Inference of ICD Codes from Japanese Medical Records by Searching Disease Names

Masahito Sakishita
Faculty of Informatics, Shizuoka University, Japan
msakishita@kanolab.net

Yoshinobu Kano
Faculty of Informatics, Shizuoka University, Japan
kano@inf.shizuoka.ac.jp

Abstract

Importance of utilizing medical information is getting increased as electronic health records (EHRs) are widely used nowadays. We aim to assign international standardized disease codes, ICD-10, to Japanese textual information in EHRs for users to reuse the information accurately. In this paper, we propose methods to automatically extract diagnosis and to assign ICD codes to Japanese medical records. Due to the lack of available training data, we dare employed rule-based methods rather than machine learning. We observed characteristics of medical records carefully, writing rules to make effective methods by hand. We applied our system to the NTCIR-12 MedNLPDoc shared task data where participants are required to assign ICD-10 codes of possible diagnosis in given EHRs. In this shared task, our system achieved the highest F-measure score among all participants in the most severe evaluation criteria. Through comparison with other approaches, we show that our approach could be a useful milestone for the future development of Japanese medical record processing.

1 Introduction

In these years, more medical institutes adopt EHRs of electronic media replacing paper media. However, natural language processing (NLP) technologies in medical fields tend to be underdeveloped; hospitals and clinics have been extremely reluctant to allow access to clinical data for researchers from outside the associated institutions (Chapman et al., 2011).

In order to develop NLP technologies of medical field, various shared tasks (contests, competitions, challenge evaluations, critical assessments) have been organized. One of the well-known medical-related shared tasks is the Informatics for Integrating Biology and the Bedside (i2b2) by the National Institutes of Health (NIH), which started in 2006 (Uzuner, 2008) now brought in SemEval as Clinical TempEval 2015 (Bethard et al., 2015) and Clinical TempEval 2016 (Bethard et al., 2016). The Text Retrieval Conference (TREC), which addresses more diverse issues, also launched the Medical Reports Track (Voorhees et al., 2012). The first European medical shared task was the ShARe/CLEF eHealth Evaluation Lab (Goeuriot et al., 2015; Kelly et al., 2014; Suominen et al., 2013). While they are mainly targeted at English, medical reports are written in native languages in most countries. Therefore, information retrieval techniques in individual languages are required to be developed.

As a first step of our research for the development of Japanese medical NLP field, we propose methods that automatically extract diagnosis from Japanese EHRs, assigning ICD (International Classification of Diseases) codes[1]. ICD is made by the World Health Organization (WHO) to record, analyze, interpret and compare medical data (disease and cause of death) that has been collected all over the world. The latest version is ICD-10. An ICD code consists of a single letter prefix and numbers (e.g. “I48”). Single letter prefix mostly represents a kind of disease (e.g. “I” stands for *ischemic heart disease*) and numbers represent detailed information of disease (e.g. “I48” stands for “*atrial fibrillation* and *flutter*”). ICD could be used to create machine readable data.

Even a human expert has difficulty assigning an appropriate ICD code. Only doctors with actual clinical experiences could understand real intention of diagnosis. In other words, expert techniques

[1] World Health Organization, International Classification of Diseases (ICD), available from : http://www.who.int/classifications/icd/en/

Proceedings of the Clinical Natural Language Processing Workshop, pages 64–68, Osaka, Japan, December 11-17 2016.

and experiences are required if a non-professional guesses the intention to assign codes without examining an actual patient. This point makes the automatic ICD coding tasks difficult.

We describe details of our methods in Section 2. Section 3 describes our experiments and results where we applied our system to the shared task data of the NTCIR-12 MedNLPDoc task (Aramaki et al., 2016). Our system achieved the best performance regarding the *Sure* match score of this MedNLPDoc task. Section 4 describes future works of our research and concluding this paper.

2 Method

We suggest five methods that output appropriate ICD code given a Japanese medical record text. In our system, method 2.1 is our base method. We defined methods 2.2-2.4 assuming results of method 2.1. Method 2.5 and part of method 2.4 are independent of method 2.1. We describe our methods one by one below.

2.1 Decision of target sentence

We define a "sentence" as a line of text marked off by the Japanese periodical symbol, "。".

We suggest that there are two types of sentences in medical records: sentences that include diagnosis, and sentences that do not include any diagnosis. The latter type of sentences may include disease names which are not related to any diagnosis.

When a sentence contains diagnosis, and when that sentence contains a name of disease, our system output a corresponding ICD code of that disease name. We describe details of our method below.

We extract sentences that contain a keyphrase to narrow candidate sentences down. For example, the previous example sentence with diagnostic result "検査の結果で慢性化膿性中耳炎と診断され、手術目的に入院となる。(As a result of medical check, diagnosed as *chronic suppurative otitis media*, and hospitalization is needed for an operation.)" has a keyphrase of "と診断され (be diagnosed)" with its diagnosis name of disease before the keyphrase. In addition to the keyphrase "と診断され", we listed and used keyphrases of "の診断 (diagnosis of)" , etc. 30 keyphrases in total. We chose these keyphrases by manually verifying medical records written in Toba (2006) and medical records of MedNLPDoc training data, which details are described later. If a sentence contains a negation, e.g. "認めない (not see)", this sentence is discarded from the candidate sentences.

After selecting sentence candidates, morphological analysis is performed by Kuromoji morphological analyzer[2] with a custom dictionary where Wikipedia entry words and disease names are registered. Disease names are taken from Japanese Standard Disease-Code Master (Hatano et al., 2003). We changed the weight of words in the dictionary in order to make disease names of the dictionary appear preferentially. When a disease name is included in the morphological analysis result, we assign a corresponding ICD code in the table of Japanese Standard Disease-Code Master.

2.2 Translation of medical technical words from English to Japanese

There are many English words used as technical terms in the Japanese medical records, written in alphabets. Because these English words are often not registered in our custom dictionary, we cannot deal with it directly. We used Life Science Dictionary (Ohtake et al., 2008) to translate English words into Japanese words. In this method, we only use dictionary entries which exactly matched with the English words in the medical record.

2.3 Unification of paraphrase words

There are many inconsistent spelling variations appear in the medical records. We deal with this problem by our method below. We use the redirection relations of Wikipedia to make such normalizations, i.e. redirected words correspond to normalized words.

2.4 Assigning ICD codes to disease names including various body parts

In our method described in section 2.1, descriptions like "XX に癌,YY に損傷 (*cancer* of XX,

[2] http://www. atilika.org/

damage to YY)" will only output corresponding ICD codes of *damage* or *cancer*, ignoring "XX" and "YY". However, these ignored words could include information required to output appropriate ICD codes. We decided to focus on "*malignant neoplasm*" and "*damage*" in our method. Our system outputs ICD codes from combination of words.

We define rules to detect ICD codes using combination of words that express various parts of body, and the words which represent *malignant neoplasm* and *damage*. We manually made a list of body parts using the Japanese Standard Disease-Code Master.

If a sentence contains both a word of the body parts and a word which represents *malignant neoplasm* or *damage*, our system outputs a corresponding ICD code.

In case of *damage*, we only check sentences selected by our method described in section 2.1, while we used the whole medical record in case of *malignant neoplasm*. This is because there are special keyphrases used for *malignant neoplasm*.

Our system covered almost all ICD codes of "*malignant neoplasm*" and "*damage*", including various body parts. We removed words which represent *malignant neoplasm* or *damage* from the dictionary used in method 2.1, because these words e.g. "癌 (*cancer*):C80" are sometimes used to refer specific concepts e.g. "肺癌 (*lung cancer*):C349" but not for the general meaning.

2.5 Inferring ICD codes from XML tags

We suggest another method that outputs ICD codes using information in XML tags of the MedNLPDoc task dataset. We focused on tags of *anamnesis* (既往歴) and *family clinical history* (家族歴), because there are categories of ICD codes directly correspond to these two types. If there is a tag of *anamnesis* or *family clinical history*, our system outputs an ICD code by extracting clues from words inside these tags. Then we apply the same method described in 2.4 to the extracted words.

3 Experiment and Result

3.1 Experiment Setting

We applied our system to the NTCIR-12 MedNLPDoc task. MedNLP is a shared task series for Japanese medical record texts in NTCIR (NII Testbeds and Community for Information access Research). Previous tasks include three sub tasks: named entity removal task (de-identification task), disease name extraction task (complaint and diagnosis), and normalization task (ICD coding task)(Morita et al., 2013). The MedNLPDoc task is more advanced and practical. In this task, participants' systems infer disease names in ICD. Due to this practical setting, task participants' systems could directly support actual daily clinical services and clinical studies in various areas (Aramaki et al., 2016).

Task organizers created a medical record corpus as a training dataset for this task which includes 200 individual medical records. The average number of sentences per record is 7.82. The average number of codes per record is 3.86. 552 code types appeared in the corpus.

Test dataset consists of 78 clinical texts, which were randomly selected from the past National Examination for Medical Practitioners [3]. Question sentences and graphics were eliminated from the original documents. Then, three professional human coders (more than one-year experience) individually added ICD-10 codes (Aramaki et al., 2016) to the same documents in parallel.

The MedNLPDoc task provides three evaluation metrics. *Sure* metric regards ICD codes which all of three annotators agreed to annotate, *Major* metric for more than two annotators, *Possible* metric

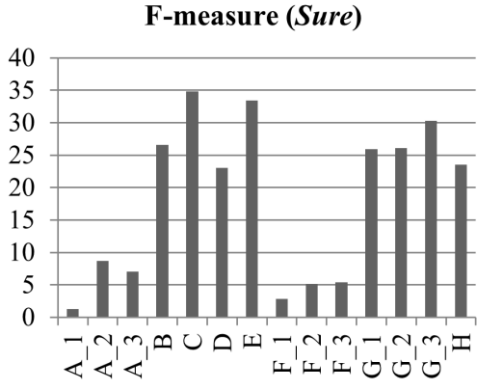

Figure 1. Comparison with other teams in F-measure (*Sure*), where C indicates our result

[3] Ministry of Health, Labour and Welfare, Question and the correct answer of the 108th national medical examination, available from : http://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp140512-01.html

for more than a single annotator. Because the inter-annotator discrepancy is quite low in this dataset, the *Sure* metric is considered as most reliable.

3.2 Result

We measured our system performance by participating in the MedNLPDoc task. Figure 1 shows results of all participants in the *Sure* evaluation metric. Our result is shown as Team C, which is the best score in F-measure *Sure* metric. Team C is rule-based, while others use machine learning methods, like CRF (Team B, E), CRF and SVM (Team G) (Aramaki et al., 2016).

Combination of Methods	**# of system output**	**# and scores of perfect match**				**# and scores of 3-digits match**			
		#	**P**	**R**	**F**	**#**	**P**	**R**	**F**
2.1	424	101	23.82	13.08	16.89	161	37.97	20.85	26.92
2.1+2.2	450	110	24.44	14.25	18.00	176	39.11	22.80	28.81
2.1+2.3	479	107	22.34	13.86	17.11	170	35.49	22.02	27.18
2.1+2.4	494	120	24.29	15.54	18.96	208	42.11	26.94	32.86
2.1+2.5	446	111	24.89	14.38	18.23	174	39.01	22.54	28.57
2.1+2.2+2.3+2.4+2.5	597	145	24.29	18.78	21.18	245	41.04	31.74	35.79

Table 1. Evaluation for combinations of methods in Precision (P), Recall (R) and F-measure (F)

3.3 Effect Analysis of Methods

As gold standard annotations of the test dataset are not provided, we conducted another experiment using the training data to show effectiveness of each of our methods. Table 1 shows result of this experiment. “perfect match” means the number of codes perfectly matched with the correct ICD codes. “3-digits match” means the number of output codes which three digits (first letter and next two numbers) are matched. Total number of correct answers was 772. We compared a couple of different combinations of our sub-methods, each described in section 2.1, 2.2, 2.3, 2.4, and 2.5, respectively.

Because the F-measure becomes better when methods 2.2-2.5 are added to 2.1, each individual method can be regarded as effective. When the method 2.4 is added, the growth of F-measure is the largest. Regarding *malignant neoplasms* and *damage*, we can write coding rules easier by hand because corresponding ICD descriptions explicitly discriminates “[body_part] and *damage*”, “[body_part] and the *cancer*”, etc. Additionally, *malignant neoplasms* and *damage* are frequently appeared in the training data, which made the contribution larger.

When method 2.3 is added, the growth of F-measure is the smallest. Reasons would be that coverage of paraphrases is insufficient with Wikipedia. Another reason is that the training data does not contain many paraphrases.

4 Future work and Conclusion

There should be two criteria required to achieve the ultimate goal of this ICD codes assignment study. The first criterion is whether symptoms are explicitly described or not in medical records. This decision would have almost been achieved by our approach except for *cancers*. Regarding *cancers*, our system could not select candidate sentences effectively in some cases because there were no keyphrases found as other phrases are used. Extracting such indirect expressions would be required.

The second criterion is whether we should output ICD codes or not, when we find out symptom or name of disease. Let us consider *cough* for example, which often appears in medical records. In order for the code of the *cough* to be assigned, we need to know how strong an effect of the *cough* gives to a patient’s diagnosis by deriving relationship of the *cough* and main diagnosis. Then we can recognize relationships between symptoms and diagnosis that could contribute to the real clinical works.

If we could properly define these two criteria, we can output more accurate ICD codes.

Japanese medical records contain language specific features like inclusion of diagnosis names, paraphrases, etc. From such features, we made five rule-based methods consisting our system that output ICD codes accurately. Our system performed best among participants in the MedNLPDoc task. However, it is still difficult to output ICD codes perfectly. In order to make better ICD coding in future, it will be required to analyze relationships between a patient’s symptom and his/her disease.

References

Aramaki, E., Morita, M., Kano, Y. and Ohkuma, T. (2016). Overview of the NTCIR-12 MedNLPDoc Task, 167–179.

Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J. and Verhagen, M. (2015). SemEval-2015 Task 6: Clinical TempEval. *Proceedings of the 9th International Conference on Semantic Evaluation (SemEval 2015)*, 806–814.

Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J. and Verhagen, M. (2016). SemEval-2016 Task 12: Clinical TempEval. *Proceedings of the 10th International Conference on Semantic Evaluation (SemEval 2016)*, 1052–1062.

Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K. and Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association : JAMIA*, *18*(5), 540–543. doi:10.1136/amiajnl-2011-000465

Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J. and Zuccon, G. (2015). Overview of the CLEF eHealth Evaluation Lab 2015.

Hatano, K. and Kazuhiko Ohe. (2003). Information Retrieval System for Japanese Standard Disease-Code Master Using XML Web Service.

Kelly, L., Goeuriot, L., Schreck, T., Leroy, G., Suominen, H., W.Chapman, W., Martinez, D., Velupillai, S., Mowery, D. L., et al. (2014). Overview of the CLEF eHealth evaluation lab 2014. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9283*, 429–443. doi:10.1007/978-3-319-24027-5_44

Morita, M., Kano, Y., Ohkuma, T., Miyabe, M. and Aramaki, E. (2013). Overview of the NTCIR-10 MedNLP task.

Ohtake, H., Fujita, N., Kaneko, S., Morren, B. and Kawamoto, T. (2008). Anatomy of Life Science English: Lists of common collocates of, (3).

Shibuki, H., Sakamoto, K., Kano, Y., Mitamura, T., Ishioroshi, M., Itakura, K. Y., Wang, D., Mori, T. and Kando, N. (2014). Overview of the NTCIR-11 QA-Lab Task. *the 11th NTCIR (NII Testbeds and Community for information access Research) workshop*, (Task 1), 518–529.

Suominen, H., Salanterä, S., Velupillai, S., W.Chapman, W., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., et al. (2013). Overview of the CLEF eHealth Evaluation Lab 2013, 1–20.

Toba, K. (2006). *ICD Coding Training(Second edition) in Japanese*. Igakushoin.

Uzuner, O. (2008). Second i2b2 workshop on natural language processing challenges for clinical records. *AMIA Annual Symposium proceedings*, 1252–1253.

Voorhees, E. M. and Hersh, W. (2012). Overview of the TREC 2012 Medical Records Track. *The Twentieth Text REtrieval Conference*.