Developing ML-based Systems to Extract Medical Information from Japanese Medical History Summaries

Shohei Higashiyama **Kazuhiro Seki** **Kuniaki Uehara**
Graduate School of System Informatics, Kobe University
{higashiyama@ai.cs., seki@cs., uehara@}kobe-u.ac.jp

Abstract

With the increase of the number of medical records written in an electronic format, natural language processing techniques in the medical domain have become more and more important. For the purpose of the development and evaluation of machine learning-based systems to extract medical information, we recently participated in the NTCIR-10 MedNLP task. The task focused on Japanese medical records and aimed at evaluating different information extraction techniques on the common data set provided by the organizers. We implemented our baseline system based on structured perceptron and have developed its extensions. In this paper, we describe our systems and report on the evaluation of and the analysis on their performance.

1 Introduction

In recent years, medical records have been increasingly written in an electronic format, which leads to a growing need for natural language processing (NLP) techniques in the medical domain. Specifically, information extraction (IE) techniques, such as named entity recognition (NER), are crucial as they serve as the basis of more intellectual and/or application-oriented tasks, including information retrieval and question answering.

Given the background, the NTCIR-10 MedNLP task (Morita et al., 2013) was recently held as a shared task to foster the NLP research for medical texts, specifically targeting Japanese. The participants of the task were provided with an annotated corpus consisting of 50 fictional medical history summary reports. The intended task was a type of NER and required the participants to identify patients' personal and medical information from the reports.

For the MedNLP task, we took part in the de-identification subtask and the complaint and diagnosis subtask summarized shortly by adapting an NER model to the medical domain. The model is based on structured perceptron (Collins, 2002) and was previously developed for the biomedical domain (Higashiyama et al., to appear).

This paper reports on the results of the structured perceptron-based model for the MedNLP task and presents their analysis. Additionally, conditional random fields (CRFs) (Lafferty et al., 2001), a popular model adopted by many participants of the task, are applied for comparison.

2 NTCIR-10 MedNLP Task

2.1 Dataset

The MedNLP task organizers prepared medical history summary reports of fictional patients written by physicians. The medical records consist of 50 documents and include 3,365 sentences. Two thirds of them (2,244 sentences) and remaining one thirds (1,121 sentences) are respectively provided as the sample set and the test set.

The sample set is annotated with personal and medical information about patients. The personal information includes *age*, *person's name*, *sex*, *time*, *hospital name* and *location* [1]. The medical information indicates *complaint and diagnosis* with a modality attribute that is taken to have four values: *positive*, *negation*, *suspicion* and *family*. Suppose that there is a mention of a particular symptom about a patient. Then, the expression representing the symptom would be annotated with the attribute value of *positive* if the patient has the symptom, *negation* if the patient does not have the symptom, *suspicion* if the patient is suspected of the symptom and *family* if a member

[1] A half of these tags in fact rarely appear in the sample set. The numbers of *persons' name*, *sex* and *location* tags are less than five while the numbers of remaining tags are respectively more than 50.

International Joint Conference on Natural Language Processing Workshop on Natural Language Processing for Medical and Healthcare Fields, pages 14–21,
Nagoya, Japan, 14-18 October 2013.

of the patient's family has a history of the symptom.

2.2 Task Description and Formulation

The NTCIR-10 MedNLP task mainly consisted of the following two subtasks.

1. De-identification (DI) task: identifying personal information about patients, such as ages and hospital names.
2. Complaint and diagnosis (CD) task: extracting patients' complaint and diagnosis by physicians and determining their modality status for the patients.

The performance of participants' systems for both subtasks was measured by the F-measure ($\beta = 1$), which is the harmonic mean of precision and recall.

These subtasks can be seen as NER tasks recognizing named entities and classifying them into predefined semantic classes. Named entities indicate particular expressions to be extracted, which are represented by proper nouns and technical terms. As for the DI task, this subtask can be formulated as classifying each word in a sentence into one of the labels consisting of a semantic class (e.g. *age*) and a chunk IOB tag, where I, O, and B respectively denote the inside, outside, and beginning of an entity. For example, if a word "64" in "64 years old" is assigned with a label "`B-age`", it means that the "64" is recognized as the beginning of an entity with a semantic class *age*. The CD task can be formulated likewise by regarding a *complaint and diagnosis* tag with a modality attribute *x* as a class *c-x*.

3 Description of Baseline System

For the MedNLP task, we applied structured perceptron (Collins, 2002), which is an online algorithm. Despite its simplicity, structured perceptron is reported to have performance that closely approximates that of support vector machines (SVMs), which has been applied successfully to various classification problems. In addition, we introduced a cost function into the perceptron framework to achieve higher performance, and used the model as our baseline system. The cost function is a type of cost-sensitive learning method which lowers the expected cost of misclassification.

In the following two sections, we describe the learning and prediction algorithms on an ordinary and a cost-sensitive version of structured perceptron.

3.1 Structured Perceptron

Let $\mathcal{X}$ be a set of instances and let $\mathcal{Y}_{\boldsymbol{x}}$ be a set of possible label sequences for an instance $\boldsymbol{x} \in \mathcal{X}$, where $\boldsymbol{x}$ denotes a token sequence (i.e., sentence) in the training or test data. Additionally, $\boldsymbol{y} \in \mathcal{Y}_{\boldsymbol{x}}$ denotes a possible label sequence of $\boldsymbol{x}$. $\mathcal{Y}_{\boldsymbol{x}}$ is equivalent to the direct product $\mathcal{L}^n$, where n is the length of $\boldsymbol{x}$ and $\mathcal{L}$ is a set of labels that includes labels such as `B-age` and `O`.

Learning on structured perceptron can be regarded as finding the weight vector $\boldsymbol{w} \in \mathbb{R}^d$ so that the discriminative function f predicts the correct label sequences of instances. The discriminative function $f : \mathcal{X} \rightarrow \mathcal{Y}$ is defined as

$$f(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{w}, \Phi(\boldsymbol{x}, \boldsymbol{y}) \rangle \ ,$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product of two arguments and $\Phi(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d$ is the feature vector of $\boldsymbol{x}$ and $\boldsymbol{y}$.

The prediction $\hat{\boldsymbol{y}}$ for $\boldsymbol{x}$ is the output of f as in

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \ f(\boldsymbol{x}, \boldsymbol{y}) \ . \tag{1}$$

During learning on the training data, we receive a training instance $\boldsymbol{x}_t$ on each round t, and output its prediction $\hat{\boldsymbol{y}}_t$ by Eq. (1). Then, $\boldsymbol{w}$ is updated by Eq. (2) if the prediction $\hat{\boldsymbol{y}}_t$ differs from the correct label sequence $\boldsymbol{y}_t$:

$$\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t + \Phi(\boldsymbol{x}_t, \boldsymbol{y}_t) - \Phi(\boldsymbol{x}_t, \hat{\boldsymbol{y}}_t) \ , \tag{2}$$

where $\boldsymbol{w}^t$ is the weight vector on round t. Learning is iterated through all the training instances T times. Label sequences of test instances can be predicted by Eq. (1) in the same manner as training instances.

3.2 Cost-Sensitive Structured Perceptron

In addition to use of structured perceptron, we exploited information on distance between a correct and a candidate label sequence of each training instance during learning based on cost-sensitive learning of an ML framework for lowering misclassification cost. Cost-sensitive approaches were, for example, applied to semantic role labeling on the study by Johansson and Nugues (2008), which used passive-aggressive (Crammer et al., 2006), and to part-of-speech tagging on that by Song et al. (2012), which used multiclass SVMs.

The cost-sensitive learning algorithm on structured perceptron updates the weight vector $\boldsymbol{w}$ using $\tilde{\boldsymbol{y}}_t$ defined below instead of $\hat{\boldsymbol{y}}_t$ in Eq. (2).

$$\tilde{\boldsymbol{y}}_t = \underset{\boldsymbol{y} \in \mathcal{Y}}{\operatorname{argmax}} \; f(\boldsymbol{x}_t, \boldsymbol{y}) + \alpha\rho(\boldsymbol{y}_t, \boldsymbol{y}) \qquad (3)$$

In Eq. (3), $\rho : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{N} \cup \{0\}$ is the cost function which returns a larger value for larger distance between $\boldsymbol{y}_t$ and $\boldsymbol{y}$, and α is a parameter that is taken to have a positive real number. Here, we define the cost function ρ as

$$\rho(\boldsymbol{y}_1, \boldsymbol{y}_2) = \sum_{i=1}^{|\boldsymbol{y}_1|} \delta(y_1^{(i)}, y_2^{(i)}) \; ,$$

where $|\boldsymbol{y}|$ denotes the length of the vector $\boldsymbol{y}$ and the function $\delta : \mathcal{L} \rightarrow \{0, 1\}$ is defined as

$$\delta(y_1, y_2) = \begin{cases} 0 & (y_1 = y_2) \\ 1 & (y_1 \neq y_2) \end{cases} \; .$$

In the cost-sensitive learning framework, the weight vector can be updated to the reserve margin $\alpha\rho(\boldsymbol{y}_t, \tilde{\boldsymbol{y}}_t)$ using $\tilde{\boldsymbol{y}}_t$ instead of $\hat{\boldsymbol{y}}_t$. That is,

$$\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t + \Phi(\boldsymbol{x}_t, \boldsymbol{y}_t) - \Phi(\boldsymbol{x}_t, \tilde{\boldsymbol{y}}_t) \; .$$

3.3 Features

The following features were used in the experiments for both subtasks:

- tokens in the window of size two around the current token and
- the part-of-speech (POS) tag, the subtype of POS tag, the lemma and the pronunciation of the current token.

We applied the Japanese morphological analyzer MeCab (Kudo et al., 2004) (version 0.996) with the IPA dictionary [2] (version 2.7.0) to word segmentation and used the output of MeCab for each sentence as the latter features.

4 Evaluation and Discussion

4.1 Evaluation of Baseline System

Parameter Setting

We determined the optimal value of parameter α in Eq. (3) and the number of iterations T using the sample set as follows.

1. We used 90% of the sample set as the learning set and the remaining 10% as the validation set.
2. Varying the value of α and increasing the value of T, we learned a model for particular α and T on the learning set and evaluated it on the validation set.
3. Values of α and T that yielded the best F-measure were regarded as optimal.

Consequently, the optimal α and the number of iterations T were respectively set to 30 and 20. By use of the cost function, both precision and recall on the validation set improved by around four points, compared with the method without the function. We used these values for producing our official runs on the test set submitted to the MedNLP organizers.

Results on Test Set

Table 1 shows the performance of our system using the test set. Table 1 (a) shows the overall performance and Table 1 (b) shows the performance of each entity class. The performance was measured by precision, recall, the F-measure ($\beta = 1$), and accuracy. Recall was always lower than precision for all classes of both tasks, and especially lower in the family and the suspicion classes, which led to degraded F-scores. In addition, the lower performance for the total on the CD task than 2-way indicate difficulty of modality classification.

4.2 Error Analysis of Baseline System

For error analysis, we evaluated our system on the sample set using a five-fold cross-validation method. Subsequently, we analyzed the results on the validation sets for five iterations. As compared with the performance on the test set, the performance on the validation sets was worse by several points for the CD task, and almost equivalent for the DI task. The reason of the former is the fewer training instances, and that of the latter was that the targeted entities for the DI task have much in common as we discuss shortly.

Analysis on De-identification Task

Despite the smaller number of positive instances of entity classes for the DI task than that for the CD task, the performance for the former classes was relatively high on the whole. The

[2] http://chasen.naist.jp/stable/ipadic/

Table 1: Results of both de-identification (DI) task and complaint and diagnosis (CD) task on the test set. The “2-way” is a result of recognition of complaint/diagnosis or not. The “total” is a result including classification of modality classes. P, R, F and A indicate precision, recall, F-measure ($\beta = 1$), and accuracy, respectively.

(a) Overall performance on test set.

subtask	P	R	F	A
DI	82.09	76.39	79.14	99.38
CD (2-way)	82.37	72.29	77.00	95.48
CD (total)	74.72	65.58	69.86	94.50

(b) Performance of each entity class on test set.

subtask	class	P	R	F
DI	age	80.65	78.12	79.37
	time	84.56	81.56	83.03
	hospital	72.73	63.16	67.61
CD	c-positive	72.87	67.04	69.83
	c-negation	82.35	68.02	74.50
	c-suspicion	55.00	36.67	44.00
	c-family	66.67	36.36	47.06

reason is that a large portion of these entities fit typical patterns. For example, over 70 percents of the instances of the age class in the sample set match a simple regular expression, “[１-９]?[０-９]歳[時頃(ごろ)]?[ー〜(から)(より)(まで)]?” (“[(from)(to)]?(about)?[1-9]?[0-9](years old)”). For misclassified cases, we found two major types of errors across all classes in this task: (1) recognition of incorrect boundaries of entities; and (2) undetection of entities (false negatives).

Specifically, the most frequent errors on the age class was found to be the first type, such as “４７歳” (47 years old) for a correct boundary “２７歳〜４７歳” (27 to 47 years old) and “１０代’ ’ (10s) for “１０代前半” (early 10s). Because words or expressions co-occurring with or including ages themselves as numerical values are limited, it may be effective to fix system outputs by rule-based post-processing.

On the other hand, most errors on the hospital class was the second type. For example, entities such as “同院” (the hospital) and “総合病院” (general hospital) were often undetected. The reason is that these words rarely appeared in the sample set in contrast to frequently appearing words, such as “当院” (our hospital) and “近医” (local hospital), which were correctly detected.

As for the time class, both types of errors were often observed. A large portion of boundary errors were recognizing narrower scopes for entities than their correct ones, e.g., “１０月２９日” (October 29) against a correct boundary “１０月２９日夕刻まで” (until the evening on October 29). Many false negatives were found to be expressions using slashes, such as “７／２０”. More formal expressions, such as “７月２０日” (July 20), are more often used in the sample set. For dealing with the errors of the hospital and the former type of the time, constructing and using dictionaries composed of expressions which often constitute or co-occur with those type of entities may be beneficial. For the latter type of the time, rule-based post-processing may be effective, similarly to the age class.

Analysis on Complaint and Diagnosis Task

In addition to the two types of errors discussed for the previous task, there were mainly two types of errors in detecting complaint entities: (3) misclassification of the modality classes; and (4) misdetection of non-entities (false positive).

The most frequent errors were undetection of entities through all classes, and this type of errors frequently observed in the positive and the negation classes. In order to reduce such false negatives and improve recall, we plan to use external knowledge resources such as public dictionaries in future work.

The second most frequent errors were misclassification of entities whose boundaries were correctly recognized. They accounted for a major portion of errors on the three classes except the positive class. Especially, the low performance on the family and the suspicion classes was due to misclassification in addition to undetection which occur similarly as the other modality classes. For these modality classes, it was found that there exist typical keywords which often co-occur with entities. Entities of the family class co-occur with family relation names. In particular, most of them in the sample set co-occur in itemized sentences, such as “父：心筋梗塞” (Father: cardiac infarction). Entities of the negative class and the suspicion class occur ahead of expressions of negations, such as “なし” (be absent), and expressions of uncertainty, such as “考えられる” (be concerned), “疑いがある” (be suspected), and “可能性がある”

Input: 薬剤性肺炎 の可能性を 考え

(consider the possibility of drug-induced pneumonia)

⇓

Output: 薬剤/性/肺炎/の/可能/性/を /考え

Figure 1: An example of a parsed sentence including a suspicion entity by MeCab. The underlined part in the input sentence indicates an entity annotated with the suspicion class. The parts segmented by slashes in the output indicate words segmented by the tagger.

(be possible).

However, our system could not exploit these keywords because of the limited window size of two around the current token, and entities often occur at a distance from keywords, especially in the suspicion class. For example, Figure 1 shows an input sentence containing a suspicion entity "薬剤性肺炎" (drug-induced pneumonia) and its parsed output by the MeCab morphological analyzer. Two out of three tokens constituting the entity (i.e., "薬剤" (drug) and "性" (-induced)) are more than two tokens away from the uncertainty keywords (i.e., "可能", "性" (possibility) and "考え" (concern)). To improve classification performance for modality classes, specifically recall, it is crucial to increase the window size to, for example, sentence boundaries. Alternatively, it may be effective to take advantage of dependency parsing.

The other causes of the observed errors were incorrect boundary errors and misdetection errors. The reasons require a further study.

4.3 Post-submission Experiments

To achieve higher performance, we have developed our medical information extracting systems also after implemented and submitted our baseline system. Specifically, we used CRFs as an alternative ML algorithm to structured perceptron. Moreover, we introduced domain-specific terms in medical fields into the default dictionary of the morphological analyzer.

In the following subsections, we describe the above conversion and extension from the baseline system and the experiments on those.

Alternative ML Algorithm: Conditional Random Fields

To improve the performance of the baseline system, we employed CRFs (Lafferty et al., 2001) as an alternative ML algorithm. CRFs are extensions of maximum entropy to structured prediction. Additionally, the algorithm has been widely applied to both NER (McCallum and Wei, 2003; Settles, 2004; Finkel et al., 2005) and other NLP tasks, such as part-of-speech tagging (Lafferty et al., 2001), noun phrase chunking (Sha and Pereira, 2003) and morphological analysis (Kudo et al., 2004). Particularly, we utilized CRF++ [3] , which is an open source implementation of CRFs and allows easy customizability of features by describing in the feature template file. We used the same features as those in the baseline system.

Use of Medical Lexicon

When analyzing texts in a specific domain, morphological taggers with default dictionary in general domain often unsuccessfully analyze sentences that contain domain-specific terms. Consequently, they make errors attributed to unknown words in word segmentation or other processing such as POS tagging and pronunciation prediction. These errors can be negatively affect on NER that is a higher-level task than morphological analysis. Then, we enhanced the regulation dictionary of MeCab by addition of domain-specific terminology from life science dictionary (LSD) (Kaneko et al., 2003), which consists of a broad range of life science terms such as names of anatomical concepts, biological organisms, diseases and symptoms.

By addition of a domain-specific dictionary, not only the morphological tagger can achieve tagging error reduction, but also finely segmented morphemes that are component of domain-specific terms tend to be segmented more coarsely because expressions contained in the dictionary are more frequently regarded as one morpheme. For instance, "薬剤性肺炎" (drug-induced pneumonia) is segmented into "薬剤" (drug), "性" (-induced) and "肺炎" (pneumonia) before the addition of terms in LSD to the original dictionary and into "薬剤性肺炎" after the addition. Similarly, "ＰＩＰ関節裂隙狭小化" (joint space narrowing at the proximal interphalangeal (PIP) joints) is seg-

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

Table 2: Comparison of systems based on two algorithms with or without the enhanced dictionary using the sample set. SP denotes cost-sensitive structured perceptron and dic indicates using the enhanced dictionary.

(a) Performance for de-identification (DI) task.

system	P	R	F
SP	82.72	86.97	84.79
SP+dic	84.06	86.02	85.03
CRFs	91.01	82.32	86.45
CRFs+dic	89.26	82.61	85.81

(b) Performance for complaint and diagnosis (CD) task.

system	P	R	F
SP	66.29	72.76	69.37
SP+dic	65.05	77.02	70.53
CRFs	78.85	68.26	73.17
CRFs+dic	81.91	66.06	73.14

mented into “ＰＩＰ”, “関節” (joints), “裂隙” (space), “狭小” (narrow) and “化” (-ing) before and into “ＰＩＰ関節”, “裂隙” and “狭小化” after. The latter segmentation can be beneficial for exploiting information about strings distant from the token in question in the case of fixed window size around the token. Therefore, in addition to reduction errors in morphological analysis, NER systems can obtain benefit from coarse segmentation, by use of the tagger with the richer language resource.

Results and Discussion

To measure the performance of CRFs, which we used as an alternative algorithm to structured perceptron, and to evaluate the effectiveness of the enhanced dictionary, we compared four systems based on the two algorithms with or without the enhanced dictionary. Table 2 shows the results on the sample set using five-fold cross-validation. Table 2 (a) and (b) show the overall performance for the DI task and the CD task, respectively. For both subtasks, while recall of structured perceptron was higher than that of CRFs, CRFs outperformed structured perceptron by around 10 points in terms of precision. Additionally, CRFs also outperformed by a few points in terms of F-measure.

The both algorithms consider the overall sequence of tokens when predicting their labels, but they defer in the respective training methods. More precisely, structured perceptron minimizes the loss defined by the difference between correct and predicted label sequences. This process can be regarded as the training by a simple (sub) gradient method with fixed step size, which is a first-order gradient method. On the other hand, CRFs are trained by maximizing the log-likelihood of a given training set. The implementation of CRFs used in our experiment was based on limited-memory BFGS (L-BFGS), which is a second-order gradient method. We believe that the more sophisticated optimization algorithm of CRFs resulted in the higher performance. In fact, Sha and Pereira (2003) empirically showed that CRFs based on second-order methods, such as L-BFGS and conjugate gradient, outperformed structured perceptron on a noun phrase chunking task.

Contrary to our expectation, use of the morphological analyzer with enhanced dictionary had a little or negative effect for the performance of both algorithm and for both subtasks, except that recall of structured perceptron for the CD task was improved. We believe that this result was due to loss of common characteristics among segmented tokens. Focusing on the complaint entity “薬剤性肺炎” (drug-induced pneumonia), various expressions occur in the sample set preceding “肺炎” (pneumonia), e.g. “細菌性” (bacterial), “間質性” (interstitial), “器質化” (organizing), “強膜炎” (pleuritic) and “ニューモシスチス” (Pneumocystis), in addition to “薬剤性” (drug-induced). Furthermore, there are variety of entities containing expressions that co-occur with “肺炎”, e.g. “薬剤性肺障害” (drug-induced pulmonary disorder), “細菌感染” (bacteria infection), “器質化血栓” (organizing thrombus), “胸膜炎” (pleuritis) and “ニューモシスチス・カリニ” (Pneumocystis carinii). As we discussed previously, morphemes tend to be segmented more coarsely after augmented terms in the dictionary of the morphological analyzer. Then, entities enumerated above became to be recognized as distinct tokens without common characteristics, by segmented to one or a little larger numbers of morphemes. We consider that this affected the performance negatively and disturbed learning of classifiers.

To fix this problem, it may be effective to use prefix and suffix features derived form expressions that are often contained by or co-occurred with entities. After the processing, classifiers may come to be able to exploit information about strings that are distant from the current token and to obtain benefit by reduction errors in morphological analysis.

5 Related Work

To the NTCIR-10 MedNLP task, both rule-based and ML-based approaches were applied among the participants. Almost all systems for the DI task and over a half of all systems for the CD task were based on ML, especially supervised learning. It should be note that greater part of systems that achieved higher performance were based on ML and moreover a large portion of them employed CRFs. Specifically, systems of the top three teams for the CD task and of the second and third ranked teams for the DI task were based on CRFs. By contrast, the system that had the highest performance for the DI task was a rule-based approach. As other ML-based approaches than CRFs, structured perceptron, language models and bootstrapping were applied.

As to features, general-purpose NER features were widely applied, such as word surface (token) and POS features. Pronunciation and character type features were also used. Besides, domain-specific features including dictionary matching features or heuristic features of data-specific expressions were used. These features are derived from medical knowledge resources such as LSD and MEDIS standard masters [4], or manually constructed lexica consisting of expressions that are specific to each entity class. Among the features incorporated in the ML-based systems, particularly, those that achieved higher performance, dictionary or heuristic features provided high benefit for their performance. Specifically, Laquerre et al. (Laquerre and Malon, 2013) reported that heuristic features for the DI task improved the F-measure by around three points and heuristic and dictionary features for the CD task improved by around 4.5 points. Miura et al. (Miura et al., 2013) also reported that dictionary features for the CD task improved the F-measure by around two points.

Nevertheless the limited size of the dataset, the overall performance for the subtasks of the top systems were high: they achieved over 90% and 75% F-measure for the DI task and the CD task, respectively. As regards the performance for each entity type, that for the family entities were over 80% F-measure, which is highest of all entity types for the CD task, in spite of smaller numbers of entities in the sample set. This is due to the features for the family class such as family names could capture the characteristics of this entities well. By contrast, the F-measure was only around 50% for the suspicion entities, which occurred less frequently similarly to the family entities. This suggests that the suspicious expressions used for extracting the suspicion entities (e.g. "疑い" (suspicious) and "可能性" (possibility)) were insufficient or there exists other reasons that make it difficult to identify this type of entities.

6 Conclusions

This paper described our systems to extract personal and medical information from medical texts. We implemented a simple system based on structured perceptron as a first step toward more effective Japanese medical text processing systems, and extended it to systems based on another machine learning algorithm and on a morphological analyzer with a domain-specific dictionary. Moreover, we analyzed its performance and issues for achieving the goal. The result on the MedNLP dataset indicates that classification of medical entities into their modality classes, especially the suspicion class, is difficult. However, our analysis revealed that the terms and expressions in medical texts have useful patterns and characteristics that could be exploited for more accurate extraction.

Although it found that it was not very effective to use output of the morphological analyzer with domain-specific dictionary, we are aiming to use knowledge resources in more effective ways, e.g. incorporating dictionary features into classifiers. Additionally, we plan to explore more useful features such as suffix and prefix features for development of more advanced systems.

References

M. Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP)*, pages 1–8.

K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. 2006. Online passive-aggressive algorithms. *The Journal of machine learning research (JMLR)*, 7:551–585.

J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information

[4] http://www.medis.or.jp

extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics (ACL)*, pages 363–370.

S. Higashiyama, M. Blondel, K. Seki, and K. Uehara. (to appear). Named entity recognition exploiting category hierarchy using structured perceptron. *IPSJ Transactions on mathematical modeling and its applications*.

R. Johansson and P. Nugues. 2008. Dependency-based semantic role labeling of propbank. In *Proceedings of the 2008 conference on empirical methods in natural language processing (EMNLP)*, pages 69–78.

S Kaneko, N Fujita, Y Ugawa, T Kawamoto, H Takeuchi, M Takekoshi, and H Ohtake. 2003. Life science dictionary: a versatile electronic database of medical and biological terms". *Dictionaries and Language Learning: How can Dictionaries Help Human & Machine Learning?*, pages 434–439.

T Kudo, K Yamamoto, and Y Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 conference on empirical methods in natural language processing (EMNLP)*, pages 230–237.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th international conference on machine learning (ICML)*, pages 282–289.

P. F Laquerre and C Malon. 2013. NECLA at the medical natural language processing pilot task (MedNLP). In *Proceedings of the 10th NTCIR conference*, pages 725–727.

A. McCallum and L. Wei. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the 7th conference on natural language learning (CoNLL-2003)*, pages 188–191.

Y Miura, T Ohkuma, H Masuichi, E Yamada, E Aramaki, and K Ohe. 2013. UT-FX at NTCIR-10 MedNLP: incorporating medical knowledge to enhance medical information extraction. In *Proceedings of the 10th NTCIR conference*, pages 728–731.

M. Morita, Y. Kano, T. Ohkuma, M. Miyabe, and E. Aramaki. 2013. Overview of the NTCIR-10 MedNLP task. In *Proceedings of the 10th NTCIR conference*, pages 696–701.

B. Settles. 2004. 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 (JNLPBA)*, pages 104–107.

F. Sha and F. Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language (HLT-NAACL)*, pages 213–220.

Hyun-Je Song, Jeong-Woo Son, Tae-Gil Noh, Seong-Bae Park, and Sang-Jo Lee. 2012. A cost sensitive part-of-speech tagging: differentiating serious errors from minor errors. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1025–1034.