M selected pseudo medical domain Chinese sentences generated by M1	200 post-edited medical domain sentences

Table 1: Resources used in each hybrid MT method

Method	Bleu	Method	Bleu	Method	Bleu	Method	Bleu	Method	Bleu	Method	Bleu
B	28.04	M2	39.45	M3	32.03	M4	34.86	M5	35.09	M6	40.48
M1	39.72	M12	39.72	M13	32.85	M14	35.11	M15	35.52	M16	40.71

Table 2: BLEU of each hybrid MT method

M2, 804 post-edited sentences are used to train a new language model, without changing the translation model. In M3, paper abstracts in medical domain are used to derive a new language model. M4, M5 and M6 are similar except that different sizes of corpora are used. M4 and M5 use 1.1M and 2.1M sentences, respectively, while M6 uses 0.95M sentences chosen by using the selection strategy. M12-M16 are combinations of M1 and M2-M6, respectively. Translation rules, translation model and language model are refined by using different resources. Total 200 of the 1,004 post-edited sentences are selected to tune the parameters of Moses in the advanced methods.

Table 2 shows the BLEU of various MT methods. The BLEU of the MT system without employing simplification-translation-restoration methodology (Chen et al., 2012a) is 15.24. Apparently, the method B, which employs the methodology, achieves the BLEU 28.04 and is much better than the original system. All the enhanced systems are significantly better than the baseline system B by t-test (p<0.05). Comparing M1 and M12-M16 with the corresponding systems, we can find that introducing the mined patterns has positive effects. M1 is even much better than B. Although the number of the post-edited sentences is small, M2 and M12 show such a resource has the strongest effects. The results of M3 and M13 depict that 30,000 sentences selected from medical literature are not quite useful for medical record translation. Comparing M4 and M5, we can find larger pseudo corpus is useful. M6 shows using the selected pseudo subset performs much better. Comparing the top 4 methods, the best method, M16, is significantly better than M12 and M1 (p<0.05), but is not different from M6 significantly (p=0.1662).

We further analyze the translation results of the best methods M6 and M16 from two perspectives. On the one hand, we show how the mined rules improve the translation. The following list some examples for reference. The underlined parts are translated correctly by new mined patterns in M16.

(1) Example: Stenting was done from distal IVC through left common iliac vein to external iliac vein.
M6: 支架置入術 是 從 遠端 下腔靜脈 通過 從 左髂總靜脈 到 髂外靜脈 。
M16: 完成 支架置入術 從 遠端 下腔靜脈 通過 從 左髂總靜脈 到 髂外靜脈 。

(2) Example: We shifted the antibiotic to cefazolin.
M6: 我們 把 抗生素 頭孢唑啉 。
M16: 我們 把 抗生素 更換 為 頭孢唑啉 。

(3) Example: Enhancement of right side pleural, and mild pericardial effusion was noted .
M6: 增強 方面 的 權利 胸腔 、 和 發現 有 輕微 的 心包積液 。
M16: 增強 的 右 胸腔 、 輕微 心包積液 被 注意到 。

On the other hand, we touch on which factors affect the translation performance of M16. Three factors including word ordering errors, word sense disambiguation errors and OOV (out-of-vocabulary) errors are addressed as follows. The erroneous parts are underlined.

(1) Ordering errors
Example: Antibiotics were discontinued after 8 days of treatment.
M16: 抗生素 中斷 後 8 天 的 治療 。
Analysis: The correct translation result is "8 天 的 治療 後 抗生素 中斷。" The current patterns are 2-5 grams, so that the longer patterns cannot be captured.

(2) Word sense disambiguation errors
Example: After tracheostomy, he was transferred to our ward for post operation care.
M16: 氣管切開術 後 , 他 被 轉送到 我們 病房 為 員額 關懷 行動 。
Analysis: The correct translation of "post operation care" should be "術後照護". However, the 1,004 post-edited sentences are still not large enough to cover the possible patterns. Incremental update will introduce more patterns and may decrease the number of translation errors.

(3) OOV errors
Example: Transcatheter intravenous urokinase therapy was started on 1/11 for 24 hours infusion.
M16: transcatheter 靜脈 尿激酶 在 1/11 開始 進行 治療 24 小時 輸液 。
Analysis: The word "transcatheter" is an OOV. Its translation should be "導管".

4 Conclusion

This paper considers different types of resources in cross-domain MT adaptation. Several methods are proposed to integrate the mined transla-

tion rules, translation model and language model. The adaptation experiments show that the rules mined from the monolingual in-domain corpus are useful, and the effect of using the selected pseudo bilingual corpus is significant.

Several issues such as word ordering errors, word sense disambiguation errors, and OOV errors still remain for further investigation in the future.

Acknowledgments

This work was partially supported by National Science Council (Taiwan) and Excellent Research Projects of National Taiwan University under contracts NSC101-2221-E-002-195-MY3 and 102R890858. We are very thankful to National Taiwan University Hospital for providing NTUH the medical record dataset.

References

N. Bertoldi and M. Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189.

H.B. Chen, H.H. Huang, H.H. Chen and C.T. Tan. 2012a. A simplification-translation-restoration framework for cross-domain SMT applications. In *Proceedings of COLING 2012*, pages 545–560.

H.B. Chen, H.H. Huang, J. Tjiu, C.Ti. Tan and H.H. Chen. 2012b. A statistical medical summary translation system. In *Proceedings of 2012 ACM SIGHIT International Health Informatics Symposium*, pages. 101-110.

J. Civera and A. Juan. 2007. Domain adaptation in statistical machine translation with mixture modeling. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 177–180.

G. Foster and R. Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135.

G. Foster, C. Goutte, and R. Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of EMNLP 2010*, pages 451–459.

S. Matsoukas, A.I. Rosti, and B. Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of EMNLP 2009*, pages 708–717.

H. Schwenk. 2008. Investigations on large-scale lightly-supervised training. In *Proceedings of IWSLT 2008*, pages 182–189.

N. Ueffing, G. Haffari and A. Sarkar. 2007. Transductive learning for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 25–32,

K. Woodsend and M. Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of EMNLP 2011*, pages 409–420.

S. Wubben and A. van den Bosch, and E. Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of ACL 2012*, pages 1015–1024.

B. Zhao, M. Eck, M. and S. Vogel. 2004. Language model adaptation for statistical machine translation via structured query models. In *Proceedings of COLING 2004*, pages 411–417.

Z. Zhu, D. Bernhard, and I. Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of COLING 2010*, pages 1353–1361.