Identifying Symptoms and Diseases in MedNLP Japanese Materials Using Chinese Resources

Lun-Wei Ku
Institute of Information Science, Academia Sinica
128 Academia Road, Section 2
Nankang, Taipei 115, Taiwan
+886-2-27883799 ext 1808
lwku@iis.sinica.edu.tw

Edward T.-H. Chu Cheng-Wei Sun Wan-Lun Li
CSIE, National Yunlin University of Science and Technology
123 University Road, Section 3
Douliou, Yunlin 64002, Taiwan
edwardchu@yuntech.edu.tw; chengwei.kenny.sun@gmail.com;
u9817022@yuntech.edu.tw

ABSTRACT
In this paper, we describe the Sinica-Yuntech system (TeamID: SinicaNLP) at the NTCIR-10 MedNLP task. Materials of the MedNLP task are in Japanese. However, having only Chinese resources and knowledge, we need to translate these materials into Chinese. Two preprocessing approaches, different in the timing of translation, were taken. One was to translate Japanese sentences in to Chinese ones, and then to perform segmentation and part of speech tagging on these Chinese sentences; the other was to segment and tag parts of speech on Japanese sentences, and then to translate the composite words. After knowing words and their parts of speech, we identified symptoms and diseases by a vocabulary matching approach. The Internet searching results and parts of speech patterns were also utilized to recognize out of vocabulary symptoms. After recognizing the targets in Chinese, a reverse translation was performed in order to label the original Japanese materials. We merged the tags from vocabulary matching, Internet searching and pattern mapping to obtain the performance of our best run: an f-score 53.88 and an accuracy 91.46.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Test analysis.*

General Terms
Algorithms, Performance, Experimentation.

Keywords
Cross-lingual, medical natural language processing, part of speech pattern, symptom, disease.

Team Name
SinicaNLP

Subtasks
Complaint and diagnosis

1. INTRODUCTION
Medical information management has drawn a lot of attention recently[1]. In bioinformatics research domain, extracting relations between medical terms has been widely studied. For example, relations between diseases and durgs [2], relations between diseases and genes [3] and drug-drug interactions [4] were also extracted. In these researches, the medical terms were usually predefined. Therefore, most researches extract relations from limited sources and formal documents, such as papers or databases. To extend the applicability of these technologies to the medical reports, descriptions in medical records or the Web to find more information, automatic medical term extraction is always the first research problem encountered. Hence, the related techniques are indispensable.

MedNLP is a pilot task in NTCIR-10 [5]. The short-term objective of this pilot task is to evaluate basic techniques of information extraction in medical fields. We participated in the complaint and diagnosis subtask, for which we should extract the complaint and diagnosis from medical reports written by physicians. As all our resources were in Chinese and experimental materials were in Japanese, we tried to translate them into Chinese and then extracted terms. Two methods which translated materials in different timing were proposed. Terms extracted by vocabulary matching, part of speech (POS) pattern matching, and Internet searching were all collected to report final results.

2. SYSTEM FRAMEWORK
Figure 1 shows the system flow of preprocessing first and then translating, and Figure 2 shows that of translating first and then preprocessing. Here preprocessing includes word segmentation and POS tagging. As in Figure 1, if we translate after preprocessing, we don't have to translate the materials back to Japanese as we have Japanese word tokens to be translated and the final labeled Chinese string can be aligned to its original Japanese word tokens. In Figure 2, we need to translate the identified Chinese strings back into Japanese in order to search these words in the original Japanese sentence for labeling. Translation was performed by Excite Translate[2], Google Translate and Bing Translate. All translation results for each word were used in vocabulary matching. The database for vocabulary matching included 7,796 symptoms, 7,498 diseases, and 749 additional medical terms. All these terms were collected from medical web pages[3,4].

[1] https://sites.google.com/site/datadrivenwellness/cfp

[2] http://www.excite.co.jp/world/chinese/

[3] http://cht.a-hospital.com/w/%E9%A6%96%E9%A1%B5

[4] http://zz.qqyy.com/

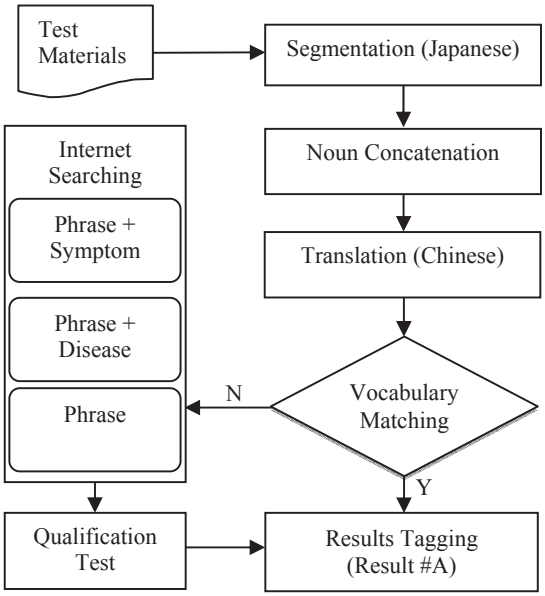

Figure 1. System Flow: Preprocess and Then Translate

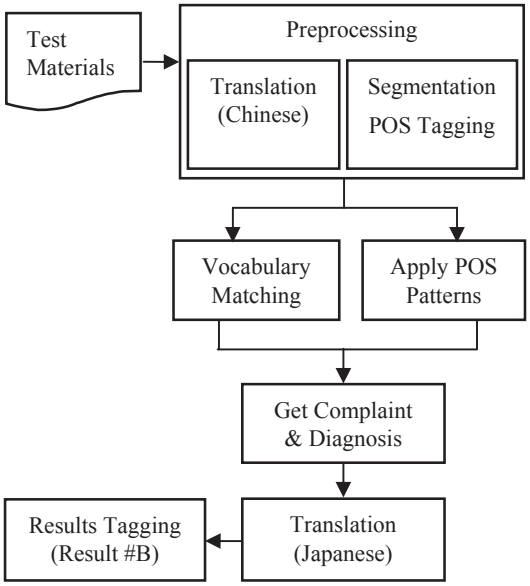

Figure 2. System Flow: Translate and Then Preprocess

We submitted two runs. Run 1 merged Result #A and Result #B, while Run 2 reported only Result #B. As to identifying out of vocabulary (OOV) symptoms and diseases, the major differences between the approaches in two figures were *Internet Searching* and the usage of *POS Patterns*. The details are described as follows.

Internet Searching. In Figure 1, Japanese test materials were segmented and POS tagged by MeCab[5]. We first concatenated sequential nouns to generate a phrase for matching. In *Vocabulary Matching*, if the concatenated phrase includes any terms in our database, it was labeled as a result string. Otherwise, *Internet Searching* will determine whether it should be included in the answer set. What *Internet Searching* did was to query three different strings: the concatenated phrase, the concatenated phrase plus a suffix "症" (symptom), and the concatenated phrase plus a suffix "病" (disease), and we then obtained the numbers of query result *num_p*, *num_s*, and *num_d* respectively. After that, we calculated the qualification degree *q*deg by formula (1):

$$q\deg = \frac{num\_p}{(num\_s + num\_d)/2} \quad (1)$$

The larger the value of *q*deg, the more possible that the phrase itself is a meaningful term and is a symptom or disease without the suffix. We set a threshold *th* 0.7, which was the heuristically best value from a grid search when experimenting on the traning data. If the value of *q*deg was bigger than *th*, we labeled the concatenated phrase as the result string.

POS Patterns. From observations we have known that many symptoms and diseases are compound nouns. POS patterns were used to find these nouns. A total of 14 patterns were designed to identify target strings. To use these patterns, first the segmentation and part of speech tagging were performed by the CKIP Chinese Segmentation System[6][1]. The general nouns of POS (Na) didn't have to be in the database in order to be concatenated, while words of the other POS must be found in the database to be considered candidates for pattern matching. The longest string matched to these patterns was labeled as the result string. The matching process scanned from left to right, and then the matched string was concatenated and its part of speech was labeled as (Na). For example, (Na)(VC)(VH) matches the pattern (Na)(VC) first, so the first two tokens were concatenated and labeled as a (Na) string. Then the original string became (Na)(VH) which matched another pattern and was labeled as one result string. Figure 3 shows the information of POS patterns.

$(Na)^+$	(VH)	(Na)(A,b,Nb,VG,VC,Nc,VJ,VE,VH)
(Na)(Suffix)	(VA)	(A,b,Nc,VH,N,D,Vc,VA,VJ)(Na)
(VH)(VF)	(VAC)	(VC,Nc)(VH)
(VHC)	(N)(A)	(Ncd)(Neqa)
(Dfb,D)(VA)	(A)(N)	

$(Na)^+$: when tokens of POS (Na) appeared sequentially, they were concatenated into one (Na) token.

(N): (Nb, Ncd, Neu, Nep, Neqb, Ng, Nh, Nv)

(Suffix): (症,症候群,病,病毒,病變,癌,炎,疹,疱,瘤,患,結石,障害,障礙,兆候)

Figure 3. POS Patterns

3. EXPERIMENT RESULT

Table 1 shows the experiment results of our system in NTCIR-10 MedNLP task. From Table 1 we found that when merging the results of searching from the Internet and adopting POS patterns for matching, recall was improved by 8.87 (20.1%) but precision dropped only 2.91 (5.6%). From the results we can say that these

[5] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[6] http://ckipsvr.iis.sinica.edu.tw/

two approaches for finding OOV targets are complementary approaches which can find different symptoms and diseases.

Table 1. Experiment Results

Run ID	Precision	Recall	F-measure	Accuracy
#1	54.75	53.03	53.88	91.46
#2	51.84	44.16	47.69	91.09

Table 2, 3 and 4 show the performance while experimenting on the training set. From these tables we can see that the performance of Result #A+#B was better than Result #B and Result #A was the worst, which were the same with the performance on the testing set.

Table 2. Performance on Training Set (Result #A)

#A	Precision	Recall	F-measure	Accuracy
w/o modality	50.48	46.72	48.53	96.77
All	44.97	41.62	43.23	97.61
<c>	41.28	42.16	41.72	
negation	58.81	39.09	46.96	
suspicion	43.28	40.28	41.73	
family	57.14	62.50	59.70	

Table 3. Performance on Training Set (Result #B)

#B	Precision	Recall	F-measure	Accuracy
w/o modality	60.51	51.82	55.83	93.14
All	54.56	46.72	50.34	93.94
<c>	54.77	48.10	51.22	
negation	53.35	41.07	46.41	
suspicion	44.93	43.06	43.98	
family	80.00	87.50	83.58	

Table 4. Performance on Training Set (Result #A+#B)

#A+#B	Precision	Recall	F-measure	Accuracy
w/o modality	62.41	66.34	64.32	95.26
All	55.46	58.95	57.15	96.00
<c>	55.10	61.26	58.02	
negation	57.96	51.98	54.81	
suspicion	41.86	50.00	45.57	
family	68.18	93.75	78.95	

4. ERROR ANALYSIS

As the gold standard for the testing set has not been released yet, we analyzed the system errors on the training set. Generally speaking, there were two kinds of errors found when applying our approaches. One was related to the lack of knowledge in the processing of foreign language materials as follows.

Difficulty of word matching in bi-directional translations. When preprocessing after translation as in Figure 2, there were difficulties to find the original Japanese terms for labeling after targets were identified in their Chinese form. For example, "麻痺症状" was translated into Chinese as "麻痺症" and then identified as a symptom. However when translated back into Japanese, it was still "麻痺症" and the final annotation was wrong though it was identified correctly.

Change of word sense in bi-direction translations. Sometimes the word sense would be altered after translations, which also caused incorrect annotations. For example, "嘔氣" was translated into Chinese as "嘔氣" or "噁心" in two different translation tools, respectively. However, when they were translated back into Japanese, these two terms became "吐く息" and "吐き気", respectively, which have different meanings with the original Japanese term.

The other kind of errors was related to the approaches, in which we did not enable the system to find some diseases or symptoms:

Symptoms with degrees or properties. We haven't handled these degrees or properties yet as they were general terms instead of proper names. Symptoms have different properties and the developed system will need to learn automatically to report them correctly. Here are some examples: 炎症反応高値 (high value), 透過性低下 (low), 抗体陰性 (negative), 咳嗽症状減少 (decrease), 四肢冷感 (cold).

Diseases or symptoms composite of English, numbers or punctuation marks. This kind of diseases or symptoms might be segmented into two or more words in the pre-processing. Therefore it is difficult to recognize them afterwards. For example, ＥＳＲ, ＰＩＰ関節裂隙狭小化, ＣＰＫ高値, 腎症Ⅱ～Ⅲ期, ２型糖尿病 are those with English or numbers; "腹部平坦，軟" and " 膨隆、軟" contains punctuation marks and hence system reported "腹部平坦" and "膨隆", respectively.

General disease or symptom categories. These terms were reported by the system as false alarms. For example, In the sentence "〔自律神経系〕排尿・排便障害：排尿困難、便秘、起立性低血圧なし", "排便障害" was reported wrongly by the system; in the term "縦隔肺門に石灰化", "石灰化" was reported wrongly by the system.

5. CONCLUSION AND FUTURE WORK

In this paper, we have developed a system to find complaint and diseases in a foreign language, Japanese, for the MedNLP task in NTCIR-10. We translated Japanese materials into Chinese, process them, and translated them back into Japanese for annotation. In the whole process, we suffered from cross-lingual issues beside technical difficulties in this task, such as translation errors, incapable to match, limited resources, and so on. However, we still tried two different approaches of translating materials and proposed a web search method to assist the identification of the targets. Possible issues in processing foreign medical materials were also reported.

As we did not adopt machine learning methods, resources were very critical to us. For example, without appropriate knowledge, word segmentation before the identification might have caused the mismatch of complaints and diseases and hence deteriorated the performance. In the future, we would like to collect more Japanese resources in our methods to avoid the translation process and obtain the baseline performance. We would use our own translation techniques to keep the word alignment information. We would also try doing the longest match of targets without word segmentation to see whether the performance can be improved. As we found that contexts surrounding different targets varied, we may consider the information within a specific context window to help the identification.

6. ACKNOWLEDGMENTS

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC 101-2628-E-224-001-MY3.

7. REFERENCES

[1] CKIP (Chinese Knowledge Information Processing Group). 1995/1998. *The Content and Illustration of Academica Sinica Corpus*. Technical Report no 95-02/98-04. Taipei: Academia Sinica.

[2] Elizabeth S. Chen, George Hripcsak, Hua Xu, Marianthi Markatou and Carol Friedman. 2008. Automated Acquisition of Disease–Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. *Journal of the American Medical Informatics Association*, 15(1), pp.87-98.

[3] Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki and Jun'Ichi Tsujii. 2006. Extraction of Gene-Disease Relations From Medline Using Domain Dictionaries and Machine Learning. In *Proceedings of Pacific Symposium on Biocomputing* 11. pp.4-15.

[4] Heleen van der Sijs, Laureen Lammers, Annemieke van den Tweel, Jos Aarts, Marc Berg, Arnold Vulto and Teun van Gelder. 2009. Time-dependent Drug–Drug Interaction Alerts in Care Provider Order Entry: Software May Inhibit Medication Error Reductions Original Research Article. *Journal of the American Medical Informatics Association*, 16(6), pp.864-868.

[5] Mizuki Morita,Yoshinobu Kano,Tomoko Ohkuma, Mai Miyabe, Eiji Aramaki. 2013. Overview of the NTCIR-10 MedNLP Task. In *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, June 16-19, 2013, Tokyo, Japan.