A Pilot Statistical Medical Summary Translation System

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

Keywords: Domain Adaptation, Medical Summary, Machine Translation, Pattern Identification

1. Introduction

In cross domain MT applications, the differences of vocabularies and linguistic structures between the two domains affect the performance of MT systems. Therefore, adapting MT systems from a generic domain to a specific domain needs various in-domain resources, such as bilingual dictionary and corpus. However, even those MT systems whose source and target languages are common in real world may still face the resource poor problems. This demo considers the translation of medical summaries as the research target.

In Taiwan hospitals, medical summaries are usually written in English. That results in the absence of English-Chinese aligned medical summaries. Thus, medical summary translation is a very good example to investigate the cross domain MT problem. Moreover, Department of Health of Taiwan government announced a new law on March 12, 2010. Based on the new law, every hospital should provide patients with Chinese medical summaries. Thus, how to utilize MT to assist hospitals and patients to reduce the language barrier becomes a very important, practical and emergent issue.

In this research, we develop a general English-Chinese SMT system with Moses toolkit [2], and present a framework to adapt it to build a medical summary SMT system [1].

2. A Framework

Figure 1 shows a MT framework 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 bilingual patterns is the goal. We are provided with raw medical summaries and terminological resources. These in-domain data is organized into 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 extracts the 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 hand, post-editing and log analyses are beneficial for tuning and optimizing the system by machine learning techniques.

The system is been developing at the first two stages. The kernel of Stage 1 is discussed in Section 3.

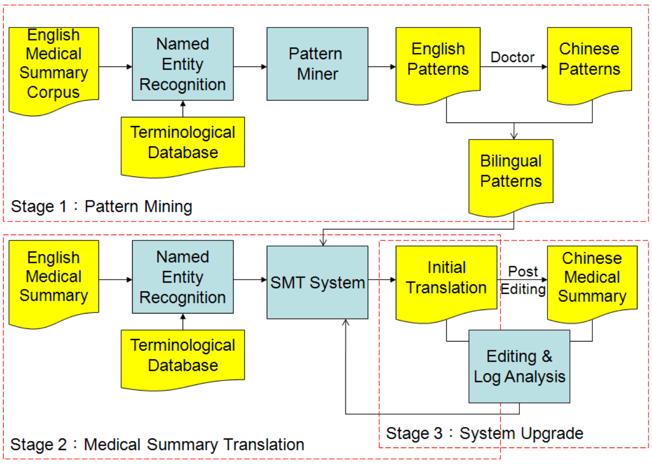
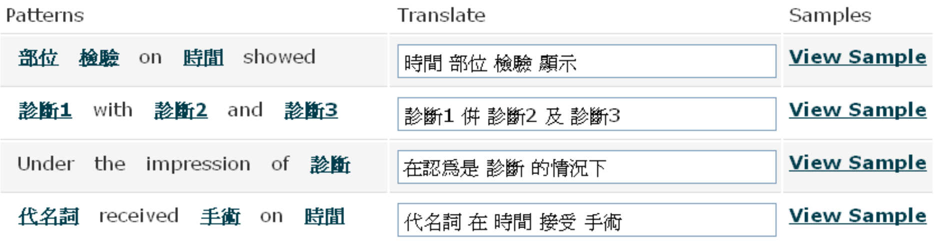


Figure 1. An Overall Framework for Medical Summary Translation



Patterns	Translate	Samples
部位 檢驗 on 時間 showed	時間 部位 檢驗 顯示	View Sample
診斷1 with 診斷2 and 診斷3	診斷1 併 診斷2 及 診斷3	View Sample
Under the impression of 診斷	在認爲是 診斷 的情況下	View Sample
代名詞 received 手術 on 時間	代名詞 在 時間 接受 手術	View Sample

Figure 2. Web UI for Translating Patterns

3. Pattern Mining and Translation

Provided with a large English medical summary corpus and terminological databases, we aim to (1) extract the significant patterns to capture the domain specific writing style as much as possible; (2) reduce the size of the pattern set to minimize the cost of doctors in translating the patterns. The overall steps for pattern mining 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 (n=2~5) in the new corpus to extract a set of frequent patterns [3].

(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 relationship 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.

For pattern translation, doctors are involved in translating the mined patterns. Figure 2 gives a snippet of the online annotation UI for collecting the bilingual patterns.

Acknowledgements

This work was partially supported by National Science Council, Taiwan under grant NSC 101-2221-E-002-195-MY3. We would like to thank NTUH doctors for their help in translating the medical summary patterns.

References

[1] Han-Bin Chen, Hen-Hsen Huang, Jengwei Tjiu, Ching-Ting Tan and Hsin-Hsi Chen, “A Statistical Medical Summary Translation System,” *Proceedings of 2012 ACM SIGHIT International Health Informatics Symposium*, Miami, Florida, USA, January 2012, pp. 101-110.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constrantin, and E. Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation,” *Proceedings of ACL 2007, Demonstration Session*, Prague, Czech Republic, June 2007, pp. 177-180.

[3] S. Banerjee and T. Pedersen, “The Design, Implementation, and Use of the Ngram Statistics Package,” *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, February 2003, pp. 370-381.