A Statistical Medical Summary Translation System

Han-Bin Chen[1], Hen-Hsen Huang[1], Ching-Ting Tan[2], Jengwei Tjiu[2], and Hsin-Hsi Chen[1]
[1] Department of Computer Science and Information Engineering
National Taiwan University
[2] National Taiwan University Hospital
Taipei, Taiwan
{hbchen, hhhuang}@nlg.csie.ntu.edu.tw; {tanct5222, p99748036, hhchen}@ntu.edu.tw

ABSTRACT
In a hospital, a medical summary is indispensable for both a clinician and a patient. However, it is written in English in some non-English native countries and becomes a barrier for a patient to read. In this paper we propose a framework for rapid acquisition of bilingual medical summaries using machine translation (MT) techniques. We describe a medical summary corpus and some terminological databases prepared for the framework. We then touch on the challenging issues of MT adapted from generic to specific domains, and propose a pattern translation scheme to achieve domain adaptation based on a background statistical MT system. We identify the significant patterns to capture the specific writing styles in a medical summary. The patterns are then translated with the involvements of doctors. Our major concern is to reduce the cost of translation and better allocate the efforts made by the domain experts. The experimental results show the proposed methods are effective in terms of the significance and diversity of the patterns. The approaches to integrate the mined patterns into background MT are also discussed.

Categories and Subject Descriptors
J.3 [**Life and Medical Sciences**]: Medical information systems; I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Machine translation*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Clustering*

General Terms
Algorithms, Design, Experimentation

Keywords
Medical Summary, Machine Translation, Pattern Identification

1. INTRODUCTION
A health record such as a medical summary is a document which keeps track of a patient's history, present illness, treatments, etc. in a hospital. On the one hand, a medical summary helps doctors quickly understand the overall status of an incoming patient. On the other hand, medical summaries are personal documents and each individual patient has the right to acquire his own information. In other words, patients have rights to know the treatments during their stay in a hospital by reading these documents. To this end, it is common to carry out electronic medical summaries for efficient retrieval and management in hospitals. However, a medical summary is always written in English in some countries (e.g., Taiwan) where the official languages are not English. That becomes barriers for patents to know what treatments have been done and infringes their right to know. Therefore, it is important to make such personal medical information available in their native language.

An intuitive solution is to convert the existing medical summaries into the target language counterpart with the involvements of human translators. However, translating these documents by medical personnel at hospitals is impractical. In our experimental corpus of medical summaries, the average length of a patient's history is 30 sentences. Translating millions of medical summaries over the past decades with human efforts therefore involves tremendous costs in terms of both time and money. Under the circumstances, a machine translation (MT) system, which automatically translates documents in one language into another, may play important roles in medical summary translation.

In recent years statistical machine translation (SMT), which builds MT model with a large corpus, becomes the mainstream in MT researches due to the rapidly growing computing speed and storage size. Many SMT models such as phrase-based model, syntax-based model, and example-based model have been proposed. In addition, some typical search engine portals also provide translation services such as Google Translate and Yahoo Babelfish. However, these MT models or MT services cannot resolve medical summary translation directly because of the specific medical domain. Consider an example. Sentences like "Spine MRI on 2009/2/8 showed compression fracture" are frequently used in medical summaries to describe the observations after performing a diagnostic procedure. Google Translate reports its Chinese translation as "對 2009 年 2 月 8 日，MRI 表現爲脊柱壓縮性骨折". One of the proper translations would be "於 2009 年 2 月 8 日進行的腰椎核磁共振顯示爲壓縮性骨折". In this example, the online translator fails to recognize "spine MRI" as a procedure and mistranslates the verb and preposition. Furthermore, it gives the incorrect word order in the target language translation.

Such cross-domain problems emerge when statistical modeling is applied. The language usages in different domains vary significantly. The differences come from different linguistic aspects such as lexical choice, writing style, and so on. These varieties affect the term distribution in corpora used for training and thus greatly change the statistical model for a specific domain. A straightforward way to deal with the domain-specific problem

IHI'12, January 28–30, 2012, Miami, Florida, USA.

is to train the model using in-domain data. However, such a domain-dependent corpus is not always available. This problem is much more serious in cross-language cross-domain applications, particularly for those applications in highly specific domains such as biochemistry and medical science.

In the past, the parallel corpus used to train an MT system mainly comes from fixed domains such as parliamentary and news articles. The bilingual resources for a specific language pair or a specific domain usually come in small size, even unavailable. One of the challenging issues in a cross-domain MT application is to realize an in-domain MT model in a resource-poor environment where a bilingual training corpus is not applicable.

To tackle this problem, we propose a framework to build an English-to-Chinese medical summary SMT system in this paper. In an SMT model, an English-Chinese parallel corpus is indispensable for training the translation model (TM). However, only an English medical summary corpus is available in this domain. Here, we first develop a general English-Chinese SMT system with Moses toolkit [11] and a general English-Chinese parallel corpus. Next, we mine the common patterns in a medical summary to capture its writing styles. The problem is that the learned patterns are still monolingual. It is necessary to involve domain experts in setting up bilingual patterns. With the domain specific and cross-language knowledge, we adapt our general purpose SMT system to serve as the medical summary translator.

There are some core issues addressed in our framework. Medical summaries written by doctors are raw texts in inconsistent formats, which are inconvenient to process and analyze by programs. Furthermore, documents in such a specific domain contain plenty of proper nouns and named entities which are hard to identify without in-domain data resources. Therefore, it is necessary to pre-process the format of medical summary data and build the terminological databases in medical domain as the initial stage. In addition, the cost of domain experts involved in translating the patterns is another major concern. For this reason, a series of research issues are touched in this paper: (1) to prepare for the in-domain data resources, (2) to identify significant patterns from an English medical summary corpus, (3) to find their coverage relations and decide which patterns should be translated by experts, and (4) to introduce these patterns into the out-domain general MT system.

Depending on the type of in-domain resources at hand, various domain adaptation techniques in MT have been proposed. Foster and Kuhn [6] proposed a mixture-model approach to deal with the case where bilingual in-domain text is available but in a relatively small size. A training corpus was divided into several components to train several models. Each model was weighted to estimate the similarity between components and in-domain development data. Based on this work, Foster et al. [7] incorporated instance weighting that learned the weights of bilingual phrase pairs to capture the degree of relevance to the target domain. Similarly, a mixture-model approach was also applied in word-alignment task [4]. In their work, domain related parameters were added in the standard HMM training to derive an alignment model sensitive to the topic for each sentence.

In some applications, a bilingual in-domain corpus is simply unavailable while the in-domain monolingual text (either source or target side) is relatively easy to acquire. Zhao et al. [16] combined the baseline language model (LM) and the in-domain LM which was trained by retrieving documents from large text collections using query models. Besides LM, Bertoldi and Federico [3] exploited a monolingual corpus to train TM. They generated a synthetic parallel corpus from a monolingual one and used it for domain-specific training.

Our work is close to the monolingual scenario. Provided with a monolingual domain-specific corpus, we adapt our background MT into the one that is suitable for translating medical summaries. There are some major differences among our work and those proposed previously. First, the related works exploited the entire in-domain training data to adapt the existing LM or TM by model mixture and parameter tuning. Instead, we identify and translate significant patterns from large-scale in-domain source texts and introduce them to our SMT system. Second, the significant patterns are translated with the involvement of expert knowledge to deal with the large domain difference between background training corpus and medical summaries. To reduce the labor cost of experts, filtering, clustering, and ranking the patterns are the major issues.

The rest of this paper is organized as follows. Section 2 proposes the overall framework and briefly goes through each component. Section 3 gives details of the in-domain data, including a medical summary corpus and medical terminological resources. How these data are refined and organized for further use is also described. Sections 4, 5 and 6 describe pattern identification, translation and further integration in detail, respectively. Section 7 evaluates the performance of these algorithms. Finally, Section 8 concludes the remarks.

2. Framework

Translating articles in a specific domain using a general domain MT is challenging. In this section, we introduce the motivation behind the pattern translation scheme and propose the overall framework for the medical summary SMT system.

In hospitals, medical summaries are in special written styles and are usually short. In this study, for example, our experimental dataset is a collection of the English medical summaries from National Taiwan University Hospital (NTUH). The average length of a sentence is 10 words. In contrast, the background general domain corpus is Hong Kong Parallel Text purchased from the Linguistic Data Consortium (LDC). The average sentence length is 29 words on the English side.

A number of patterns frequently repeat in medical summaries. For example, the sentence "Port-A implantation was performed on 2009/10/9" contains a frequent medical pattern with many instances:

paracentesis was performed on **2010-01-08**
repositioning was performed on **2008/04/03**
incision and drainage was performed on **2010-01-15**
tracheostomy was performed on **2010/1/11**

The pattern states a kind of surgery was performed on some particular date, shown as follows:

SURGERY was performed on **DATE** (1)

Here, **SURGERY** represents a class of medical terms denoting surgeries, and **DATE** is a class of date expressions.

The general SMT systems are unable to properly recognize common patterns like pattern (1) shown above, and tend to produce improper translations of texts which contain these patterns. One of the major reasons is that highly specific terms in

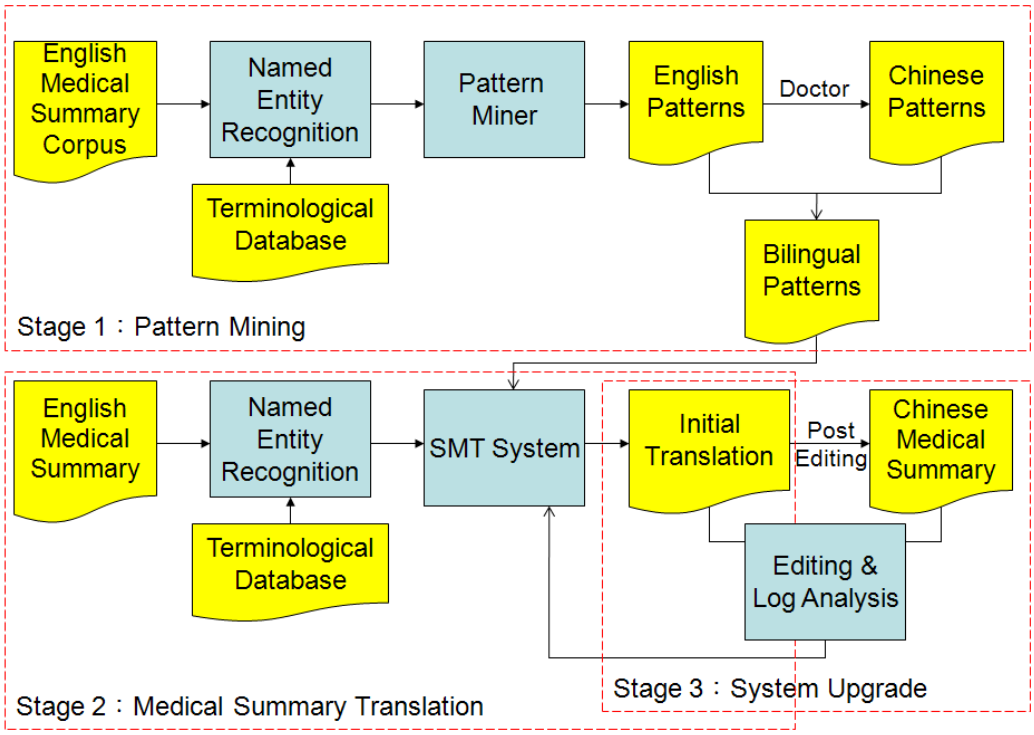

Figure 1. An overall framework for medical summary translation.

Table 1. Examples of bilingual patterns.

English Patterns	Chinese Patterns
SURGERY was performed on **DATE**	在 **DATE** 施行 **SURGERY**
was admitted for scheduled **SURGERY**	住院預定接受 **SURGERY**
received **SURGERY** on **DATE**	在 **DATE** 接受 **SURGERY**
agreed to go on **SURGERY**	同意進行 **SURGERY**

the target domain are rare or even unseen in the bilingual training corpus. Another reason is that the specific writing styles of the target domain are unknown. As a result, instances of a pattern in a medical summary are translated by the generic SMT systems with poor qualities and inconsistent styles.

By identifying and translating these common patterns, we obtain bilingual patterns which are applied during the translation of an input medical summary. Table 1 gives the examples of translated patterns, including pattern (1).

In this paper, we present a machine translation framework shown in Figure 1 to deal with the translation issues in the medical domain. These components include the medical data resource preparation, the technical term identification, the significant pattern extraction and translation, the integration of bilingual patterns into a general SMT system, and the log analysis together with feedback mechanism of the post-edited medical summaries. There are 3 major stages in the framework for building the medical summary translation system.

At Stage 1, setting up a set of bilingual patterns is the goal. We are provided with raw documents of medical summaries and terminological resources. These in-domain data is pre-processed and organized in an accessible format for the later stages. With the derived terminological databases, named entities in medical summaries are identified and labeled with medical classes. The pattern miner then estimates and extracts the English significant patterns from medical summaries. These patterns are translated with the involvements of domain experts (i.e., doctors) and a set of bilingual patterns are produced.

At Stage 2, we adapt the bilingual patterns to the background SMT system. During the runtime translation, we apply this domain specific translator for each input medical summary, and output the translation result. Since medical summaries are health records and of great importance for patients, further review and modification of MT results by doctors is necessary.

At Stage 3, an interface is designed for doctors to post-edit the translation results produced by the medical summary translator. The modified translations serve as the Chinese medical summaries to help non-English speakers, which is our primary purpose. On the other side, post-editing and log analyses are beneficial for tuning and optimizing the system by machine learning techniques.

The details of Stage 1 will be discussed in Sections 3-5, which are data description, pattern identification and pattern translation, respectively. Section 6 describes the ideas for Stages 2 and 3.

3. Data Description

3.1 A Medical Summary Corpus

To translate medical summaries, an in-domain training corpus is necessary. We obtain the corpus of NTUH medical summaries written from January to June, 2010. It is composed of 60,448 medical summaries with 1.8M sentences and 18M words.

There are 3 main parts in an NTUH medical summary: chief complaint, brief history, as well as course and treatment. Chief

Table 2. A sample of medical summary.

Chief Complaint Hematemesis on 1/22.
Brief History This 70-year-old male **has history of** 1.Hypertension, 2.Diabetes mellitus, 3.Chronic kidney disease. He **regularly followed up** at our OPD. His long term prescription included Tapal 1 tab QD. ccording to his wife's statement, he felt nausea since 1/22 morning, and then he **had sudden onset of** syncope. A lot of blood was vomited on 1/23 accompanied with general convulsion. Then he was sent to our emergency department by Ambulance. In our emergency department, **laboratory data revealed** low level hemoglobin (7.5 mg/dl) and normal range of cardiac enzyme and normal range of PT 、PTT. Blood transfusion and Pantoloc **were given for suspected** upper gastrointestinal bleeding. Endoscopy **was performed on** 1/23 with results of a 0.4cm A1 duodenal ulcer and a gastric erosion. Bosmin injection and heater probe were done during endoscopy procedure. **Under the impression of** duodenal ulcer with bleeding, he was admitted to our ward for further management.
Course and Treatment After admission, we continued intravenous fluid supply and PPI. We switched PPI to oral form since 2010/01/26. Following hemoglobin was stationary (1/27 Hb:10.3). He was discharged on 2010/01/28 with further GI OPD follow up.

complaint includes the symptoms for which a patient seeks medical care, and it is written in patient's own words. Brief history states the present illness and past medical history of a patient. Course and treatment describes the disease status of a patient and the progress of treatments.

A typical medical summary is sampled and shown in Table 2. As illustrated in the table, for chief complaint as well as course and treatment, they are briefly documented and short in length. In contrast, brief history records a patient's past history and how he/she suffers from the present illness in detail. It constitutes the major part of a medical summary and contains many specific writing styles. Table 2 highlights the examples of frequent patterns with the strings underlined in bold.

Based on this argument, we perform the experiments on brief histories, from which we identify and translate significant patterns. Note in Table 2 that with punctuations, these raw documents are inconvenient to gather statistics and analyze. Therefore, for each medical summary, we perform tokenization of words and detect sentence boundaries using the Natural Language Toolkit (NLTK) [12]. Sentences are then processed as the basic units of a medical summary by the pattern miner and SMT system in Stages 1 and 2.

3.2 Terminological Databases

We obtain terminological resources in medical domain and build the monolingual and bilingual databases. Monolingual databases are employed for identifying medical named entities, while bilingual ones serve as dictionaries that translate English medical terms into Chinese counterparts.

Table 3. Four basic medical classes in NTUH databases.

Class Names	Descriptions
SURGERY	Treatments including surgical and non-surgical
DIAGNOSIS	Diseases and symptoms
TEST	Laboratory and diagnosis procedures
DRUG	Drug names and pharmaceutical substances

Table 4. Mappings from Semantic Types to medical classes.

Examples	Semantic Types	Medical Classes
wide excision	Therapeutic or Preventive Procedure	**SURGERY**
hypertension	Disease or Syndrome	**DIAGNOSIS**
Lanoxin	Pharmacologic Substance	**DRUG**
GnRH test	Laboratory Procedure	**TEST**

3.2.1 Monolingual Terminology

Recall in Section 2 that a pattern in medical summary contains medical classes such as SURGERY. Thus, given a medical summary, identifying domain specific terms and classifying them into suitable classes is the first step toward the extraction of significant patterns. This procedure is similar to computer-based coding. Medical coding transforms medical terms into their corresponding medical code numbers with classification systems, such as ICD and DSM. However, these classification systems are sophisticated and mainly applied for clinical text analysis, knowledge extraction and expert system. In this work, we are interested in a coarse-grained medical classification, and aims to extract patterns at a more general level. For example, for the translation of the pattern "he was diagnosed with DIAGNOSIS", it makes little difference whether the DIAGNOSIS represents mental or physical diseases. Such simplified classifications are applied in medical entity recognition [1][14] and semantic relation extraction [5].

We are provided with lists of terms frequently used by several hospital departments at NTUH. Terms of a variety of subjects such as diagnosis, surgery, pharmacy, and laboratory medicine are used in these departments. They thus form the basic medical classes in our patterns: DIAGNOSIS, SURGERY, DRUG and TEST. Class names and the sorts of terms these classes include are described in Table 3.

In addition to these non-public terminological resources, we incorporate larger amount of public resources in to further extend our knowledge base for medical term classification. While online medical dictionaries are free to consult, they are mainly built for explaining the meaning of medical terms, without explicit information of general concept that a medical term represents.

Here, we apply the resources from the Unified Medical Language System (UMLS) maintained by National Library of Medicine. The UMLS covers a wide range of terms in medical domain and relations between these medical terms. Among these resources, the Metathesaurus organizes medical terms into groups of concepts. Moreover, each concept is assigned to at least one

Table 5. Number of classified medical terms.

Medical Classes	NTUH	UMLS
SURGERY	7,837	12,240
DIAGNOSIS	17,556	36,734
DRUG	2,781	35,890
TEST	2,673	8,970
BODY	NULL	16,005

Table 6. Ambiguous medical terms.

English terms	Chinese terms
Breda's disease	熱帶莓疹
	莓疹病
	雅司病
confusion	混亂
	精神混亂
	紊亂
colloid	膠狀質
	膠質
	膠體

Semantic Type. Semantic Types provide categorization of concepts at a more general level and are well-suited to be incorporated. Table 4 shows examples of how Semantic Types correspond to NTUH medical classes.

Merging existing ontologies is another research issue. In this paper we propose a mapping from 133 Semantic Types of UMLS to our 4 medical classes. Terms with some unmapped Semantic Types, such as animals and plants, are not classified. On the other side, terms with Semantic Types related to organs and body parts are mapped to an additional medical class BODY. We create this new class because these terms are frequently used to describe patients, and are parts of many patterns.

We process these monolingual terminological resources and build the databases that store and classify each medical term. The statistics of each medical class are presented in Table 5. Notice in the table that by introducing public UMLS, we greatly enhance the non-public terminological resources from NTUH.

3.2.2 Bilingual Terminology

During the translation, we detect patterns in medical summaries and translate them into target language. First we apply our bilingual patterns as the translation rules illustrated in Table 1. We then determine the translation of medical classes in these target language patterns.

Take pattern (1) for an example. First we translate "Port-A implantation was performed on 2009/10/9" into "在 DATE 施行 SURGERY" by applying the first bilingual pattern in Table 1. Next, we translate "2009/10/9" and "Port-A implantation" into "2009 年 10 月 9 日" and "人工血管置放術". Finally, we derive the translation of this pattern, that is "在 2009 年 10 月 9 日施行人工血管置放術".

In the above example we can observe that a bilingual dictionary is needed to translate the source medical terms in to the target ones. To this end, we collect bilingual resources and build a database that is applied during the translation in Stage 2. The NTUH term lists described in the last section provide translations of each term, and thus contribute to the non-public bilingual dictionary. In addition, we collect the public dictionaries, including those provided by Department of Health.

Most of the medical terms have one-to-one correspondences between the English and Chinese. However, merging the bilingual dictionaries causes ambiguous problem, where a term has the same translations but in different styles. Table 6 shows some of these ambiguous terms. To achieve the consistent translations in medical summaries, ambiguous translations are reviewed and edited by the staffs at NTUH. Thus far we have 71,687 pairs of bilingual terms in our database.

4. Pattern Identification

Provided with a large English medical summary corpus and terminological databases, we aim to (1) estimate and extract the significant patterns to capture the domain specific writing style as much as possible; (2) refine and reduce the size of the pattern set to minimize the cost of expert involvements in reviewing and translating the patterns. The overall steps are summarized as follows.

(a) Medical Entity Classification
Recognize medical named entities including surgeries, diseases, drugs, etc., transform them into the corresponding medical classes, and derive a new corpus.

(b) Frequent Pattern Extraction
Employ n-gram models in the new corpus to extract a set of frequent patterns.

(c) Linguistic Pattern Extraction
For each pattern, randomly sample sentences having this pattern, parse these sentences, and keep the pattern if there is at least one parsing sub-tree for it.

(d) Pattern Coverage Finding
Check coverage relations among higher order patterns and lower order patterns, and remove those lower patterns being covered.

(e) Pattern Clustering
Cluster the remaining patterns of the same order, and output the representative patterns from each cluster for pattern translation.

Steps (a)-(c) deal with the first issue, i.e., to extract patterns of high qualities. Steps (d)-(e) touch on the second issue, i.e., to select patterns of high diversities.

4.1 Medical Entity Classification

As discussed in the earlier sections, a pattern may include classes representing the general concept of a group of terms. With the named entity recognition (NER) techniques, as well as the established monolingual terminological databases, we identify and classify named entities into suitable classes.

Named entities such as medical terms, hospital names, date/time expressions, etc. are our targets. Recognition of traditional named entities like organization names and date/time expressions has been discussed intensively before, so that they are neglected in this paper. Here, we focus on the classification of medical terms only. To identify and classify medical terms in our domain specific corpus, we examine each sentence from left to right and adopt a longest-first strategy to replace medical terms with classes. As described in Section 3.2.1, here we apply our monolingual

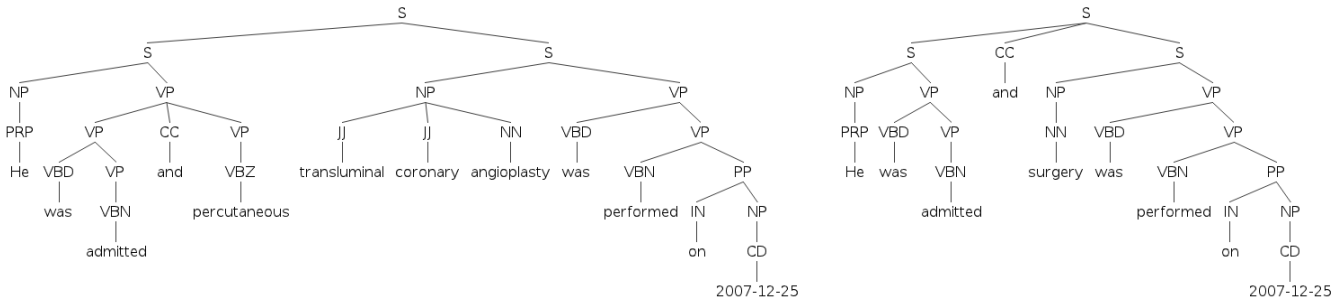

Figure 2. Linguistic completeness of pattern (1).

terminological databases to classify medical terms. In this way, a set of medical summaries are transformed into a new corpus.

In this study, we employ n-gram (i.e., string of consecutive tokens) to represent pattern. Length of n-grams shows some limitation on their usages in natural language processing, including machine translation. On the one hand, lower order n-grams only capture local cues in a restricted scope only. On the other hand, we need more training data to achieve reliable statistics of higher order n-grams. Recognizing word strings of specific semantics and replacing them with classes is useful to resolve the locality issue of n-grams. Thus, patterns are in terms of combinations of words and classes rather than words only. That will enlarge the scopes of patterns in some senses.

4.2 Frequent Pattern Extraction

To address the domain adaptation problem in MT, we extract patterns from an in-domain medical summary corpus to capture domain specific writing styles. These patterns are translated and will be applied in the run-time translation. Accordingly, we prefer the format of patterns that is easy to be integrated into an SMT system for the target application.

The phrase-based model [9][10] is one of the state of the art translation models, in terms of both accuracy and speed. The phrase-based SMT translates source phrases into target ones with phrase table, which consists of bilingual phrase pairs and feature scores estimated from word-to-word alignments. Since a phrase (i.e., a string of consecutive words) is served as the basic unit of translation, integrating n-gram based patterns into the background phrase-based SMT system is a natural choice.

We enumerate all the n-grams from the sentences of our in-domain corpus that contains words and medical classes. In this way, two kinds of patterns are extracted: (1) **class patterns** that contain at least one medical class and (2) **lexical patterns** that contain only words. Note that both patterns are easy to be integrated into a phrase-based SMT, by embedding lexical patterns into the phrase table, and by serving class patterns as translation rules that are applied when a pattern occurs in the medical summary. After all the patterns are extracted from the medical summaries, they are ranked by frequencies.

4.3 Linguistic Pattern Extraction

The main role of doctors in our framework is to translate patterns extracted by our algorithms. This includes reviewing the patterns, neglecting insignificant ones, and translating the patterns considered to be important. However, more than 7M distinct patterns were extracted from medical summaries consisting of 1.8M sentences. It is infeasible to judge the significances through this enormous number of patterns by these domain experts. Therefore, further filtering the patterns to an acceptable size is necessary before the expert involvements.

The linguistic meaningfulness of patterns is proposed to judge their significance. For example, "SURGERY was performed" is a linguistic constituent, while "SURGERY was performed on" is not complete. Accordingly, we filter out patterns that do not meet the requirements of complete linguistic constituents. A parser is adopted to determine the linguistic completeness of patterns.

A cross-domain issue arises when applying a general purpose parser to the domain specific corpus, because the parser built from a general domain training set may suffer from parsing the text with domain specific terms, such as diagnoses and drug names, especially when a term spans across multiple words. Here we take advantage of medical entity classification introduced in Section 4.1. For each named entity in a sentence, we replace it with a common word in favor of our general purpose parser. In this way, we reduce not only the out-of-vocabulary (OOV) words, but also the length of sentences, and thereby facilitate the parsing procedure. For instance, we successfully determine the linguistic completeness of pattern (1) by replacing a complicated surgery name "percutaneous transluminal coronary angioplasty" with the simpler one "surgery", as illustrated in Figure 2. Note in the figure that the original instance of pattern (1) in the left tree fail to pass the requirement of linguistic completeness, while the modified instance "surgery was performed on 2007-12-25" in the right tree is a linguistic constituent.

For each extracted pattern, we select *m* distinct sentences in which it occurs. These sentences are then analyzed by a parser, and the *m* parsing trees are produced. The pattern is considered as a significant candidate, if it is a syntactic constituent in any one of these parsing trees. In this paper we apply Stanford Parser [8] and set *m* to 10 in consideration of the parsing speed.

4.4 Pattern Coverage Finding

The involvement of domain expert often guarantees the quality of annotation, but much higher cost is introduced at the same time. In this paper we try to further reduce the efforts made by doctors in translating the patterns, while keeping the diversities of the translated patterns to cover the in-domain writing styles as much as possible.

A higher order pattern *A* may be composed of two lower order patterns *B* and *C*. We call *A* covers *B* and *C* if all of them are linguistically complete. Consider an example. The 5-gram pattern (1) in Section 2 is the concatenation of the 3-gram pattern "SURGERY was performed" and the 2-gram pattern "on DATE". After pattern (1) is translated, we can derive the translations of their lower order composing components without translations by

Table 7. Examples of coverage relations.

Coverage Relation	Higher Order Patterns *A*
	Lower Order Patterns *B*
	Lower Order Patterns *C*
4=2+2	**BODY SURGERY** on **DATE**
	BODY SURGERY
	on **DATE**
5=2+3	local finding showed left **DIAGNOSIS**
	local finding
	showed left **DIAGNOSIS**
5=3+2	Elevated **TEST** level was noted
	Elevated **TEST** level
	was noted

experts. By this coverage relation of "5=3+2", we keep only pattern (1) and omit its 3-gram and 2-gram components. Other examples of coverage relations are given in Table 7.

Translating the higher order patterns not only extends the translations of its components, but also gives the correct ordering of their combination. Thus, keeping the covering patterns and ruling out the covered ones will reduce the size of extracted patterns and preserve their integrity at the same time.

4.5 Pattern Clustering

Pattern clustering partitions a set of patterns into subgroups, and output patterns from each subgroup in a specific order. This process further reduces the cost of expert involvements in translating the patterns.

Given a cluster of similar patterns, translating the most representative pattern may imply the translations of the others in the same cluster. An example of a cluster of similar patterns is illustrated below:

he received **SURGERY** on **DATE**
he received **TEST** on **DATE**
he underwent **SURGERY** on **DATE**
he underwent **TEST** on **DATE**

If the first pattern is translated by an expert, the translations of the others are easy to be inferred without doctors' help. Consequently, we prevent doctors from translating similar patterns, and thus enrich the diversity of their efforts.

In clustering, we define the similarity between two n-gram patterns to be the number of identical words in identical positions. Two n-grams are placed into the same cluster if their similarity is not less than n-1. Single-link clustering is adopted.

To achieve the diversity of patterns, we present them in a round-robin style among the groups generated by the clustering algorithm. Due to the large number of groups and the limited human resources, we present the patterns in a specific order by measuring the inter-group and intra-group scores. On the one hand, groups are ranked by sum of frequencies of their patterns. On the other hand, patterns are ranked by their frequencies in each group. In this manner, we focus on translating the most significant patterns among the groups.

5. Pattern Translation

This section introduces translation resources to build bilingual significant patterns. Doctors are involved in translating class patterns, while lexical patterns are translated by free online translation system first, and then corrected by doctors.

Table 8. WSD problems in medical summary domain. Bold text shows the words having ambiguous senses.

English Pattern	Wrong Translation by MT
	Correct Translation by Doctor
the bulging **mass** progressively enlarged	質量
	腫塊
a total of six **courses**	課程
	療程
visited our hospital for help	參觀
	到

5.1 Translation by MT system

With medical entity classification described in Section 4.1, domain specific terms are identified and transformed into medical classes. As a result, the lexical patterns extracted from the transformed corpus contain only common words, and can be translated by MT systems without the OOV problem.

We use Google Translate to translate the lexical patterns. Then doctors review and correct these translations. Building the bilingual patterns based on MT outputs can save much more time from the experts, compared to starting from scratch with only monolingual patterns.

5.2 Translation by Doctors

We deploy the doctors to translate the class patterns, which contain medical classes and require in-domain knowledge from experts. We design a Web UI for the doctors. To focus on the translation quality, we make efforts on the user friendly interface to reduce the editing steps of manual translations, and to help the doctors understand the meanings of the patterns. Figure 3 gives a snippet of the online annotation UI.

The first column lists candidate patterns in the order we arrange as described in Section 4.5. The classes in the pattern are shown in traditional Chinese characters, with the mapping shown as follows.

手術：**SURGERY**
診斷：**DIAGNOSIS**
藥物：**DRUG**
檢驗：**TEST**
部位：**BODY**
時間：**DATE**
代名詞：**Pronouns**

The doctors inspect these patterns and edit the translations in the second column. To facilitate the editing, each class is output with a single click. Consider the example of the first pattern in the figure. The lexical part of the target language pattern ("顯示") is edited by the doctor. On the other hand, the classes BODY (部位), TEST (檢驗) and DATE (時間) are output by mouse clicks on the corresponding classes of source language pattern to save editing time. Note a translation is left blank in Figure 3. Patterns are denied the translations if they are considered unimportant by doctors.

Some patterns are relatively hard to understand and translate, and therefore we provide the doctors with several instances for each

Figure 3. Web UI for translating class patterns.

pattern for reference. For each source pattern in our UI, up to 10 sample sentences where the pattern occurs are given, and the patterns in the sample sentences are highlighted. By clicking on the third column, these sample sentences are shown in a popup window.

5.3 Review by Doctors

For lexical patterns translated by Google Translate, the translation quality may be sabotaged due to the domain specific usages. For example, word sense disambiguation (WSD) problem often causes translation errors by such a general purpose translation system, as illustrated in Table 8.

Each lexical pattern is reviewed by doctors and corrected to the domain specific usages. Modifying these patterns is faster than translating class patterns, since the doctors make corrections only on the error parts, and keep the others untouched. The results will be analyzed in detail in Section 7.

6. Pattern Integration

In this paper we extract and translate significant patterns from the medical summary corpus, and attempt to achieve domain adaptation by integrating these bilingual patterns into a general domain SMT system. Since we use n-gram patterns that are consecutive sequences of tokens, the integration can be carried out without major changes to the background SMT system. We set up a phrase-based SMT system using Moses. Lexical patterns can serve as a separate phrase table, as proposed in [3], to provide domain specific translation options during the decoding stage. Because class patterns are mostly used in the medical domain and their translations by doctors are unlikely to be ambiguous, we adopt the translations of these class patterns in each input medical summary, and start decoding from the partial hypothesis. This is feasible with the support of some advanced functions, such as XML markup and continuing partial translation, in the current version of Moses.

The adapted SMT system serves as the prototype medical summary translator that translates medical summaries from English to Chinese. In Stage 3 of our framework, doctors modify translation results into Chinese medical summaries, which are not only read by patients, but also learned by the current SMT system. With these development data, the system will be tuned to the medical domain, and updated with new patterns in an iterative style. On the one hand, parameters are adjusted for each component, such as language model, reordering model and phrase table [13]. On the other hand, new patterns might be discovered and integrated into the system, through either statistical analysis or identification by doctors. The updated medical summary translator is expected to produce translations with better quality in the next iteration.

7. Experiments

Acquiring the bilingual patterns constitutes one of the most expensive parts of the overall framework due to the involvements of doctors. The cost of expert translation depends on the algorithms for mining the significant patterns. We evaluate the performance of our pattern miner from two aspects: significance (Section 7.1) and diversity (Section 7.2), which are addressed in steps (a)-(c) and steps (d)-(e) respectively in Section 4. We then discuss the quality of translated lexical patterns (Section 7.3).

We conduct experiments on NTUH medical summaries in 2010. For pattern identification as described in Section 4, we perform medical entity classification, and use Ngram Statistics Package (NSP) [2] to enumerate patterns and calculate their frequencies. Then we filter out non-linguistic with Stanford Parser and calculate coverage relations.

Table 9. Number of patterns after each step.

	NSP	Parser	Coverage
5-gram	2.6M	7.6K	7.6K
4-gram	2.3M	14.7K	10.8K
3-gram	1.6M	19.1K	12.5K
2-gram	0.7M	15.8K	9.2K
Total	7.2M	57.2K	40.1K

Table 10. Results of pattern clustering for class patterns.

N-gram	#Patterns	#Clusters	Avg. Cluster Size
5	4,634	2,149	2.17
4	6,229	1,957	3.18
3	5,099	753	6.77
2	2,097	14	149.79

Table 11. Results of translating class patterns.

N-gram	Accept	Wrong	Ignore	Accuracy
5	642	6	432	59.22%
4	348	3	152	69.00%

Table 9 shows the number of n-grams after each step. The patterns after the Parser step are below 1% of those extracted by the NSP tool. Most of the patterns filtered by Stanford Parser contain conjunctions, prepositions or adjectives at their end. The patterns are further reduced after the Coverage step. Note that 5-gram patterns remain unchanged in this step, since 5 is the highest order among the extracted n-gram patterns. After the Coverage step, we translate the class patterns and lexical patterns to obtain bilingual patterns.

For class patterns, we perform pattern clustering to these patterns and present them to the doctors as described in step (e). The statistics of class patterns and derived clusters are shown in Table 10. Here we sample 5-grams and 4-grams for translations. Each cluster contains 2.17 and 3.18 patterns on the average for 5-grams and 4-grams, respectively. Note that for lower order patterns, the average cluster sizes are much larger and patterns in each cluster are less similar. We plan to investigate other clustering methods for lower order n-grams in future work.

We ask 32 NTUH residents to translate class patterns in larger clusters to achieve diversity. These domain experts are instructed to use the Web UI by an on-site tutorial. Then, they examine each pattern in the order we present. Based on their expertise, they translate common patterns in medical summaries, while ignoring those considered unimportant.

For lexical patterns, we experiment on 5-grams which are translated by Google Translate first. These bilingual patterns are then reviewed by one NTUH visiting staff. The translations of the patterns are either accepted or modified based on the doctor's expert knowledge.

Table 12. Common error patterns. Error parts are underlined.

English Pattern	Error Type
under the impresison of **DIAGNOSIS**	typo
was admitted for schedualed **SURGERY**	typo
DIAGNOSIS and **DIAGNOSIS** was noted	grammar
DRUG and **DRUG** was given	grammar

Table 13. Extension of bilingual class patterns

N-gram	Doctor	Coverage	Clustering	Total
5	636	+0	+1,303	1,939
4	345	+1,141	+1,750	3,236

7.1 Significance of Class Patterns

As illustrated in Figure 3, the doctors use the Web UI to translate class patterns by either accepting or ignoring them. They consider the former as significant patterns and translate them into Chinese. In contrast, the latter that cannot be translated are non-significant. After the translation, we collect the bilingual patterns and have them inspected again by the system administrator.

Table 11 shows the overall results of the translation. Except for few errors that give the wrong translation, most of the translated patterns are applicable for integration into the background SMT. The accuracy of the presented patterns is defined as

$$\text{Accuracy} = \frac{\text{Accept - Wrong}}{\text{Accept + Ignore}}$$

which is 59.22% and 69.00% for 5-gram and 4-gram patterns, respectively. This demonstrates the effectiveness of our strategy to select linguistic patterns. Among the ignored patterns, some contain incorrect medical classes, due to misclassified terms in our terminological databases. Parsing errors also produce some non-linguistic patterns.

7.2 Diversity of Class Patterns

As shown in Table 11, initially fewer than 1000 patterns are translated by the doctors. We further exploit the diversity of these patterns by extending them based on coverage relation and pattern clustering in steps (d) and (e) of Section 4. For each translated 5-gram pattern, we produce a new 4-gram bilingual pattern if such a coverage relation exists. For each cluster with at least one pattern translated by a doctor, we manually uncover the translations of other similar patterns. During the manual extension, we discard some improper patterns, which contain errors such as typos and incorrect grammars. Table 12 shows some examples of the discarded patterns. These common errors may be incorporated into auxiliary modules of our MT system or post-editing system, such as grammar checker.

Table 13 reports the results of our extension methods, showing the newly discovered patterns after the Coverage and Clustering steps. The 5-gram and 4-gram patterns after the extensions are 3.05 and 9.38 times more than those translated by the doctors. Compared to average cluster sizes in Table 10, we better allocate our expert efforts and achieve high diversity among the translated patterns.

Table 14. Error analysis of bilingual lexical patterns. Note that a translation may have multiple errors.

Error Type	Count	Percent
WSD	411	50.12%
Style	205	25.00%
Reordering	163	19.88%
Other	232	28.29%

7.3 Error Analysis of Online Translator

After the visiting staff reviewed the lexical patterns, we analyze the data in order to examine the effectiveness of general online translator applied to the specific domain. Among the 1,174 reviewed bilingual 5-gram lexical patterns, 354 of them are remained unchanged, while the others are modified. In other words, the accuracy of Google Translate is only 30.15%.

We perform manual analysis on these 820 corrected translations, and give the results in Table 14. Half of the translations have WSD errors as illustrated in Section 5.3. This suggests domain gap is a practical issue in SMT applications. Disagreements with writing styles mainly come from regional variation of language between mainland China and Taiwan. For example, both "烟鬼" and "老煙槍" are the translations of "heavy smoker", but Taiwanese prefers the later. In addition, the lack of training corpora in medical domain causes some reordering errors (incorrect word orders between source and target language). Other translation errors include miss words, extra words, etc. Evaluation and analysis of MT output is itself an important research issue, and readers can refer to [15] as a gentle introduction.

8. Conclusion and Future Work

We proposed a framework to build a statistical medical summary translation system. We collected and organized the in-domain resources. Significant patterns in medical summary were identified and translated with the expert involvements. The approaches to incorporate bilingual patterns into the background SMT system were also discussed. One of the main concerns throughout the proposed framework is to reduce the cost of expert translation. We identified and arranged the significant patterns with high quality and diversity. We designed a user friendly interface and applied an online translator to save the translation time of the doctors.

The experiments were performed on the NTUH medical summaries. The results showed the significance of the presented patterns, and the diversity of the translated bilingual patterns. In future work, we will build a medical summary SMT system, based on the acquired bilingual patterns. We will also investigate ways for tuning the adapted system by supervised learning techniques, with the continuous help from the doctors.

9. ACKNOWLEDGMENTS

We would like to thank NTUH doctors for their help in translating the medical summary patterns.

10. REFERENCES

[1] 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.

[2] Banerjee, S. and Pedersen, T. 2003. The design, implementation, and use of the Ngram Statistics Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381.

[3] 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.

[4] 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.

[5] Embarek, M. and Ferret, O. 2008. Learning patterns for building resources about semantic relations in the medical domain. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2006-2002.

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

[7] 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.

[8] Klein, D. and Manning, C. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL 2003*, pages 423-430.

[9] Koehn, P., Och, F. J., and Marcu D. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL 2003*, pages 127-133.

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

[11] 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.

[12] Loper, E. and Bird, S. 2002. NLTK: the natural language toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*.

[13] Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

[14] Shadow, G. and MacDonald, C. J. 2003. Extracting structured information from free text pathology reports. In *Proceedings of AMIA 2003 Annual Symposium,* pages 584-588.

[15] Vilar, D., Xu, J., D'Haro, L. F. and Ney H. 2006. Error analysis of statistical machine translation output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 697-702.

[16] 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.