De-identifying Free Text of Japanese Electronic Health Records

Kohei Kajiyama[1], Hiromasa Horiguchi[2], Takashi Okumura[3], Mizuki Morita[4] and Yoshinobu Kano[5]
[1,5]Shizuoka University
[2]National Hospital Organization
[3]Kitami Institute of Technology
[4]Okayama University
[1]kajiyama.kohei.14@shizuoka.ac.jp
[2]horiguchi-hiromasa@hosp.go.jp
[3]tokumura@mail.kitami-it.ac.jp
[4]mizuki@okayama-u.ac.jp
[5]kano@inf.shizuoka.ac.jp

Abstract

A new law was established in Japan to promote utilization of EHRs for research and developments, while de-identification is required to use EHRs. However, studies of automatic de-identification in the healthcare domain is not active for Japanese language, no de-identification tool available in practical performance for Japanese medical domains, as far as we know. Previous work shows that rule-based methods are still effective, while deep learning methods are reported to be better recently. In order to implement and evaluate a de-identification tool in a practical level, we implemented three methods, rule-based, CRF, and LSTM. We prepared three datasets of pseudo EHRs with de-identification tags manually annotated. These datasets are derived from shared task data to compare with previous work, and our new data to increase training data. Our result shows that our LSTM-based method is better and robust, which leads to our future work that plans to apply our system to actual de-identification tasks in hospitals.

1 Introduction

Recently, healthcare data is getting increased both in companies and government. Especially, utilization of Electronic Health Records (EHRs) is one of the most important task in the healthcare domain. While it is required to de-identify EHRs to protect personal information, automatic de-identification of EHRs has not been studied sufficiently for the Japanese language.

Like other countries, there are new laws for medical data treatments established in Japan. "Act Regarding Anonymized Medical Data to Contribute to Research and Development in the Medical Field" was established in 2018. This law allows specific third party institute to handle EHRs. As commercial and non-commercial health data is already increasing in recent years [1], this law promotes more health data to be utilized. At the same time, developers are required to de-identify personal information. "Personal Information Protection Act" was established in 2017, which requires EHRs to be handled more strictly than other personal information. This law defines personal identification codes including individual numbers (e.g. health insurance card, driver license card, and personal number), biometric information (e.g. finger print, DNA, voice, and appearance), and information of disabilities.

[1] (Ministry of Internal Affairs and Communication International Strategy Bureau, Information and Communication Economy Office, 2018)

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 65–70
Brussels, Belgium, October 31, 2018.

De-identification of structured data in EHRs is easier than that of unstructured data, because it is straightforward to apply de-identification methods e.g. k-anonymization (Latanya, 2002).

In the i2b2 task, automatic de-identification of clinical records was challenged to clear a hurdle of the Health Insurance Portability and Accountability Act (HIPAA) states (Özlem, Yuan, & Peter, 2007). There have been attempts to make k-anonymization for Japanese plain texts (Maeda, Suzuki, Yoshino, & Satoshi, 2016). Shared tasks of de-identification for Japanese EHRs were also held as MedNLP-1 (Mizuki, Yoshinobu, Tomoko, Mai, & Eiji, 2013) and MedNLP-2 (Aramaki, Morita, Kano, & Ohkuma, 2014).

While rule-based, SVM (Corinna & Vlandimir, 1995) and CRF (Lafferty, McCallum, & Pereira, 2001) were often used in these previous NER tasks, deep neural network model has shown better results recently. However, rule-based methods are still often better than machine learning methods, especially when there is not enough data, e.g. the best system in MedNLPDoc (Aramaki, Morita, Kano, & Ohkuma, Overview of the NTCIR-12 MedNLPDoc Task, 2016). The aim of the MedNLPDoc task was to infer ICD Codes of diagnosis from Japanese EHRs.

In this paper, we focus on de-identification of free text of EHRs written in the Japanese language. We compare three methods, rule, CRF and LSTM based, using three datasets that are derived from EHRs and discharge summaries.

We follow the MedNLP-1's standard of person information which require to de-identify "age", "hospital", "sex" and "time".

Methods

We used the Japanese morphological analyzer kuromoji[2] with our customized dictionary, as same as the best result team (Sakishita & Kano, 2016) in the MedNLPDoc task.

We implemented three methods as described below: rule-based, CRF-based, and LSTM-based.

1.1 Rule-based Method

Unfortunately, details and implementation of the best method of the MedNLP1 de-identification task (Imaichi, Yanase, & Niwa, 2013) are not publicly available. We implemented our own rule-based program based on their descriptions in their paper. Our rules are shown below. For a target word *x*,

option1	main rule		option2
翌 (next)	一昨年	two yeas ago	**より** (from)
前 (before)	昨年	last year	**まで** (until)
入院前 (before hospitalizetion)	先月	last month	**代** ('s)
入院後 (after hospitalizetion)	先週	last week	**前半** (ealry)
来院から (after visit)	昨日	yesterday	**後半** (last)
午前 (a.m.)	今年	this year	**～** (from)
午後 (p.m.)	今月	this month	~ (from)
発症から (after onset)	今週	this week	**以上** (over)
発症してから (after onset)	今日	today	**以下** (under)
治療してから (after care)	本日	today	**から** (from)
	来年	next year	**時** (when)
	来月	next month	**頃** (about)
	来週	next week	**ごろ** (about)
	翌日	tomorrow	**ころ** (about)
	再来週	the week after next	上旬 (early)
	明後日	day after tommorow	中旬 (mid)
	同年	same year	下旬 (late)
	同月	same month	春 (spring)
	同日	same day	夏 (summer)
	翌年	following year	秋 (fall)
	翌日	the next day	冬 (winter)
	翌朝	the next morning	朝 (morning)
	前日	the previous day	昼 (Noon)
	未明	early morning	夕 (evening)
	その後	after that	晩 (night)
	xx年	xx(year)	早朝 (early morning)
	xx月	xx(month)	明朝 (early morning)
	xx週間	xx(week)	以前 (before)
	xx日	xx(day)	以降 (after)
	xx時	xx(o'clock)	夕刻 (evening)
	xx分	xx(minutes)	ほど (about)

Table 1: our extraction rules for "age"

[2] https://www.atilika.com/en/kuromoji/

age (subject's years of age with its suffix**)**

- If the detailed POS is *number*, apply rules in Table 1

hospital (hospital name**)**

- If one of following keywords appeared, then mark as *hospital*: 近医 (a near clinic or hospital), 当院 (this clinic or hospital), 同院 (same clinic or hospital)
- If POS is *noun* and detailed-POS is not *non-autonomous word*, or *x* is either "●", "○", "▲" or "■" (these symbols are used for manual de-identification due to the datasets are pseudo EHRs), then if suffix of *x* is one of following keywords, mark as *hospital*: 病院 (hospital or clinic), クリニック (clinic), 医院 (clinic)

sex

- If either 男性 (man), 女性 (woman), men, women, man, woman, then mark as *sex*

time (subject's time with its suffix**)**

- If detailed-POS is *number* and *x* is concatenation of four or two, or one digit number, slash and two-digit number (e.g. yyyy/mm or mm/dd) then mark as *time*
- If detailed-POS of *x* is *number* and followed with either 歳 (old), 才 (old), 代 ('s), mark as *time*
    - If it is further followed with either "より"," まで "," 前半"," 後半"," 以上"," 以下"," 時"," 頃","ごろ","ころ","から", "前半から", "後半から", "頃から", "ごろから","ころから" and so on include these words in the marked *time*

1.2 CRF-based Method

As a classic machine learning baseline method of series labelling, we employed CRF. Many teams of the MedNLP1 de-identification task used CRF, including the second best team and the baseline system. We used the mallet library[3] for our CRF implementation. We defined five training features for each token as follows: part-of-speech (POS), detailed POS, character type (Hiragana, Katakana, Kanji, Number,), whether the token is included in our user dictionary or not, and a binary feature whether the token is beginning of sentence or not.

1.3 LSTM-based Method

We used a machine learning method that combines bi-LSTM and CRF using character-based and word-based embedding, originally suggested by other group (Misawa, Taniguchi, Yasuhiro, & Ohkuma, 2017). In this method, both characters and words are embedded into feature vectors. Then a bi-LSTM is trained using these feature vectors. Finally, a CRF is trained using the output of the bi-LSTM, using character level tags.

The original method uses a skip-gram model to embed words and characters by seven years of Mainichi newspaper articles of almost 500 million words. However, we did not use skip-gram model but GloVe[4], because GloVe is more effective than skip-gram (Pennington, Socher, & Manning, 2014). We used existing word vectors[5] instead of the pre-training in the original method. Our training and prediction is word based while the original method is character based. Our implementation is based on an open source API[6].

2 Experiment

2.1 Data

Our dataset is derived from two different sources. We used the MedNLP-1 de-identification task data to compare with previous work. This data includes pseudo EHRs of 50 patients. Although there were training data and test data provided, the test data is not publicly available now, which makes direct comparison with previous work impossible. However, both training and test data are written by the same writer and was originally one piece of data. Therefore, we assume that the training data can be regarded as almost same as the test data in their characteristics.

Another source is our dummy EHRs. We built our own dummy EHRs of 32 patients, assuming that the patients are hospitalized. Documents of our dummy EHRs were written by medical professionals (doctors). We added manual annotations for de-identification following a guideline of the MedNLP-1 task. These annotations were assigned by ourselves.

[3] http://mallet.cs.umass.edu/sequences.php
[4] https://nlp.stanford.edu/projects/glove/
[5] http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/
[6] https://github.com/guillaumegenthial/sequence_tagging

All of these data are assigned five types of de-identification tag; *age*, *hospital*, *sex*, *time* and *person*. MedNLP-1 data includes 2244 sentences and our dummy EHRs include 8327 sentences. Writers hold doctor's licenses in both sources, assuming fake patients to describe pseudo medical records. However, descriptions are not similar between the two sources, probably because of the difference of the writers.

	Rule based	CRF	CRF Mix	LSTM	LSTM Mix
ALL	**84.23**	82.62	26.40	80.61	66.25
age	**93.43**	71.12	32.55	88.49	91.68
hospital	84.73	87.09	26.02	**92.90**	84.82
person	N/A	N/A	N/A	N/A	N/A
sex	**50.00**	16.67	14.65	0.00	**50.00**
time	82.61	83.88	26.12	**94.32**	87.53

Table 2: F1 value testing of MedNLP1's dataset. There were no "person" annotations in this dataset..

	Rule based	CRF	CRF Mix	LSTM	LSTM Mix
ALL	43.74	66.97	67.13	77.20	**77.66**
age	51.13	48.46	38.87	75.69	**79.16**
hospital	15.98	47.85	48.62	67.57	**68.70**
person	N/A	26.96	28.36	**65.60**	65.06
sex	93.75	35.92	90.08	45.51	**98.08**
time	49.48	71.28	70.60	89.17	**90.92**

Table 3: F1 value testing of dummy-EHR dataset. We did not implement rules for "person".

2.2 Evaluation method

Our evaluation method followed MedNLP-1, using the IOB2 tagging (Tjong & Jorn, 1999). We applied four hold cross validation, while the rule-based method does not require training data. From the two sources described above, we derived three datasets: MedNLP-1, dummy EHRs, and both of MedNLP1 and dummy EHRs (mixture). We trained CRF and LSTM by this mixture data. We divided each data source for our cross-fold validation to hold the same balance of these two sources. Our evaluation metrics is strict match of named entities.

3 Result and Discussion

3.1 Result of MedNLP-1 dataset

Table 2 shows the evaluation results. The best F1 score is by the rule-based method. This is because the rules were tuned for the MedNLP-1 data. In both of datasets, CRF and LSTM are not significantly different from the rule-based one. LSTM performed best for the *hospital* tag and the *time* tag, probably because they might have typical patterns of less variations. Total occurrence of *sex* is very small, *person* is zero, in the MedNLP-1 dataset.

	MedNLP1	dummy	Mix
ALL	26.40	67.13	47.10
age	32.55	38.87	36.28
hospital	26.02	48.62	32.27
person	N/A	28.36	18.04
sex	14.65	90.08	53.83
time	26.12	70.60	51.01

Table 4: F1 value of trained Mix dataset by CRF

	MedNLP1	dummy	Mix
ALL	66.25	77.66	76.21
age	91.68	79.16	86.35
hospital	84.82	68.70	72.18
person	N/A	65.06	65.06
sex	50.00	98.08	98.08
time	87.53	90.92	90.55

Table 5: F1 value of trained Mix dataset by LSTM

3.2 Result of Dummy-EHR dataset

The result is shown at Table 3. The best score is performed by LSTM trained by the mixture dataset. Despite the data size is four times larger than that of MedNLP-1, the result is a little better. Regarding CRF, training with mixture dataset is worse than the dummy her dataset only. This is not true for LSTM, which shows better results when trained by mixture dataset.

3.3 Overall

We trained CRF and LSTM by the mixture dataset and evaluated on MedNLP-1, dummy-EHR and mixture dataset individually. These results are shown in Table 4 and Table. Regarding CRF, there is 26 point difference in average between evaluations with MedNLP-1 and dummy-EHR datasets. On the other hand, LSTM shows 7 point difference in average. These results suggest that the datasets are quite different, but LSTM absorbed these differences well.

4 Conclusion and Future Work

We implemented three different de-identification methods for Japanese EHRs. We applied these

methods to three datasets derived from two different pseudo EHR sources with de-identification tags manually annotated. Our results show that LSTM is better than other methods also shows robustness between different sources compared with CRF. Machine learning methods could extract named entities of de-identification comparable to the rule based method that is manually tuned to specific target data. However, machine learning method is still weak for expressions with low occurrences. Combination of LSTM and rule-based method could be a future work.

Because the current performance is enough high among publicly available Japanese de-identification tools, we plan to apply our system to actual de-identification tasks in hospitals. Although it is still difficult to make real EHRs publicly available, we could use our large amount of EHRs inside our hospitals. Increasing the annotated dataset for such internal usage would be another future work.

5 Acknowledgement

This work was partially supported by Japanese Health Labour Sciences Research Grant and JST CREST.

References

Aramaki, Eiji, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. "Overview of the NTCIR-11 MedNLP-2 Task." *Proceedings of the 11th NTCIR conference*, 2014: 147-154.

Aramaki, Eiji, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. "Overview of the NTCIR-12 MedNLPDoc Task." *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies* (Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies), 2016: 147-154.

Corinna, Cortes, and Vapnink Vlandimir. "Support-Vector Networks." *Machine Learning*, 1995: 20:273-297.

Hochreiter, Sepp, and Jürgen Schmidhunber. "LONG SHORT-TERM MEMORY." *NEURAL COMPUTATION 9(8)*, 1997: 1735-1780.

Imaichi, Osamu, Toshihiko Yanase, and Yoshiki Niwa. "A Comparison of Aramaki, E., Morita, M., Kano, Y., & Ohkuma, T. (2014). Overview of the NTCIR-11 MedNLP-2 Task. *Proceedings of the 11th NTCIR conference*, 147-154.

Aramaki, E., Morita, M., Kano, Y., & Ohkuma, T. (2016). Overview of the NTCIR-12 MedNLPDoc Task. *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, 147-154.

Corinna, C., & Vlandimir, V. (1995). Support-Vector Networks. *Machine Learning*, 20:273-297.

Hochreiter, S., & Schmidhunber, J. (1997). LONG SHORT-TERM MEMORY. *NEURAL COMPUTATION 9(8)*, 1735-1780.

Imaichi, O., Yanase, T., & Niwa, Y. (2013). A Comparison of Rule-Based and Machine Learning Methods for Medical Information Extraction. *International Joint Conference on Natural Language Processing Workshop on Natural Language Processing for Medical and Healthcare Fields*, 38-42.

Lafferty, J. D., McCallum, A., & Pereira, F. C. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *ICML2001*, 282-289.

Latanya, S. (2002). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *Int J Uncerainty, Fuzziness Knowledge-Based Systems*, 10:557-570.

Ma, X., & Hovy, E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1064-1074.

Maeda, W., Suzuki, Y., Yoshino, K., & Satoshi, N. (2016). Anonymization technique for Unstructured text data considering inference from context. *Forum on Information Tecnology Vol.2*, 47-48.

Ministry of Internal Affairs and Communication International Strategy Bureau, Information and Communication Economy Office. (2018, 7 8). *Survey research report on the weekly report on information distribution / accumulation volume.* Japan, Tokyo. Retrieved from http://www.soumu.go.jp/johotsusintokei/linkdata/h25_03_houkoku.pdf

Misawa, S., Taniguchi, M., Yasuhiro, M., & Ohkuma, T. (2017). Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition. *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, 97-102.

Mizuki, M., Yoshinobu, K., Tomoko, O., Mai, M., & Eiji, A. (2013). *Overview of the NTCIR-10 MedNLP task.* Tokyo, Japan: In Proceedings of NTCIR-10.

Özlem, U., Yuan, L., & Peter, S. (2007). Evaluating the State-of-the-Art in Automatic De-identification. *J Am Med Inform Assoc, Sep-Oct*(14), 550-563.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543.

Sakishita, M., & Kano, Y. (2016). *Inference of ICD Codes from Japanese Medical Records by Searching Disease Names.* Clinical Natural Language Processing Workshop at the 26th International Conference on Computational Linguistics (COLING 2016).

Tjong, K. S., & Jorn, V. (1999). *Representing Text Chunks.* Proceedings of EACL '99.