HCRL at NTCIR-11 MedNLP-2 Task

Osamu Imaichi
Hitachi, Ltd., Central Research Laboratory
osamu.imaichi.xc@hitachi.com

Masakazu Fujio
Hitachi, Ltd., Central Research Laboratory
masakazu.fujio.kz@hitachi.com

Toshihiko Yanase
Hitachi, Ltd., Central Research Laboratory
toshihiko.yanase.gm@hitachi.com

Yoshiki Niwa
Hitachi, Ltd., Central Research Laboratory
yoshiki.niwa.tx@hitachi.com

ABSTRACT
This year's MedNLP-2 [1] has two tasks: Extraction task (Task 1) and Normalization task (Task 2). We tested both machine learning based methods and an ad-hoc rule-based method for the two tasks. For the Extraction Task, a two-stage approach (first, the machine learning based method is applied to identify c tags, and second, the rule-based method is applied to modality features) obtained higher results. For the Normalization Task, the machine learning based method obtained higher results for training data, but the simple pattern-matching method obtained higher results for test data.

Team Name
HCRL

Subtasks
Task 1 (Extraction Task)

Task 2 (Normalization Task)

Keywords
sequential labeling, CRF, SVM, unsupervised feature learning.

1. INTRODUCTION
Machine learning based and rule-based methods are the two main approaches for extracting useful information from natural language texts. To clarify their pros and cons, we applied both approaches to this year's MedNLP-2 tasks: Extraction Task (Task 1) and Normalization Task (Task 2).

2. Task 1 (Extraction Task)
2.1 Approach
We formalized the information extraction task as a sequential labeling problem. We adopted a conditional random field (CRF) [2] as the learning algorithm. We used CRFsuite [3], which is an implementation of first order linear chain CRF. We also used MeCab [4] as a Japanese morphological analyzer and CaboCha [5] as a Japanese dependency parser.

2.2 Basic Features
We used the following features to capture the characteristics of the token: surface, part-of-speech, and dictionary matching. 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 version of Wikipedia, Chinese characters related to diseases, expressions of suspicion, and negative expressions. These dictionaries were created on the basis of the rules from training data and Wikipedia. We also used a sentence feature based on the field name to which the sentence belongs.

2.3 Unsupervised Feature Learning
In addition to the basic features, we used clustering-based word features [6] to estimate clusters of words that appear only in test data. These clusters can be learned from unlabeled data by using Brown's algorithm [7], 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 finished.

We examined two kinds of Brown features: those created from training and test data related to the MedNLP-2 task (1,000 categories) and those created from Japanese Wikipedia (1000 categories).

2.4 Rule-based method
We applied the machine learning method to identify c tags and then the rule-based method to identify time tags and modality tags. The patterns are developed for NTCIR-10 MedNLP Task [8]. The details are explained in our previous work [9].

3. Task 2 (Normalization Task)
3.1 Approach
We formalized the normalization task as a chunk classification problem, and an information retrieval problem. We adopted a support vector machine (SVM) as the learning algorithm. We used libsvm [10], which is an implementation of the SVM. We also tested the simple pattern-matching method as a baseline.

3.2 Basic Features
We used the following features to capture the characteristics of the chunk: surface, part-of-speeches of the tokens in each chunk or neighbors of each chunk, and dictionary matching. The surfaces and part-of-speeches were converted into the feature vectors, which are the same as those for Task 1. For the dictionaries, we used the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD10). The first two to three layers of the matched ICD10 entries were used for the dictionary matching features.

3.3 Unsupervised Feature Learning

In addition to the basic features, we used clustering-based word features, which are the same as those for Task 1. In Task 2, we also used a neural-network based word vector representation [11]. We used word2vec [12], which is an implementation of the neural network based on the skip-gram model. We examined 50 to 800 dimensions of word2vec features learned from MedNLP-2 training, test data, and Japanese Wikipedia.

4. RESULTS

4.1 Task 1 (Extraction Task)

The results for Task 1 are shown in Tables 1, 2, and 3. All values in the tables are F1 values.

Table 1: Results for training data (c tags and modality)

	c	c+ family	c+ negation	c+ suspicion
HCRL-1 (ML only)	85.36	78.12	78.10	46.49
HCRL-2 (ML + Rule)	85.26	88.89	80.79	72.75
HCRL-3 (ML only)	85.94	75.97	78.40	50.62

Table 2: Results for test data (F-value)

	t	c
HCRL-1 (ML only)	85.68	82.61
HCRL-2 (ML + Rule)	87.07	82.54
HCRL-3 (ML only)	85.39	83.33

Table 3: Results for test data (NE + modality of c tag)

	NE+ positive	NE+ family	NE+ negation	NE+ suspicion
HCRL-1 (ML only)	74.06	86.84	72.87	48.78
HCRL-2 (ML + Rule)	75.94	86.08	76.67	60.50
HCRL-3 (ML only)	75.74	89.74	73.59	44.44

HCRL-3 used ICD10 entry names as an additional dictionary.

4.2 Task 2 (Normalization Task)

Table 4: Results for training data and test data

	Training data	Test data
HCRL-1 (SVM)	75.7	62.3
HCRL-2 (SVM + w2v)	75.8	63.5
HCRL-3 (SVM + Brown)	72.9	60.5
HCRL-4 (pattern match)	70.7	72.2

The results of HCRL-4 are not submitted because the score for training data is relatively low.

5. Conclusion

For the Extraction Task, a two-stage approach (first, the machine learning based method is applied to identify c tags, and second, the rule-based method is applied to modality features) obtained higher results. For the Normalization Task, the machine learning based method obtained higher results for training data, but the simple pattern-matching method obtained higher results for test data.

6. REFERENCES

[1] Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T. 2014. Overview of the NTCIR-11 MedNLP-2 Task. In P*roceedings of the 11th 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] http://www.chokkan.org/software/crfsuite/

[4] https://code.google.com/p/mecab/

[5] 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.

[6] 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.

[7] 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.

[8] 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.*

[9] Imaichi, O., Yanase, T., and Niwa, Y. 2013. HCRL at NTCIR-10 MedNLP Task, In *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*.

[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[11] Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient Estimation of Word Representation in Vector Space, In *Proceedings of workshop at ICLR*.

[12] https://code.google.com/p/word2vec/