Incorporating Unsupervised Features into CRF based Named Entity Recognition

Yuki Tawara
Nara Institute of Science and Technology
tawara.yuki.tn7@is.naist.jp

Mai Omura
Nara Institute of Science and Technology
omura.mai.oz5@is.naist.jp

Mirai Miura
Nara Institute of Science and Technology
miura.mirai.me1@is.naist.jp

ABSTRACT

We participated in the extraction of complaint and diagnosis Task and the normalization of complaint and diagnosis Task of MedNLP2 in NTCIR11. In the extraction Task, we use CRF based Named Entity Recognition method. Moreover, we incorporate unsupervised features learned from raw corpus into CRF. We show such unsupervised features improve system performance.

Team Name

CL

Subtasks

Task 1 (Extraction task)
Task 2 (Normalization task)

Keywords

Named Entity Recognition, Conditional Random Fields, Brown Clustering, Word Representation

1. INTRODUCTION

In medical fields, applications of electronic media to information management have been increasing. For example, clinical records have been shifted to electronic media. As a result, utilizing clinical records is desired strongly. Most of information in clinical records is written in natural language, so utilizing electrical record requires Natural Language Processing (NLP) techniques. However, NLP technique in medical fields is far from well developed.

We developed a system for the extraction of complaint and diagnosis Task and the normalization of complaint and diagnosis Task. The extraction of complaint and diagnosis Task is a task to extract expressions which represent complaint, diagnosis and time expressions related to a patient, from clinical records prepared for this competition. Normalization of complain and diagnosis Task is task to assign ICD-10 class tags to extracted complaint and diagnosis. In addition, we constructed a system to assign modality tags to extracted complaint and diagnosis. For detail of two task and ICD-10, see [1].

The rest of this paper is organized as follows. Section2 explains the method used in our system. Section3 describes experiments we conduct for evaluate our system and its results. Finally Section4 concludes this paper.

2. PROPOSED METHOD

B I I O O O O
多発 | 性 | 筋炎 | と | 考え | られる | 。

Figure 1: An example of assignment of IOB tags to sequence of morphemes

2.1 Extraction of Complaint and Diagnosis

Extracting terms related to some domain is refered to Named Entity Recognition (NER). Popular methods used in NER include rule based method, machine learning method such as Maximum Entropy Model, Conditional Random Fields (CRF) [4]. In this paper, we use CRF, which is reported to archive high performance [8], in extracting named entities.

2.1.1 Named Entity Recognition using CRF

NER can be considered to assigning IOB tags to sequence of morphemes like the Figure 1, and in such a way it is formalized as sequential labeling.

B tag represents its token is located at the beginning of named entity, I tag represents it is located in inside of named entity, O tag represents it is located in outside of named entity. CRF is a statistical model which is used in sequential labeling. It is a discriminative model and has an advantage in flexibility of incorporating features. In CRF, label sequence is predicted so as to maximize conditional probability of label sequence $\boldsymbol{y}$ given tokens $\boldsymbol{x}$ as below:

$$\boldsymbol{y} = \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{\boldsymbol{Z}\boldsymbol{x}} \exp \sum_{i=1}^{n} \sum_{k} \lambda_k f_k(y_{i-1}, y_i, \boldsymbol{x})$$

$$\boldsymbol{Z}_{\boldsymbol{x}} = \sum_{\boldsymbol{y}} \sum_{i=1}^{n} \sum_{k} \lambda_k f_k(y_{i-1}, y_i, \boldsymbol{x})$$

Where $\boldsymbol{Z}_{\boldsymbol{x}}$ is normalizing constant. f_k is feature function which is defined by tokens and labels. By feature function, we can incorporate various kind of information to the model. λ_k is a weight to be learned from annotated corpus.

2.1.2 Unsupervised Features

Generally, training of CRF needs annotated data. However, amount of annotated date is limited and preparing annotated data, especially large amount of annotated data, requires large human power. On the other hand, there exist many documents which contain expressions of complaint and diagnosis mainly on the Web. Therefore, it is worth

Table 1: Some examples of marker words for each modality

Modality	Marker words
Positive	出現, みられたが, を伴っており, （+）, を感じる, 所見あり, のみ, 軽快しない
Suspicion	疑われ, 疑わしかった, あると思われ, も示唆され, 考慮され, 可能性
Negative	示唆されない, の可能性は低い, 認めず, 検出されず, 改善し, は消失, 見られていない
Family	父, 母, 長男, 妹, 息子, 妻, 夫, 祖父, 祖母

utilizing such unannotated data for training of supervised learning.

We try to utilize such unannotated data for training of CRF by using unsupervised method such as Brown Clustering and word representation. Brown Clustering [2] is a method which clusters word hierarchically by minimizing mutual information. Brown Clustering has been reported to improve accuracy of some task in fields of NLP [7], [6], [3]. Word representation is learned from raw corpus based on frequency of co-occurrence of words. Word representation is said to present word's meaning [5]. We incorporated two cluster information, one is from Brown Clustering, another is from result of clustering word representation, into features of CRF.

2.2 Assignment of Modality Tag

We assign modality tags using rules based method. Construction of assignment rules is made by hand, referencing to training data.

At first, we list marker words for each modality, which occur in front of or back of modality-assigned expression. At the stage of assignment, we check surrounding words of the assignment target. If some marker word for some modality is present, we assign a modality tag corresponding to marker word, if not, we assign a "positive" tag.

With respect to a "family" tag, we prepare another rule based on the document structure. We assign a "family" tag if the assignment target occurs in section of family history.

We show some examples of marker words in Table 1.

2.3 Assignment of ICD-10 Tag

We assign ICD-10 tags mainly based on dictionary matching method. At first, based on ICD-10 standard master[1], we construct a dictionary which consists of disease names as lemma and ICD-10 tags as content. This dictionary permits one disease lemma has several ICD-10 tags.

At the time of assignment of the tags, we consult the dictionary. If there exists a lemma which matches with assignment target expression in the dictionary and ICDs corresponding to matched lemma is only one, we assign its ICD-10 tag to target expression. If either there exists no matching lemma in dictionary or matching lemma has several corresponding ICD-10 tags, we assign ICD-10 tag t to expression e which maximize the score:

$$t = \arg\max_{t} Score_e(t)$$

This score consists of 4 subscores: DicScore, ContextScore, TextTypeScore and WebScore. Whole score is calculated by weighted sum of these subscores:

$$\begin{aligned} Score_e(t) = &\alpha\, DicScore_e(t) \\ &+ \beta\, ContextScore_e(t) \\ &+ \gamma\, TextTypeScore_e(t) \\ &+ \delta\, WebScore_e(t) \end{aligned}$$

Where α, β, γ and δ are the weights of subscores, which take non-negative real values.

Below, we explain details of four subscores.

DicScore
: This subscore calculates how much assignment target is similar as string to disease name in dictionary which has the concerned ICD-10 tag. This subscore plays a role like a similarity search.

ContextScore
: Disease names which have the same ICD-10 tag are considered to appear in similar context. This subscore is calculated based on similarity of two contexts, a context of assignment target and a context of disease names which have concerned ICD-10 tag.

TextTypeScore
: There exits "type" attribute in training corpus and test corpus of MedNLP2. This "type" attribute represents what kind of disease concerned patient has, therefore it can be important cue to assign ICD-10 tags. This subscore is calculated based on similarity of two contexts of "type", a context of "type" attribute of the record in which assignment target occurs and a context of "type" attributes of the records in which concerned ICD-10 tag occurs.

WebScore
: This subscore is calculated like "DicScore" except that a consulting dictionary is different: the consulting dictionary is based on the Web page[2].

3. EXPERIMENTS AND RESULTS

3.1 Experimental Settings

We conduct experiments to examine performance of our system. We use distributed training set for training corpus, distributed formal-run data for test corpus.

For CRF, we use open source software CRF++[3]. Addition to the training corpus, we use two medical dictionaries, ICD-10 standard master and basic master from Health Insurance Claims Review & Reimbursement Services[4], to training of CRF. These dictionaries contains about 24,000 disease names. By using medical dictionaries, we incorporate to feature of CRF whether morphemes in training corpus is contained in entries of dictionaries. Since technical term usually consists of several morphemes, we not only check morphemes in training corpus match with the entries, but also check morphemes in training corpus match with the morphemes contained in the entries. We show basic features of CRF in the Table 2.

In the actual training of CRF, we use several combinational features of basic features and tune features manually.

[1] http://www2.medis.or.jp/stdcd/byomei/

[2] http://www.dis.h.u-tokyo.ac.jp/byomei/icd10/

[3] http://crfpp.googlecode.com/svn/trunk/doc/index.html

[4] http://www.ssk.or.jp/tensuhyo/kihonmasta/index.html

Table 2: Basic feature of CRF

Feature Name	Brief Description
SURFACE	The surface form of a morpheme.
DIC-E	Flag whether token is contained in medical dictionary as a entry.
DIC-M	Flag whether token matches with a morpheme contained in entries of medical dictionary.
DIC-S	Flag whether token matches with a substring of entries of medical dictionary.
POS1	Large Part-of-Speech of a morpheme.
POS2	Middle Part-of-Speech of a morpheme.
POS3	Small Part-of-Speech of a morpheme.
POS4	Micro Part-of-Speech of a morpheme.
INF-T	Inflectional type of a morpheme.
INF-F	Inflectional form of a morpheme.
GOSYU	Unidic's information about kind of morpheme type.
CHAR-T	Character type of a morpheme.

Table 3: Extraction results of complaint and diagnosis

	Precision	*Recall*	F_1
without unsupervised	85.67 %	73.88 %	79.34 %
with brown	86.02 %	74.02 %	79.57 %
with word2vec	86.14 %	75.09 %	80.24 %
with brown and word2vec	86.51 %	75.37 %	80.56 %

For the training of Brown Clustering and word representation, we use about 10,000 articles from Wikipedia, especially among medical domains. We use open source software disclosed in GitHub for the training of Brown Clustering[5]. We use word2vec for the training of word representation[6].

3.2 Experimental Results

3.2.1 Extraction of Complaint and Diagnosis

We show the results of extraction task in the Table 3 and the Table 4.

To examine efficiency of unsupervised features, we also show results of extraction performance without unsupervised features. From the results, both unsupervised features, ones from brown clustering and ones from word2vec, contribute to improvement of performance.

3.2.2 Assignment of Modality Tag

We show results of assignment of modality tags both to gold data and automatically extracted data in Table 5.

Though we take the rule-based simple approach to assign modality tags, we can achieve intermediately good performance. We guess this is because medical records are written almost in the same format, so we can cover most of cases without complex rules.

However, among the four modality tags, "suspicion" is relatively low in both precision and recall. A likely cause of this fact may be there exist many case of "suspicion" which require understanding the whole sentence to predict a modality tag correctly and is difficult only with surrounding words. For example, in the following sentence,

> <c>食道静脈瘤</c>の可能性を考えたが、複数回にわたって施行した上部消化管内視鏡検査にて明らかな<c>静脈瘤所見</c>を認めなかった。

, only with first part of whole sentence, we may be likely to misunderstand the modality tag of "食道静脈瘤" is "suspicion", but with whole sentence, we can predict its correct modality, "negation".

[5]https://github.com/percyliang/brown-cluster/
[6]https://code.google.com/p/word2vec/

Table 4: Extraction results of time expressions

	Precision	*Recall*	F_1
without unsupervised	91.51 %	78.86 %	84.72 %
with brown	91.25 %	79.13 %	84.56 %
with word2vec	90.12 %	79.13 %	84.27 %
with brown ans word2vec	90.43 %	79.40 %	84.56 %

Table 5: Assignment results of modality tags

Tag	*Precision*	*Recall*	F_1
positive(gold)	92.06 %	94.97 %	93.49 %
positive(auto)	79.79 %	71.53 %	75.43 %
negation(gold)	93.82 %	86.10 %	89.79 %
negation(auto)	87.14 %	65.32 %	74.67 %
suspicion(gold)	65.52 %	69.09 %	67.26 %
suspicion(auto)	70.45 %	56.36 %	62.63 %
family(gold)	73.68 %	97.67 %	84.00 %
family(auto)	71.70 %	56.36 %	62.63 %

Table 6: Assignment results of ICD-10 tags

	Accuracy(*gold*)	*Accuracy*(*auto*)
all	69.99 %	56.22 %
DicScore-only	66.10 %	53.04 %

3.2.3 Assignment of ICD-10 Tag

We show the results of assignment of ICD-10 tags both to gold data and automatically extracted data in the Table 6.

To examine efficiency of context information for assigning correct ICD-10 tags, we also show results of tagging performance using only one subscore, DicScore. When using all subscore, we set all the weights of subscores to 0.25. From the results, context information improve the system performance.

4. CONCLUSIONS

In this paper, we have developed a system for two tasks of MedNLP2. For the extraction task, we used CRF based NER method and incorporated unsupervised features into that of CRF. And we show such features improve system performance.

5. REFERENCES

[1] E. Aramaki, M. Morita, Y. Kano, and T. Ohkuma. Overview of the NTCIR-11 MedNLP-2 Task. In

Proceedings of the 11th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, 2014.

[2] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[3] T. Koo, X. Carreras, and M. Collins. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, pages 595–603, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[4] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289, 2001.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] S. Miller, J. Guinness, and A. Zamanian. Name tagging with word clusters and discriminative training. In *HLT-NAACL*, volume 4, pages 337–342. Citeseer, 2004.

[7] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[8] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics, 2004.