A Simplification-Translation-Restoration Framework for Cross-Domain SMT Applications

Han-Bin Chen[1], Hen-Hsen Huang[1], Hsin-Hsi Chen[1], and Ching-Ting Tan[2]
(1) Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
(2) National Taiwan University Hospital, Taipei, Taiwan
{hbchen, hhhuang}@nlg.csie.ntu.edu.tw,
{hhchen, tanct5222}@ntu.edu.tw

ABSTRACT

Integration of domain specific knowledge into a general purpose statistical machine translation (SMT) system poses challenges due to insufficient bilingual corpora. In this paper we propose a **s**implification-**t**ranslation-**r**estoration (*STR*) framework for domain adaptation in SMT by simplifying domain specific segments of a text. For an in-domain text, we identify the critical segments and modify them to alleviate the data sparseness problem in the out-domain SMT system. After we receive the translation result, these critical segments are then restored according to the provided in-domain knowledge. We conduct experiments on an English-to-Chinese translation task in the medical domain and evaluate each step of the *STR* framework. The translation results show significant improvement of our approach over the out-domain and the naïve in-domain SMT systems.

用於跨領域統計式機器翻譯系統之簡化-翻譯-還原架構

摘要

因為雙語語料的不足，將特定領域知識整合到一般用途的統計式機器翻譯(SMT)系統具有相當挑戰性。在本篇論文中，我們提出一個簡化-翻譯-還原(*STR*)的架構，藉由簡化特定領域的片段來達成SMT的領域調適。對於一篇領域內的文字，我們首先辨識其中重要的片段再做修改以減輕領域外SMT系統的資料稀疏問題。我們取得翻譯結果後，根據提供的領域內知識將這些重要片段還原。最後我們進行了醫療領域的英中翻譯的實驗，並且評估*STR*架構內的每一步驟。翻譯結果顯示我們的方法顯著地優於領域外的SMT系統，以及簡易型的領域內SMT系統。

KEYWORDS : Cross-Domain SMT, Domain Adaptation, Statistical Machine Translation
關鍵詞 : 跨領域統計式機器翻譯, 領域調適, 統計式機器翻譯

Proceedings of COLING 2012: Technical Papers, pages 545–560,
COLING 2012, Mumbai, December 2012.

1 Introduction

Over the past decades, the rapid growth of available parallel corpus makes SMT development feasible, and SMT system has gradually moved toward practical use because of its relatively acceptable translation speed and quality. A phrase-based SMT system (Koehn et al., 2003; Koehn, 2004), for example, trains a phrase table from a large bilingual corpus as its translation model, and decodes source language input in polynomial time with greedy algorithms such as beam search. It translates phrases as basic units, and thus captures short-range reordering phenomena between source and target languages. Generally phrase-based SMT models outperform word-based ones (Koehn et al., 2003). However, an SMT system fails to capture long-range contextual knowledge due to the limited horizon and the sparseness nature of lexical n-grams. These drawbacks reduce the translation quality in terms of translation and reordering errors, especially when limited bilingual corpus is available for estimating translation model.

The data sparseness problem may worsen when we build an SMT system for a specific domain but have small or no bilingual in-domain corpus. The translation performance could be seriously degraded since the SMT system cannot gather the statistical evidence of a segment containing out-of-vocabulary (OOV) words. For some domains, a bilingual dictionary containing source-target term pairs is available. With such in-domain knowledge, one can force an SMT system to translate OOV terms according to the dictionary. In this way, we correctly translate in-domain terms. However, translation quality may still be unsatisfying under this naïve integration because in-domain terms are rare or unseen in the background SMT model. Hence the context of these in-domain terms can hardly be captured. FIGURE 1 shows an incorrect English-Chinese translation of a sentence in a medical record by the online Google Translate service. In this example, it correctly translates the diagnosis term "crystal induced arthritis". However, it mistranslates its nearby phrasal verb "suffered from" by translating these two words separately. This example shows an SMT system may recognize the domain specific terms with either bilingual dictionary or its background SMT model, but still gives improper translations of the surrounding words due to insufficient context knowledge of these in-domain terms.

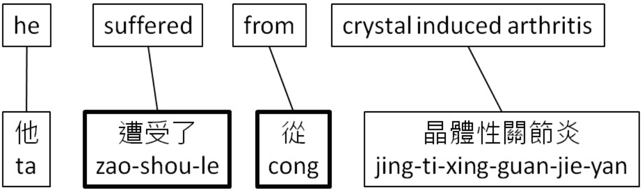


FIGURE 1 – Translating domain specific text with Google Translate. The bold target phrases are the inappropriate translations.

Our work originates from the following idea: modifying an in-domain segment in a source text such that the background SMT model not only recognizes it but also translates it together with its context words. Based on the motivation, for example, we modify the complicated diagnosis term "crystal induced arthritis" in FIGURE 1 into a more general term that occurs more frequently in the background SMT model, such as "cancer", "pneumonia" or "hypertension". Since these terms are more common in the general corpora, the general SMT system is able to better handle the modified text. Take the common word "cancer" as an example, the Google Search engine shows that Web pages containing "suffered from cancer" significantly outnumber those containing "suffered from crystal induced arthritis".

In this paper, we propose a **s**implification-**t**ranslation-**r**estoration (*STR*) framework to address domain adaptation in SMT by modifying in-domain text in favour of the background SMT system. The *STR* framework includes four steps to produce a cross-domain translation with higher quality than a general purpose SMT system. More specifically, to translate an in-domain source text, we identify the domain-specific segments and simplify them into more general expressions. The modified text is then translated by an out-domain SMT system. After that, we manipulate the translation result by replacing the modified parts and their translations with the correct bilingual segments in our in-domain knowledge (e.g., bilingual dictionary). Our framework is suited to the cross-domain SMT scenario where an out-domain SMT system and bilingual in-domain dictionaries are available. We show the effectiveness of the framework through a case study by building an SMT system for the medical domain. In this domain, OOV is a frequent problem if a general purpose SMT system is applied. On the other hand, there are plenty of bilingual dictionaries in this area. We implement the *STR* framework and report the experimental results of the translation tasks on medical summaries in a hospital. The central issues in this framework include (1) collecting bilingual in-domain knowledge and identifying in-domain segments in a source text, (2) replacing the in-domain segments with the proper simplified forms, (3) translating the modified text with a background SMT system and (4) restoring the original in-domain segments after receiving the translation results from the background SMT system.

We are not the first to rephrase source language text in order to improve SMT output. As a pilot study, Resnik et al. (2010) proposed a targeted paraphrasing approach which identifies the critical source segments difficult for the background SMT system to translate. These segments are then manually paraphrased in many ways in order to provide the SMT system with more choices of decoding paths. Different from their work, we automatically identify the critical segments with in-domain knowledge, simplify them with linguistic information, and restore these critical segments after receiving the SMT results.

This paper is organized as follows. In Section 2, we review the previous works on domain adaptation and related work of our approach in SMT. In Section 3, we formally describe our simplification-translation-restoration framework to deal with domain adaptation in SMT. In Section 4, we evaluate the effectiveness of our *STR* framework by conducting a case study of English-Chinese medical summary translation, and discuss the experimental results. Section 5 draws the conclusion, indicates the potentials of our method, and shows some future work.

2 Related Work

Building an SMT system from large scale bilingual data for a specific application has become a practical option today. On the other hand, SMT model heavily relies on the statistical evidences in the training corpus. As a result, it may learn a biased SMT model, and suffer from the data sparseness problem of the training corpus when dealing with the ambiguity nature of human language. This drawback results in some typical issues such as translation disambiguation problem (Carpuat et al., 2007; Chan et al., 2007), in which a word has several senses, but the corpus is biased towards a particular subset of the senses. The SMT model trained from such a corpus is therefore prone to give the wrong translation due to the wrong choice of sense.

When a cross-domain SMT application is concerned, data sparseness problem is worsened by limited in-domain bilingual corpus. Domain adaptation techniques therefore play a key role in building an in-domain SMT system under a resource poor environment. A number of adaptation approaches have been proposed by leveraging either bilingual or monolingual in-domain resources. Foster and Kuhn (2007) proposed a mixture-model approach that divides and trains a bilingual corpus into several models. Different models were then weighted by estimating the similarity between a model and the in-domain development data. Matsoukas et al. (2009) devised sentence-level features and weighted the domain relevance to each sentence in the bilingual training corpus by optimizing an objective function. Foster et al. (2010) further raised the granularity by weighting at the level of phrase pairs. Similarly, a mixture-model approach was also applied in word-alignment task (Civera and Juan, 2007). Zhao et al. (2004) applied information retrieval techniques to select in-domain documents from large monolingual text collections and enhanced the baseline language model. Bertoldi and Federico (2009) exploited an in-domain monolingual corpus to synthesize a pseudo bilingual corpus and trained an in-domain translation model from the synthesized corpus.

While previous works concentrated on model and parameter estimation to achieve domain adaptation, they worked on the data sets with similar lexicons. Few studies dealt with large domain gap, which is a practical issue for a cross-domain SMT system. In the medical domain, for example, a term may not even appear in training corpus and therefore SMT system gives no translation to it. This OOV problem is common when translating domain specific terms such as diagnosis and surgical names in biomedical literature or medical records using a general purpose SMT system. Different from the previous works, we address cross-domain issues in SMT across two largely distinct domains by simplifying in-domain segments to the ones that can be recognized by the out-domain or the background SMT system, and restoring the in-domain segments after receiving the SMT results.

Text simplification (Zhu et al., 2010; Woodsend and Lapata, 2011; Wubben et al., 2012) itself has some straightforward NLP applications. For example, we produce a simpler version of a text by modifying the lexical contents and shortening the grammatical structures without changing the original text at the semantic level. Such simplified contents are beneficial for language learners and people with lower levels of literacy. One of the real world applications is Simple English Wikipedia (http://simple.wikipedia.org), which uses simple English words and grammar, and thus English language learners can benefit from it. In this paper we apply sentence simplification techniques to improve machine translation quality. The source language input is simplified into the version that makes it easier for the SMT system to translate. This simplification step serves as a pre-processing module of the out-domain SMT model which has poor in-domain knowledge.

The simplification approach can be viewed as a variant of paraphrasing, which expresses the same meaning in different ways. Paraphrasing is employed for various NLP tasks, such as machine translation, natural language generation and computer assisted language learning. For SMT task, paraphrasing is often used to alleviate the data sparseness problem in translation model. For example, we paraphrase a source language text so that the paraphrased parts are easier for the background SMT system to translate. Callison-Burch et al. (2006) pioneered a pivoting approach through parallel corpus to improve phrase-based SMT model. Marton et al. (2009) proposed a monolingual framework to select paraphrases of a term by comparing its context with those of candidate paraphrases. Aziz et al. (2010) proposed a semi-automatic approach to mine paraphrases from hypernyms and hyponyms in ontology. Resnik et al. (2010)

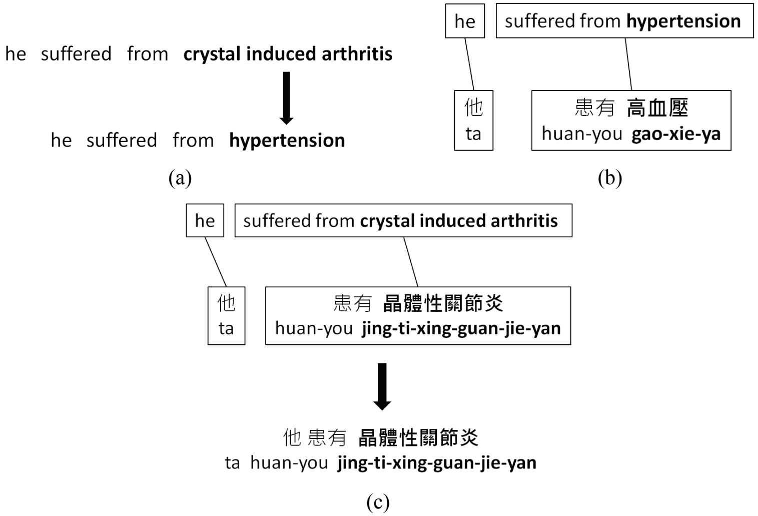

he suffered from **crystal induced arthritis**
he suffered from **hypertension**
(a)
he
suffered from **hypertension**
他
ta
患有 **高血壓**
huan-you **gao-xie-ya**
(b)
he
suffered from **crystal induced arthritis**
他
ta
患有 **晶體性關節炎**
huan-you **jing-ti-xing-guan-jie-yan**
他 患有 **晶體性關節炎**
ta huan-you **jing-ti-xing-guan-jie-yan**
(c)

FIGURE 2 – The idea behind our *STR* framework applied to the bold phrases. We (a) simplify the source diagnosis term before translation, (b) translate the simplified sentence with an SMT system, and (c) restore the original diagnosis term and produce the translation result.

conducted a pilot study of targeted paraphrasing in which monolingual speakers on both sides collaborate to improve SMT output by paraphrasing the critical segments of source text.

Different from previous studies that applied paraphrasing to SMT in the general domain, our work focuses on addressing cross-domain issues. We aim at adapting in-domain knowledge into the out-domain SMT system in a more smooth fashion than the naïve integration approach. While other works paraphrase general segments to provide more decoding options, we concentrate on domain specific segments which account for the performance degradation of a cross-domain SMT system. We try to identify these in-domain segments and simplify them to better fit the background out-domain SMT model.

3 A Simplification-Translation-Restoration (*STR*) Framework

We express the basic idea behind our *STR* framework through the example of the incorrect translation result shown in FIGURE 1. In order to fit the in-domain term to its general context, we change the obscure medical term to a more general one. In this example, the rare source medical term "crystal induced arthritis" in FIGURE 1 is thus simplified to the more public diagnosis term "hypertension", as illustrated in FIGURE 2(a). This simplified sentence is then sent to Google Translate. FIGURE 2(b) shows the translation result, which gives not only the correct translation of the diagnosis term "高血壓", but also the nearby context "患有". Moreover, the phrase "suffered from hypertension" is translated as a unit, implying that the translation model capture the context distribution of the simple diagnosis term "hypertension". Finally, to acquire the actual translation rather than the simplified one, the simplified parts are then cut out and the original bilingual in-domain terms are restored as illustrated in FIGURE 2(c). The final translation result is correct and fluent in contrast with the one in FIGURE 1.

The proposed *STR* framework is composed of four steps as follows.

1. Identifying in-domain segments $s_1, s_2, \ldots, s_n$ from an input sentence S.
2. Simplifying $s_1, s_2, \ldots, s_n$ in S and deriving a new source sentence S'.
3. Translating the source sentence S' into a target sentence T'.
4. Restoring the bilingual in-domain segments s_1-t_1, s_2-t_2, …, s_n-t_n back to S'-T' and deriving the final translation result T.

The following subsections describe each of them in detail.

3.1 Identification

An SMT system performs worse when translating segments which are rare in the background model. Therefore, in the first step of our framework, we identify the in-domain segments in an input source text for the next simplification step. To this end, we collect bilingual in-domain resources which include source-target string pairs.

Although a parallel corpus may not be available in a special domain, there are various ways to collect bilingual in-domain knowledge. For example, bilingual dictionaries can be found in the specific areas with long histories, such as medicine, physics and economics. They provide in-domain terminology with high quality and less noise. Not limited to the hand-crafted dictionaries, the bilingual in-domain knowledge may include phrase pairs or synchronous grammar rules, depending on the translation model and the decoding style of our background SMT system. Such bilingual knowledge can be collected by using automatic approaches (Wu and Chang, 2007; Haghighi et al., 2008) or semi-automatic approaches (Morin et al., 2007; Chen et al., 2011).

3.2 Simplification

In the simplification step, the identified in-domain segments of a text are transformed into the more general expressions. The modified text is then ready to be translated by the background SMT system. The simplification step serves as a pre-processing step before translation. We simplify an in-domain segment according to its type – say, **terminological unit** and **syntactic unit**, in this study.

Terminological units refer to terms that appear in domain specific dictionaries and glossaries without specifiers and modifiers. For these in-domain terms, we simplify them by finding the related terms such as hypernyms or synonyms which have relatively more occurrences and contextual information in the background SMT model. Syntactic units are linguistically meaningful segments which constitute special writing styles of a domain. These units usually bear syntactic categories at clausal or phrase levels such as S, NP, VP, PP, etc. They contain heads along with their modifiers. These syntactic categories can be derived from the parsed or the chunked results. We simplify a syntactic unit based on the rule corresponding to its syntactic category shown as follows.

(a) NP (Noun Phrase)

We keep the head of an NP and remove its specifier and modifier. If the head noun is a domain specific term, then it is further treated by the simplification rule to a terminological unit.

FIGURE 3 shows a parsing tree of a string as an example. The string is labelled as NP at the root node which contains two sub-trees with categories NP and PP, respectively. According to this simplification rule, we therefore remove its PP modifier. As a result, the string "a patient of skin rash with multiple erythematous papules" is simplified to only its head "a patient".

(b) VP (Verb Phrase)

VP → V + NP: We keep V untouched and simplify NP according to the simplification rule (a). For example, we simplify "had underlying diseases of ventricular tachycardia and dyslipidemia" to "had diseases".

VP → V + PP: We keep V untouched and remove PP if PP is a modifier. If PP is mandatory, it is further simplified based on the simplification rule (c). Whether PP is optional or mandatory is determined by the subcategory of V. For example, we simplify the sentence "he was discharged on the morning of 6/30" to "he was discharged".

(c) PP (Prepositional Phrase)

PP → P + NP: We keep P and simplify NP according to the simplification rule (a). For example, we simplify "with underlying diseases of ventricular tachycardia and dyslipidemia" to "with diseases".

(d) S (Clause)

We simplify a clause by simplifying its children recursively according to the above simplification rules.

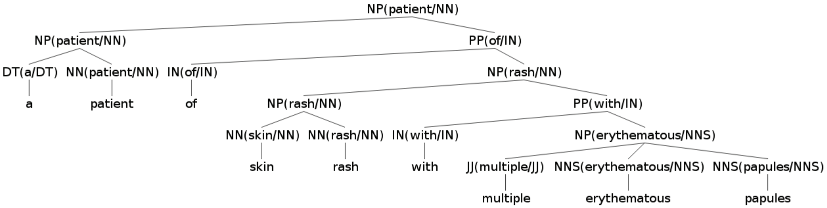


FIGURE 3 – A parse tree of a syntactic unit. A bracketed string at each non-terminal node is the head of the corresponding syntactic category.

The rule-based simplification approach is straightforward, but can be effectively applied to most of the syntactic units as discussed in Section 4. Applying transformation or rewriting rules on source sentences based on their syntactic structures has been adopted in other works. Wang et al. (2007) listed a set of prominent syntactic reordering rules that systematically describe the word order difference between the source and target languages. Based on these rules, they parsed a source language input and reordered its structure to match the target language grammar for training a better translation model and improving a phrase-based SMT system. In their work, source side syntactic reordering also serves as a pre-processing module of the SMT system. Different from their work, we simplify a source language input in favour of the background SMT system instead of changing the order of its structure.

3.3 Translation

In the translation step, the background SMT system translates the simplified in-domain text and produces its translation result. Since the input source text is simplified in favour of the translation system, the contextual distributions of the phrases can be estimated better than those of the original text, as demonstrated in FIGURE 2(b).

Our *STR* framework performs domain adaptation under the scenarios where bilingual in-domain segments are available. It is possible to be combined with other domain adaptation approaches that exploit monolingual or bilingual in-domain corpus to help further improve the translation quality. For example, if a parallel in-domain corpus is available, we can perform learning-based domain adaptation approaches described in Section 2, and tune the background translation model toward the specific domain. In this way, we may receive better translation results from the background SMT system and facilitate the next restoration step.

For a phrase-based SMT system and its variations (Chiang, 2005; Xiong et al., 2006; Huang and Chiang, 2007), we can further customize the decoder to produce the translation with higher quality and facilitate the next restoration step of our framework. For example, the in-domain segments in a text are either terminological or syntactic units in our experiments, and therefore their translations are continuous without other interleaving translations. However, an out-domain SMT system may give wrong ordering without in-domain knowledge. For instance, the translation of the medical term "bone lesion" is separated as illustrated in FIGURE 4. In our work, we set up a Moses (Koehn et al., 2007) SMT system and apply its advanced feature of specifying reordering constraint to each of the simplified phrases. Under the constraint, a simplified phrase is translated as a block and its translation is continuous on the target side.

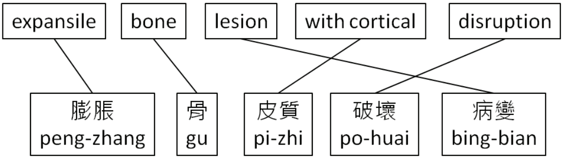
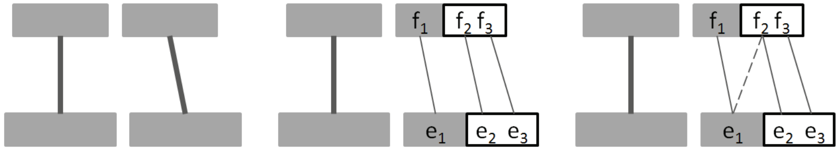


FIGURE 4 – The incorrect reordering of the in-domain segments "bone lesion".

3.4 Restoration

In the restoration step, we receive the translation result from the background SMT system and perform post-processing steps. We locate the simplified phrase pairs and replace them with their corresponding bilingual in-domain segments as illustrated in FIGURE 2(c). The resulting parallel text is the final output of our framework and its target side is the translation of the in-domain text.

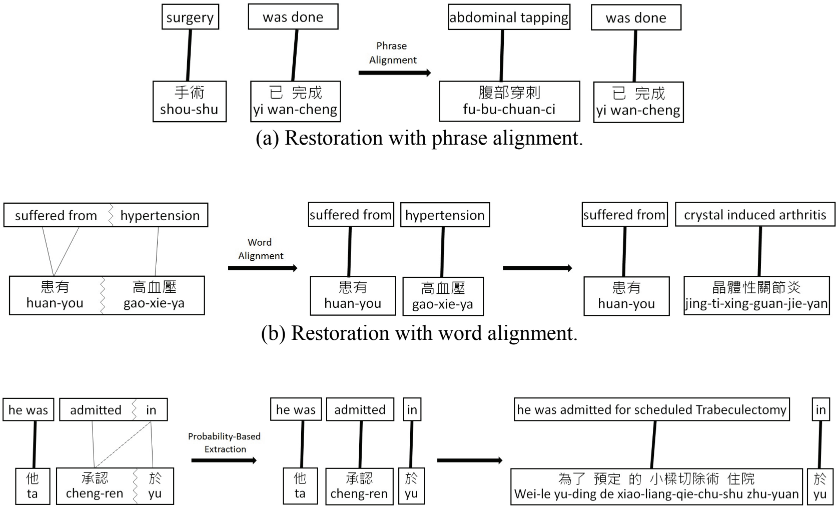
To successfully restore the bilingual in-domain segments, we need the internal alignment information between the source and the target sides. Depending on the difficulty of extracting the simplified phrase pairs, different levels of granularity including phrase alignment, word alignment and word alignment score are needed. The restoration methods under different situations are summarized in FIGURE 5. We apply these restoration approaches one by one in the order of increasing granularity: **phrase alignment**, **word alignment** and **probability-based extraction**. As will be discussed in the later section, the empirical results show that these approaches are simple but can be effective to deal with most of the cases.



(a) Phrase Alignment (b) Word Alignment (c) Probability-based Extraction

FIGURE 5 – Three restoration methods: phrase alignment, word alignment and probability-based extraction. The thick lines are phrase alignments and the thin lines are word alignments.

If a simplified source phrase is translated alone without its nearby context, phrase-level alignment information is sufficient to perform the restoration: we replace the simplified phrase pairs with the original bilingual in-domain segments. FIGURE 5(a) illustrates the phrase alignment method. The shaded blocks and thick lines denote simplified phrase pairs. By checking the phrase alignments provided by the decoder, the simplified phrase pairs can be replaced without further processing. The method is exemplified in FIGURE 6(a). In this example, the simplified term "surgery" is translated as a single phrase. Based on the phrase alignments, the simplified phrase pair "surgery-手術" is replaced with its original form "abdominal tapping-腹部穿刺".



(a) Restoration with phrase alignment.

(b) Restoration with word alignment.

(c) Restoration with probability-based extraction.

FIGURE 6 – Three restoration methods applied with different levels of alignment information. The thick lines are phrase alignments and the thin lines are word alignments.

In some cases, a simplified phrase is translated together with its nearby context, and therefore we need to determine the translation of the simplified phrase before we can restore its original form, such as the term "hypertension" in FIGURE 2(b). Phrase level alignments are insufficient now and higher granularity is needed, such as word-level alignments, to separate a simplified phrase from its contexts. With word alignment information, we may extract a simplified phrase pair without violating the consistency judgment of a phrase pair (Och et al., 1999). We then replace the simplified phrase pair with its original bilingual in-domain segment. Our word alignment method is illustrated in FIGURE 5(b). The hollow blocks denote the context f_2f_3, which is translated together with its nearby simplified term f_1. The thin lines and the thick lines are word alignments and phrase alignments, respectively. Compliant with the consistency of phrase extraction, we separate the phrase pair f_1-e_1 from the phrase pair $f_1f_2f_3$-$e_1e_2e_3$, and perform the restoration. The method is exemplified in FIGURE 6(b). Based on the word alignments, the simplified term "hypertension" can be separated from its context "suffered from" without violating the consistency. Therefore, we can successfully restore the original form "crystal induced arthritis".

There are still cases in which word alignment method fails due to the fertility feature between source and target languages under IBM models (Brown et al., 1993). For a simplified term, which is usually a content word, its translation may be aligned to non-content words on the source side. In this case, extracting simplified phrase pairs would violate the consistency of phrase extraction. Here we apply a probability-based extraction of simplified phrases. This approach deletes weak word alignments based on alignment probabilities. For a source simplified phrase $f_{i,j}$ spanning from word f_i to f_j which is aligned to its target translation $e_{i,j}$, we examine each word in $e_{i,j}$. If there exists an e_k (i<=k<=j) which is aligned to two source words within $f_{i,j}$ and outside $f_{i,j}$ respectively, we try to delete one of the alignments by comparing their word alignment probabilities. FIGURE 5(c) illustrates the probability-based extraction method. The word e_1 is aligned to both f_1 and f_2. The fuzzy alignment causes the unsuccessful separation of f_1-e_1 from its context f_2f_3-e_2e_3. If the word alignment probabilities show $P(f_1|e_1) > P(f_2|e_1)$, we can delete the weak word alignment f_2-e_1, and meet the consistency judgment. The method is exemplified in FIGURE 6(c) where the word alignment method is unsuccessful. In this case, we fail to determine the translation of the simplified syntactic unit "he was admitted" because the target word "承認" is aligned to both "admitted" and "in". With the probability-based extraction approach, we delete the alignment (the dashed line) "in-承認", because P(in|承認) < P(admitted | 承認). After that, we are able to extract the phrase pair "admitted-承認". By determining the translation of "he was admitted" with two consecutive phrase pairs, the restoration of the original form "he was admitted for scheduled Trabeculectomy" can be done easily.

4 Experiments

We experiment our *STR* framework on an English-Chinese SMT application in the medical domain. Moses is built as our general domain SMT system. The translation model is trained on 6.8M sentence pairs collected from Hong Kong corpus (LDC2004T08) and UN corpus (LDC2004E12). We train the trigram language model on the Chinese part of the above parallel corpus and the Central News Agency part of the Tagged Chinese Gigaword (LDC2007T03).

We test our method with the English-Chinese translation task on the medical summaries from the National Taiwan University Hospital (NTUH). The dataset comprises 1,077 parallel sentences within 18 medical summaries. In our SMT application, the gap between in-domain and out-domain corpora is very large in terms of vocabulary and writing styles. In our test set, the average length of a sentence in a medical summary is short (12.58 words) compared to the background general corpus (29 words). TABLE 1 shows some interesting statistics of the domain specific segments in the test set, including the average number of terminological units, syntactic units and OOV words per sentence, and the percentages of words they account for in the test data.

On the average, the in-domain segments (terminological and syntactic units) constitute over 36% of the experimental dataset. Nearly 21% of OOV words including surgical, diagnosis and drug terms occur in a sentence and most of them are parts of terminological and syntactic units. This justifies our motivation to alleviate the OOV problem in our cross-domain SMT system by simplifying the text in this special domain before sending it to the background SMT system.

	Occurrence	Percentage
Terminological Unit	2.04/sentence	18.37%
Syntactic Unit	0.65/sentence	17.86%
OOV Word	2.93/sentence	21.18%

TABLE 1 – Statistics of the in-domain segments and the OOV words in the test set.

To identify the in-domain segments in a text, we collect bilingual terminological and syntactic units from the medical domain. Section 3.1 describes the identification step and discusses various ways to collect the bilingual in-domain knowledge with automatic and semi-automatic approaches. In our experimental domain (i.e., English medical summaries), bilingual medical dictionaries are available from plenty of resources and thus they are sufficient for collecting bilingual terminological units. On the other hand, bilingual syntactic units are relatively hard to obtain. Only monolingual corpus can be obtained in this domain, and obtaining parallel text by manual translation incurs high cost due to the involvements of domain experts (e.g., doctors). Therefore we collect the bilingual syntactic units by taking a semi-automatic approach in order to make the best of manual efforts. The collection of terminological and syntactic units is briefly described below, and readers can refer to Chen et al. (2011) for more details.

The terminological units are collected from both public and non-public resources. They include public medical dictionaries from domestic medical colleges and Department of Health. We are also provided with frequent bilingual term pairs used by NTUH staff. Merging these bilingual dictionaries causes ambiguity in which a medical term has multiple translations with similar meanings but in different styles. Since consistent translation of terminology is desired in this application, ambiguous translations are reviewed and edited by the staff at NTUH. So far nearly 70,000 bilingual terminological units are collected and stored in our database.

The syntactic units are semi-automatically collected with a pattern mining algorithm and annotations by the in-domain experts. Since we choose phrase-based SMT as our background SMT system, the n-gram based syntactic units are preferred for easier integration. We first automatically extract the candidate n-grams from the experimental medical summary corpus with medical entity recognition (Ben Abacha and Zweigenbaum, 2009) and NLP techniques. These source language n-grams are then reviewed and fast translated by doctors with the help of a user-friendly annotation tool. The resulting bilingual n-grams are then collected and organized into the bilingual syntactic units. The NLP techniques aim to reduce the cost of annotations by doctors, and increase the coverage of bilingual syntactic units. TABLE 2 gives some samples of these n-gram based syntactic units in this domain. Both source and target n-grams are listed for each bilingual syntactic unit. The words in bold denote the medical categories which represent diseases or symptoms (**DIAGNOSIS**), medical tests (**TEST**), surgical or non-surgical treatments (**TREATMENT**), etc. These syntactic units capture in-domain writing styles and local reorderings (note the different orders of medical categories between source and target sides).

Source Syntactic Unit	Syntactic Label
Target Syntactic Unit	
underwent **TEST** on **DATE**	VP
於 **DATE** 接受 **TEST** 檢查 yu jie-shou jian-cha	
DRUG was given for **DIAGNOSIS**	S
使用 **DRUG** 用於治療 **DIAGNOSIS** shi-yong yong-yu-zhi-liao	
received **TREATMENT** with **DRUG**	VP
接受 **TREATMENT** 及 **DRUG** 治療 jie-shou ji zhi-liao	
DIAGNOSIS at the right **REGION**	NP
在右側 **REGION** 之 **DIAGNOSIS** zai-you-ce zhi	
TREATMENT of the right **REGION**	NP
右側 **REGION** 的 **TREATMENT** you-ce de	

TABLE 2 – Samples of bilingual syntactic units.

In the simplification step, we simplify the identified in-domain segments of a text. For runtime efficiency, we perform the simplification on all of the collected terminological and syntactic units in advance and store the results to avoid redundant work. The terminological units are simplified based on the ontology provided by Unified Medical Language System (UMLS). For a medical term, we search for its hypernyms and find the most frequent one in the background translation model and designate it as the simplified form.

The syntactic units are parsed with Stanford parser (Klein and Manning, 2003) and simplified with the rules described in Section 3.2. TABLE 3 gives the number of syntactic units for each syntactic label. There are parsing errors for few n-grams and they are manually corrected. As shown in TABLE 3, more than 70% of the n-grams are correctly simplified. Most of the errors come from the unexpected syntactic labels which can be processed with new simplification rules. For example, the common syntactic unit "was admitted due to **DIAGNOSIS**" can be simplified to "was admitted". However, the syntactic label ADJP which covers "due to **DIAGNOSIS**" is not considered in our rules and therefore remains unchanged after the simplification step. We plan to devise a more robust method beyond the rule-based one for future work.

In the restoration step, we receive and post-process the SMT results as described in Section 3.4. In our experiments, phrase level alignments are provided by the Moses decoder, and word level alignments are obtained as the intermediate results after training the phrase table. Our methods successfully perform the restorations on most of the test data. Total 1,004 (93.22%) of the 1,077 sentences can be successfully restored with the three proposed restoration methods. For the other 73 sentences, most of these failure cases result from translation errors of the simplified phrases by Moses, which in turn affect the word alignments and introduce difficulties. TABLE 4 shows the performance of each restoration method. The second column counts the number of applications of each restoration method on the test set, and the third column counts how many

Syntactic Unit	Count	Parsing Error	Simplification Error
NP	697	4.85%	28.15%
VP	287	2.94%	27.12%
PP	228	1.85%	8.89%
S	342	1.23%	9.59%
Other	12	33.33%	

TABLE 3 – Parsing and simplification performances on syntactic units. We manually simplify the 12 syntactic units with minority labels.

	Total Count	Sentence Count
Phrase Alignment	1,767 (60.93%)	865 (86.16%)
Word Alignment	981 (33.83%)	657 (65.44%)
Probability-based Extraction	152 (5.24%)	137 (13.65%)

TABLE 4 – Various restoration methods.

sentences include the application of each restoration method. Total 2,900 applications of the restoration methods have been done in the testing. With phrase alignments, we can deal with over 60% of the simplified phrases. For the remaining simplified phrases, the background SMT model captures their contexts with higher confidence, and therefore they are translated with the surrounding words. For these cases, we use word alignment information, and perform the word alignment and the probability-based extraction methods. Although the probability-based extraction accounts for only 5.24% of the restorations, it is applied on 13.65% of the sentences, as shown in the third column. This confirms that the approach of deleting weak word alignment is simple but effective in handling the inconsistency during phrase extraction.

We compare our *STR* framework against three baselines. These four SMT systems share the same background translation and language models which are trained from the corpus described in the beginning of this section. The Moses system is trained with its default scripts without any in-domain prior knowledge. We choose Moses' default behaviour to handle OOV words, i.e., they are copied to translation results. We also use an advanced feature of Moses to build another setting (Moses+Terminology) by adding an in-domain phrase table with the terminological units. It serves as the back-off model during the runtime decoding. That is, Moses searches the back-off model only when no translation for a phrase is found in the background model. In the naïve integration system, we apply the identification step to a source sentence and mark the in-domain segments. Rather than simplifying these segments, we use the XML markup function of Moses to force the translations of these terminological and syntactic units.

TABLE 5 shows 4-gram BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) scores of these systems. With poor in-domain knowledge, the baseline Moses achieves the low performance on the test set. On the other hand, the Moses system with terminology back-off model eliminates most of the OOV problems and gets better performance. The naïve integration system gives further improvement over the Moses+Terminology system by performing medical entity recognition on a source input before sending it to Moses. Therefore, it can identify the multiword terminological units and the syntactic units during the identification step. The system with the *STR* framework achieves the best translation performance among the four systems measured by both BLEU and TER scores. Comparing the translation results of the *STR* framework and the naïve integration system, we find the former gives better translations and reorderings on the general segments. In summary, the experimental results and observations show that our simplification approach improves the baseline SMT systems by identifying specific segments with bilingual in-domain knowledge and modifying these special segments to better fit the background SMT model.

	BLEU-4	TER
Moses	14.35	64.911
Moses+Terminology	20.56	55.712
Naïve Integration	26.29	47.683
STR Framework	35.87	40.650

TABLE 5 – Translation performances on medical summaries.

Conclusion and Future Work

This paper proposes a simplification-translation-restoration framework for cross-domain applications in SMT. We integrate bilingual in-domain knowledge into a background out-domain SMT system. That deals with the cross-domain and the data sparseness problems at the same time. The in-domain text goes through identification, simplification, translation, and restoration steps. Important issues are addressed and discussed for each step, including preparing bilingual in-domain knowledge, simplification with syntactic information, and different restoration strategies for extracting simplified phrases from the SMT results. We evaluate the performance of our framework through a cast study of medical summary translation. The empirical results show the effectiveness of our approach at each step of the framework. For the end-to-end translation task, our method outperforms the background SMT system and the systems with different integration approaches.

Currently, we apply a rule-based simplification approach to four common syntactic units, i.e., NP, VP, PP, and S. The alternative is to simplify in-domain segments based on statistical evidence. Searching for a simplified form for a phrase in a more robust way will give more hints to aid the background SMT system. That is worthy of further investigation. Besides, to introduce a feedback mechanism to improve an *STR*-based MT system is also one of the research issues. Which parts, pre-processing (i.e., identification and simplification), translation, and post-processing (i.e., restoration), have to be modified through the feedback and how they are modified are critical. In our scenario, the doctors post-edit the MT results of English medical summaries and produce the correct Chinese medical summaries. The editing logs can be analysed to improve identification, simplification, translation, and restoration steps for further domain adaptation.

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 101R890858.

References

Abacha, A. B. and Zweigenbaum, P. (2011). Medical entity recognition: a comparison of semantic and statistical methods. In *Proceedings of the 2011 Workshop on Biomedical Natural Language Processing*, pages 56–64.

Aziz, W., Dymetman, M., Mirkin, S., Specia, L., Cancedda, N., and Dagan, I. (2010). Learning an expert from human annotations in statistical machine translation: the case of out-of-vocabulary words. In *Proceedings of EAMT 2010*.

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

Callison-Burch, C., Koehn, P. and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings NAACL 2006*, pages 17–24.

Carpuat, M. and Wu, D. (2007). Improving statistical machine translation using word sense disambiguation. In *Proceedings of EMNLP 2007*, pages 61–72.

Chan, Y. S., Ng, H. T. and Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. In *Proceedings of ACL 2007*, pages 33–40.

Chen, H., Huang H., Tjiu, J., Tan, C. and Chen, H. (2011). Identification and translation of significant patterns for cross-domain SMT applications. In *Proceedings of Machine Translation Summit XIII*, pages 277–284.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, pages 263–270.

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

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

Haghighi, A., Liang, P., Berg-Kirkpatrick, T. and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL 2008*, pages 771–779.

Huang, L. and Chiang, D. (2007). Forest rescoring: Faster decoding with integrated language models. In *Proceedings of ACL 2007*, pages 144–151.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of ACL 2003*, pages 423–430.

Koehn, P., Och, F. J. and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 127–133.

Koehn, P. (2004). Pharaoh: a beam search decoder for phrased-based statistical machine translation models. In *Proceedings of AMTA 2004*, pages 115–124.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constrantin, A., and Herbst, E. (2007). Moses:

Open source toolkit for statistical machine translation. In *Proceedings of ACL 2007*, Demonstration Session, pages 177–180.

Marton, Y., Callison-Burch, C. and Resnik, P. (2009). Improved statistical machine translation Using monolingually-derived paraphrases. In *Proceedings of EMNLP 2009*, pages 381–390.

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

Morin, E., Daille, B., Takeuchi, K. and Kageura, K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of ACL 2007*, pages 664–671.

Och, F. J., Tillmann, C. and Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proceedings of EMNLP 1999*, pages 20–28.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A. and Bederson, B. (2010). Improving translation via targeted paraphrasing. In *Proceedings of EMNLP 2010*, pages 127–137.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA 2006*, pages 223–231.

Wang, C., Collins, M. and Koehn, P. (2007). Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP 2007*, pages 737–745.

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

Wu, J. and Chang, J. S. (2007). Learning to find English to Chinese transliterations on the web. In *Proceedings of EMNLP-CoNLL 2007*, pages 996–1004.

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

Xiong, D., Liu, Q. and Lin, S. (2006). Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of ACL-COLING 2006*, pages 521–528.

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

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