HCRL at NTCIR-10 MedNLP Task

Osamu Imaichi
Hitachi, Ltd., CRL
1-280, Higashi-Koigakubo
Kokubunji-shi, Tokyo 185-8601
osamu.imaichi.xc@hitachi.com

Toshihiko Yanase
Hitachi, Ltd., CRL
1-280, Higashi-Koigakubo
Kokubunji-shi, Tokyo 185-8601
toshihiko.yanase.gm@hitachi.com

Yoshiki Niwa
Hitachi, Ltd., CRL
1-280, Higashi-Koigakubo
Kokubunji-shi, Tokyo 185-8601
yoshiki.niwa.tx@hitachi.com

ABSTRACT

This year’s MedNLP[1] has two tasks: de-identification and complaint and diagnosis. We tested both machine learning based methods and an ad-hoc rule-based method for the two tasks. For the de-identification task, the rule-based method got slightly higher results, while for the complaint and diagnosis task, the machine learning based method had much higher recalls and overall scores. These results suggest that these methods should be applied selectively depending on the nature of the information to be extracted, that is to say, whether it can be simply patternized or not.

Team Name

HCRL

Subtasks

De-identification
Complaint and diagnosis

Keywords

sequential labeling, CRF, unsupervised feature learning.

1. INTRODUCTION

Machine learning based and rule-based methods are the two major approaches for extracting useful information from natural language texts. In order to clarify the pros and cons of these two approaches, we applied both approaches to this year’s MedNLP tasks: de-identification and complaint and diagnosis.

For the de-identification task, ages and times, for example, are seemingly a type of information that can be patternized quite simply. In such cases, an ad-hoc rule-based method is expected to deliver a relatively good performance. In contrast, the complaint and diagnosis task would seem much more difficult to patternize, so a machine learning approach is expected to provide an effective methodology for tackling these problems.

2. MACHINE LEARNING APPROACH

In this section, we explain how machine learning based approach works.

2.1 Sequential Labeling by using CRF

We formalized the information extraction task as a sequential labeling problem. A conditional random field (CRF)[2][1] was used as the learning algorithm. We used CRFsuite[1], which is an implementation of first order linear chain CRF.

The CRF-based sequential labeling proceeded as follows. First, we applied a Japanese morphological parser (MeCab [2]) to documents and segmented the sentences into tokens with part-of-speech and reading. Then, the relationship between tokens was estimated using CaboCha[3], which is a common implementation of the Japanese dependency parser[3]. Finally, we extracted the features of the tokens and created models using CRFsuite.

2.2 Basic Features

We used the following features to capture the characteristics of the token: surface, part-of-speech, and dictionary matching. The surface and part-of-speech of the target token were converted into numerical expressions in what is known as one-hot representation: the feature vector has the same length as the size of the vocabulary, and only one dimension is on. The dictionary feature is a binary expression that returns one if a word is in the dictionary and zero otherwise.

We prepared ten kinds of dictionaries featuring age expressions, organ names, Japanese era names, family names, time expressions, names of hospital departments, disease names from the Japanese Wikipedia, Chinese characters related to diseases, suspicious expressions, and negative expressions. These dictionaries were created based on the rules explained in Section 3.

In order to capture the local context of a target token, we combined features of several neighbor tokens. First, we merged the features of five adjacent tokens. Let w_i be the i-th token of the sentence. We concatenated the features of w_{i-2}, w_{i-1}, w_i, w_{i+1}, and w_{i+2} and created $w_{[i-2:i+2]}$ to express the i-th node. Second, we concatenated the features of $w_{[i-2:i+2]}$ and w_i^{src} (w_i^{tgt}) to denote the source (target) token of w_i.

2.3 Unsupervised Feature Learning

In addition to the basic features, we used clustering-based word features[4] to estimate clusters of words that appear only in test data. These clusters can be learned from unlabeled data by using Brown's algorithm[5], which clusters words to maximize the mutual information of bigrams. Brown clustering is a hierarchical clustering algorithm, which means we can choose the granularity of clustering after the learning process has been finished.

We examined two kinds of Brown features: those created from training and test data related to the MEDNLP task (1,000 categories) and those created from the Japanese Wikipedia (100 categories). We decreased the number of categories of the latter because clustering Wikipedia is computationally expensive. The computational time of Brown clustering is $O(VK^2)$, where V denotes the size of vocabularies and K denotes the number of categories.

[1] http://www.chokkan.org/software/crfsuite/.

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

[3] http://code.google.com/p/cabocha/.

3. RULE-BASED METHOD

In this section, we explain the rule-based method.

3.1 De-identification task

-- <a>: age,

- ✧ The basic pattern is “d1[歳才台代]”, where d1 is a positive integer, and [ABC] refers to A, B, or C.
- ✧ If an age region is followed by a specific modifiers (時|頃|[こご]ろ|代|[前後]半|以[上下]), that region is expanded to the end of the modifier. A disjunctive expression “aaa|bbb|ccc” means aaa, bbb, or ccc.
- ✧ Further details are omitted.

-- <t>: time,

- ✧ The basic pattern of time tags is “d1 年 d2 月 d3 日 d4 時 d5 分 d6 秒”, where d1 to d6 are non-negative integers. Any partial pattern starting from d1 or d2 or d3 is also eligible.
- ✧ The special numerical pattern d1/d2 (1900 <= d1 <= 2099, 1 <= d2 <= 12) is interpreted as year = d1 and month = d2. In addition, the special numerical pattern d1/d2 [に|から|より|まで|〜] (1 <= d1 <= 12, 1 <= d2 <= 31) is interpreted as month = d1 and day = d2.
- ✧ Exceptional patterns are: “[同当即翌前][日年月]|翌朝|翌未明|その後”.
- ✧ Further details are omitted.

-- <h>: age,

- ✧ First, hospital tags were added by using the below hospital-words dictionary composed of 7 words, and temporary division tags were added by using the division-words dictionary of 27 words.
  - ➢ Hospital words: 当院|近医|同院|病院|クリニック|総合病院|大学病院
  - ➢ Division words: 外科| 眼科|循環器内科|皮膚科|内科 … etc. (27 words)
- ✧ While a hospital region is preceded by any number of division regions, the hospital region is extended to the beginning of the division regions.
- ✧ Futher details are omitted.

-- <p>: person name, This tag was skipped.

-- <x>: sex, The sex tags were added only by a simple pattern: “男性|女性”.

3.2 Complaint and diagnosis task

- ✧ All <c> tags of the training data were extracted and a dictionary of complaints was made. The dictionary contains 1,068 words.
- ✧ The <c> tags were added to the test data by the longest match method using this dictionary. In case of a single character word (咳 and 痰), a tag is added only if both the preceding character and the following character are not Kanji characters.
- ✧ If a <c> tag region is followed by the cancelling expressions below, the <c> tag is cancelled.
  - ➢ postfix type cancelling expressions:
  - ➢ [歴剤量時室率]|検査|教育|反応|導入|胞診|精査|を?施行|培養|細胞|成分
  - ➢ 取り?扱|ガイ[ダド]|分類基準|[^予防]*予?防|[^療]*療法|＝[0-9]
- ✧ Further details together with the methods of adding three types of modality (negation, suspicion, family) are omitted.

4. RESULTS

4.1 De-identification task

	P	R	F	A
HCRL-1 (rule)	89.59	91.67	90.62	99.58
HCRL-2 (machine learning)	92.42	84.72	88.41	99.49
HCRL-3 (machine learning)	91.50	84.72	87.98	99.46

4.2 Complaint and diagnosis task

	P	R	F	A
HCRL-1 (rule)	72.47	58.12	64.50	93.40
HCRL-2 (machine learning)	88.98	74.24	80.94	96.08
HCRL-3 (machine learning)	88.55	75.32	81.40	96.06

5. Conclusion

For the de-identification task, the rule-based method got slightly higher results. while for the complaint and diagnosis task, the machine learning based method had much higher recalls and overall scores. These results suggest that we use these methods selectively depending on the nature of the information to be extracted, that is to say, whether it can be simply patternized or not.

6. REFERENCES

[1] Morita, M., Kano, Y., Ohkuma, T., Miyabe, M., and Aramaki, E. 2013. Overview of the NTCIR-10 MedNLP Task. In P*roceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.*

[2] Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, 282-289.

[3] Kudo, T. and Matsumoto, Y. 2002. Japanese Dependency Analysis using Cascaded Chunking, *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, 63-69.

[4] Turian, J., Ratinov L., and Bengio, Y. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 384-394.

[5] Brown, P. F., deSouza P. V., Mercer R. L., Pietra, V.J.D., and Lai, J.C. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18, 467-479.