past State Examinations of Medical Doctors from the website of the Japanese Ministry of Health, Labor and Welfare. We designed CRF feature functions based on characters and the character types.

Character-based CRF models assume that the classes of characters in a sentence have the Markov property. Namely, the class of a target character in a sentence is assumed to be dependent on the classes of n characters occurring before the target character. We trained n-order CRF models ($n \in \{1, 2, 3\}$) , and then estimated the medical terms included in the test dataset of Task-1, *test.xml*, by using those trained CRF models with the list of medical terms collected from the MEDIS Standard Master and additional dictionaries such as the Life Science Dictionary (LSD) and the T-terminology Dictionary (TDIC). We solved the medical term estimation as an integer linear programming (ILP) problem.

2.2 Modality Classification

For the MedNLP-2 task, four modality attributes of medical terms, *positive*, *negative*, *suspicion*, and *family*, were defined by the task organizers. The medical terms included in the training data were annotated with *positive*, *negative*, or *suspicion*, and thus we can regard the three modality attributes as exclusive events. Therefore, we designed a three-class single-labeled classifier that assigned medical terms to one of the three modality attributes. We applied TinySVM[6] to the classification problem by using a one-against-rest technique. We also designed a binary classifier that either assigned medical terms to the modality attribute of *family* or not, by using a TinySVM. We employed a second-order polynomial kernel for all the TinySVMs and trained those TinySVMs by using the medical document set, *ntcir11_mednlp_mednlp2-train_v0.xml*.

We designed the feature vectors of the medical terms by using three characters and two words occurring before and after the medical terms. We also used all the words included in the same sentence segment as the medical terms and all the words dependent on the medical terms. Moreover, we utilized the part-of-speech and clustering results for those words related to the medical terms.

For the feature design, we analyzed the word dependence of sentences with a dependency structure analyzer [5]. We employed the k-means++ method [2] to obtain the clustering results of words, whose feature vectors were obtained by applying an extended skip-gram model [4] to a Japanese Web N-gram corpus[7]. We expect the word clustering results to represent the similarities of words occurring around different medical terms, and assume that using the word clustering results is useful for estimating the modality attributes of medical terms co-occurring with similar words.

3. APPROACH TO TASK-2

We applied a one-against-rest method to the ICD-code classification of medical terms, and employed a logistic regression model for each ICD-code. The logistic regression model for the kth ICD-code estimates the relevant and irrelevant probabilities that show whether or not a medical term should be assigned to the ICD-code. The ICD-code to which a medical term should be assigned is estimated as the ICD-code maximizing the relevant probabilities of the medical term computed by using those logistic regression models. We designed the feature vectors of the medical terms by using the frequencies of character-based and character-type-based N-grams ($N \in \{1, 2, 3\}$).

We trained the logistic regression models by using the medical term and ICD-code pairs included in the annotated medical document set (ANDOC), *ntcir11_mednlp_mednlp2-train_v0.xml*. To increase the training data size, we added the basic medical term and ICD-code pairs defined in the basic table of the ICD-10 MEDIS Standard Master. Synonyms for the basic medical terms are included in the index table of the ICD-10 MEDIS Standard Master. We assumed that the synonyms should be assigned to the same ICD-code as the basic medical terms, and added the pairs of synonyms and inferred ICD-codes to the training data. Using the pairs of medical terms and ICD-codes collected from the ICD-10 MEDIS Standard Master, we also estimated the ICD-codes of Japanese and English synonyms included in additional medical term dictionaries (ADDIC) such as LSD, TDIC, and MedDRA, and then added them to the training data.

Table 1: Recall, precision, and F-measure scores of medical term extraction obtained with our method. Note that "test" data were used as unlabeled data for training the CRF models.

(a) Medical term extraction

Training data	Recall	Precision	F-measure
org	0.778	0.789	0.783
org with *dic*	0.781	0.782	0.782
org with *dic* and *test*	0.788	0.794	0.791

(b) Extraction of medical terms with modality attributes

Training data	Recall	Precision	F-measure
org	0.702	0.712	0.707
org with *dic*	0.706	0.707	0.707
org with *dic* and *test*	0.715	0.720	0.718

4. EXPERIMENTAL RESULTS

4.1 Task-1 Results

Table 1 (a) shows recall, precision, and F-measure scores for medical term extraction obtained with our method. With our method, we estimated medical terms included in the test data of Task-1, *test.xml*.

The "org" line in Table 1 (a) shows the experimental results obtained by using the labeled and unlabeled data described in Section 2.1. We also examined the performance obtained by using additional datasets, *dic* and *test*, for training character-based CRF models. The *dic* dataset is a list of medical terms collected from the ICD-10 MEDIS Standard Master and additional dictionaries such as LSD and TDIC. For "org with *dic*" and "org with *dic* and *test*," we used the medical term list as labeled data for training the CRF models. The *test* dataset is a raw document dataset, *test.xml*, provided for Task-1 evaluation by the task organizers. We added the test dataset to unlabeled data and then trained the CRF models to obtain the experimental results of "org with *dic* and *test*." As shown in Table 1 (a), the recall score obtained with the additional datasets, *dic* and *test*, was better than that without them.

[6]http://chasen.org/~taku/software/TinySVM/

[7]http://googlejapan.blogspot.jp/2007/11/n-gram.html

Table 2: ICD-code classification accuracies of medical terms obtained with our method. N_c shows the number of medical terms whose ICD-codes were estimated correctly. 2136 medical terms included in the golden standard test dataset were used for the evaluation.

Features	Accuracy	N_c
N-gram	0.792	1691
N-gram with *type*	0.781	1668
N-gram with *type* and *cluster*	0.777	1660

Table 3: ICD-code classification accuracies of medical terms obtained with classifiers trained by using different combinations consisting of the annotated medical document set (ANDOC), ICD-10 MEDIS Standard Master (MEDIS), and additional medical term dictionaries (ADDIC).

Training data	Accuracy	N_c
ANDOC	0.657	1403
MEDIS	0.551	1176
MEDIS and ADDIC	0.578	1235
ANDOC with MEDIS	0.789	1685
ANDOC with MEDIS and ADDIC	0.792	1691

Table 1 (b) shows the evaluation results for medical terms with modality attributes extracted by using our methods. We confirmed that using the additional datasets improved the recall scores for the extraction of medical terms with modality attributes.

4.2 Task-2 Results

Table 2 shows the ICD-code classification accuracies obtained with our method. To obtain the experimental results, we used 2136 medical terms included in golden standard test dataset, *test_goldstandard.xml*, where the medical terms were annotated with <*c*> tags by the task organizers. With our method, we estimated the ICD-codes to which the medical terms should be assigned.

The "N-gram" line in Table 2 shows the classification accuracy obtained with the feature vectors of the medical terms described in Section 3. We considered the additional features, *type* and *cluster*, for medical terms and examined the effect of those features experimentally. The *type* features were designed by using disease types provided in medical documents with *type* tag. We added the disease type information of a medical document to the feature vectors of medical terms included in the medical document, to obtain the experimental results of "N-gram with *type*" and "N-gram with *type* and *cluster*" lines in Table 2. The *cluster* features were designed by using the clustering results for character-based tri-grams included in medical terms.

As shown in Table 2, the classification accuracy of "N-gram" was better than those of "N-gram with *type*" and "N-gram with *type* and *cluster*." The additional features were not useful for improving the generalization performance of ICD-code classification in the MedNLP-2 task setting.

We also examined the classification accuracies obtained with classifiers trained by using different combinations consisting of ANDOC, MEDIS, and ADDIC, to confirm the effect of the additional training data, MEDIS and ADDIC. Note that the "ANDOC with MEDIS and ADDIC" line in Table 3 shows the same experimental result as the "N-gram" line in Table 2.

As shown in Table 3, the classifier trained by using ANDOC with MEDIS and ADDIC provided better classification accuracy than the classifiers obtained using only ANDOC or ANDOC with MEDIS. We confirmed that using MEDIS and ADDIC improved the generalization performance of the ICD-code classifier.

5. CONCLUSION

In this report, we outlined the methods we used for obtaining our experimental results for the SCT-D3 team, and discussed the results. For the Extraction of Complaint and Diagnosis task (Task-1), we employed a two-step approach where we first extracted medical terms from medical documents by using a semi-supervised CRF model and then classified the extracted medical terms into modality attributes by using SVMs. Our experimental results confirmed that using additional labeled and unlabeled data collected from dictionaries and documents improved the recall score of medical term extraction. For the Normalization of Complaint and Diagnosis task (Task-2), we employed a one-against-rest technique and logistic regression models to design an ICD-code classifier, and used the MEDIS Standard Master and additional medical term dictionaries for increasing the training data size. Our experimental results confirmed that the additional training datasets were effective in improving the generalization performance of the ICD-code classifier.

6. 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] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[3] 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 '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[5] J. Suzuki and M. Nagata. An efficient parsing framework for on-the-fly editing documents. In *Proceedings of the 19th Annual Meeting of the Association of Natural Language Processing (in Japanese)*, pages 904–907, 2013.

[6] X. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin, 2005.