Accepted Manuscript online: 2023-02-21 Article published online: 2023-04-04

An Alternative Application of Natural Language Processing to Express a Characteristic Feature of Diseases in Japanese Medical Records

Yoshinori Yamanouchi[1] Taishi Nakamura[1] Tokunori Ikeda[2] Koichiro Usuku[1]

[1] Department of Medical Information Science, Graduate School of Medical Sciences, Kumamoto University, Kumamoto, Japan
[2] Department of Pharmaceutical Sciences, Faculty of Pharmaceutical Sciences, Sojo University, Nishi-ku, Kumamoto, Japan

Address for correspondence Taishi Nakamura, MD, PhD, Department of Medical Information Science, Graduate School of Medical Sciences, Kumamoto University, 1-1-1, Honjo, Chuo-ku, Kumamoto 860-8556, Japan (e-mail: taishin@kumamoto-u.ac.jp).

Methods Inf Med 2023;62:110–118.

Abstract

Background Owing to the linguistic situation, Japanese natural language processing (NLP) requires morphological analyses for word segmentation using dictionary techniques.
Objective We aimed to clarify whether it can be substituted with an open-end discovery-based NLP (OD-NLP), which does not use any dictionary techniques.
Methods Clinical texts at the first medical visit were collected for comparison of OD-NLP with word dictionary-based-NLP (WD-NLP). Topics were generated in each document using a topic model, which later corresponded to the respective diseases determined in International Statistical Classification of Diseases and Related Health Problems 10 revision. The prediction accuracy and expressivity of each disease were examined in equivalent number of entities/words after filtration with either term frequency and inverse document frequency (TF-IDF) or dominance value (DMV).
Results In documents from 10,520 observed patients, 169,913 entities and 44,758 words were segmented using OD-NLP and WD-NLP, simultaneously. Without filtering, accuracy and recall levels were low, and there was no difference in the harmonic mean of the F-measure between NLPs. However, physicians reported OD-NLP contained more meaningful words than WD-NLP. When datasets were created in an equivalent number of entities/words with TF-IDF, F-measure in OD-NLP was higher than WD-NLP at lower thresholds. When the threshold increased, the number of datasets created decreased, resulting in increased values of F-measure, although the differences disappeared. Two datasets near the maximum threshold showing differences in F-measure were examined whether their topics were associated with diseases. The results showed that more diseases were found in OD-NLP at lower thresholds, indicating that the topics described characteristics of diseases. The superiority remained as much as that of TF-IDF when filtration was changed to DMV.
Conclusion The current findings prefer the use of OD-NLP to express characteristics of diseases from Japanese clinical texts and may help in the construction of document summaries and retrieval in clinical settings.

Keywords
- natural language processing
- text mining
- topic model
- electronic health record
- Japanese clinical texts

received
October 17, 2021
accepted after revision
April 13, 2022
accepted manuscript online
February 21, 2023
article published online
April 4, 2023

DOI https://doi.org/10.1055/a-2039-3773.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Introduction

In the past two decades after the introduction of the electronic health record (EHR) system in Japan, the implementation rate increased to 78.1%, especially in hospitals with 400 beds and above.[1] Various data types are automatically saved in the EHR systems so that the stored data can be utilized quickly in clinical practices, and be adapted for secondary use, including research works, improvement of medical services, and hospital management.[2]

The stored data are divided into two subsets: structured and unstructured data, where different data marts are built for the analysis use.[3] Structured data have a kindred feature that conforms to a certain format in defined schemes. The link by codes principally enables data models to be easily built for analysis. Conversely, unstructured data, including natural language texts, images, videos, and voice records, cannot be predefined. These data have formats such as text, DICOM, JPEG, MEPG, or WAVE, although the meaning cannot be defined; thus, the content of stored data changes depending on the situation. Contrary to structured data, unstructured data cannot combine master data and is unable to aggregate or subdivide codes, resulting to difficulties in making data mart design. However, unstructured data are considered useful for performing data analysis because they flexibly store a broad array of information and can also provide informative clinical decision making.[4,5] Therefore, analyzing the text data in EHR systems has been considered helpful for practical use.[6,7]

Notably, before conducting an analysis with text data, we need to perform natural language processing (NLP): first, a sentence is broken down into word segments; second, a syntactic analysis to confirm the relationship between respective word segments is performed, and lastly, semantic analyses using term dictionaries are required. After all these processes, we can start performing the analysis.[8] Hereof, contrary to English, Japanese is known to be not a straightforward language, especially in the first step, because there are no obvious word boundaries with clear separators.[9,10] Moreover, written Japanese utilizes three orthographical systems: hiragana, katakana, and kanji characters derived from Chinese. According to the linguistic situation, word segmentation is usually conducted by a combination of N-gram search and morphological analysis, which can establish the sequence of characters by referring to a term dictionary and analyzed with the highest coincidence.[11] However, such word segmentation often leads to misunderstanding of unknown words, abbreviations, and acronyms, thus leading to false recognition as a different set of words. Certainly, accumulating articles have reported that the accuracy of segmentation for clinical texts written in Japanese depends on the number of words in the word dictionary (WD).[12–14] Therefore, an open-end discovery (OD) method was proposed to overcome the problem. The method divides a sentence into a series of meaningful units called entities that consist of certain consecutive and semantically relevant words,[15] and potentially formulate the feature of documents. Character segmentation has not outperformed word segmentation.[16,17] However, based on the bottom-up parser without a WD, it is expected that open-end discovery-based NLP (OD-NLP) can extract more helpful information from texts than word dictionary-based NLP (WD-NLP) (►**Supplemental Fig. S1**, available in the online version). Here, we conducted two distinct NLP methods in the resource-poor language of Japanese and successfully compared them using a topic model that exhibited more characteristic features of diseases from existing medical records.

Methods

Study Design

We retrospectively collected medical records in the EHR systems from 10,520 patients who visited Kumamoto University Hospital between January 2015 and December 2017. Clinical texts at the first medical visit from subjective and objective findings in the Subjective Objective Assessment Plan format were gathered without discriminating clinical departments. Additionally, International Statistical Classification of Diseases and Related Health Problem (ICD)-10 codes in the diagnosis procedure combination, which represents the names of main disease in the Japanese medical fee system,[18] were collected. All researches were performed in accordance to the declaration of Helsinki and approval of institutional review board of the Kumamoto University (Permit Number: 959).

Natural Language Processing

As shown in ►**Fig. 1**, we applied IRIS NLP Japanese (Ver2017–02, InterSystems Co., One Memorial Drive, Cambridge, MA 02142, United States) and MeCab (Ver0.996, Kudo, Japan) for OD-NLP[19,20] and WD-NLP,[21] respectively. In case of MeCab, a major software of Japanese NLP,[22–24] ComeJisyo (Ver 5–1, Sagara, Japan), was used as a medical term dictionary for word segmentation.[25] iKnow, a library of IRIS NLP Japanese, enables the segmentation of a sentence into a series of semantic units called entities. From the three available indices for entities such as term frequency (TF), TF-inverse document frequency (TF-IDF), and dominance value (DMV), we used (TF-IDF) and (TF-IDF) for filtration in this study. TF-IDF is often used to count the number of documents in entities,[26] while DMV represents the importance of an entity in a document, which is calculated by the syntactic position and frequency of an entity. Exclusion criteria are as follows: In MeCab, useless words (example numbers and datetime) were eliminated, but nouns, verbs, adjectives, adjective verbs, and adverbs were all extracted. Consecutive and same parts of speech words were combined during the process. In IRIS NLP Japanese, entities such as numeric, datetime, and unrecognized character strings were eliminated. Words/entities were excluded when the term count was less than 10 in one document. The number of diseases less than 10 or more than 1,000 were excluded in a two-tailed manner. Additionally, inadaptable documents for ICD-10 codes were eliminated. Several datasets were generated from MeCab and IRIS NLP according to number of words/entities with filtration of TF-IDF. Only in IRIS NLP Japanese, DMV was additionally utilized to create equivalent datasets to TF-IDF.

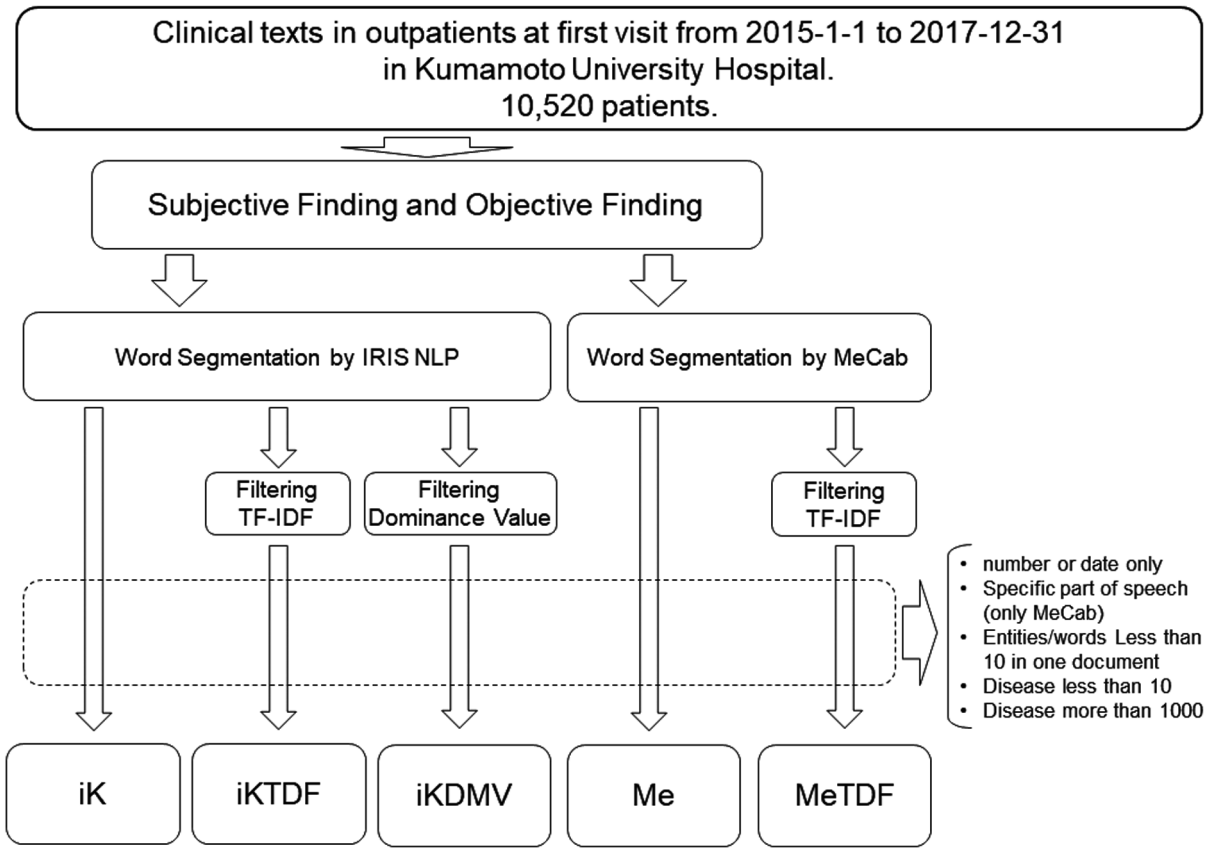

Fig. 1 Process of dataset creation and the name. Selection flows display how to do word segmentation into entities/words from the documents at the first visit to the hospital. Created dataset names are shown at the bottom for IRIS natural language processing (NLP) and MeCab, respectively. iK, iKnow; Me, MeCab; TF-IDF, term frequency and inverse document frequency.

Topic Model

The topic model is a probabilistic model that putatively generates a description of documents based on the distribution of topics and frequency of their terms. We estimated diseases using Latent Dirichlet allocation (LDA), which allowed each document to present multiple topics with different proportions, therefore, enabling the feature of each document to be described in a mixture of topics.[27] We generated an equal number of topics as in each disease name. We calculated topic probability distributions in each document-derived disease name, and gave the top 10 topics a score between 10 and 1 in the order of high points of occurrence. This step was performed in all documents, and the mean scores of topics for all diseases were given. In a cross-tabulation table, the horizontal and vertical axes represent the type of disease and topic, respectively. The relationship of the mean score between diseases and topics is presented (►**Supplemental Fig. S2A** [available in the online version]). To maximize the sum of each mean score in each dataset of cross-tabulation table, we performed the optimization by replacing the topic position on the diagonal lines (►**Supplemental Fig. S2B** [available in the online version]). From these processes, we achieved the type of topics corresponding to the respective diseases. Python version 3.4.5 (Python Software Foundation, United States) and LDA Gensim version 3.3.0 (RARE Technologies Ltd, CZ) were used in the analyses.

Practical Comparison of Topic Terms

The topic model can generate topic terms using a topic. The terms were blindly evaluated by physicians whether they related to respective diseases, if the number of entities/words was equivalent to each dataset with NLPs. From the top 30 terms generated by the topic model from dataset, the number of terms related to each disease was practically counted by the disease, and the highest counts among NLP methods were determined as the most accurate method for revealing disease characteristics. The ratio of diseases was given by dividing the number of diseases by all diseases that were common to the corresponding datasets for each comparison.

Statistical Analysis

In the existence of disease linked to each topic, we generated 2×2 contingency tables as follows: if the top topic from the document expected actual disease, then the disease was properly predicted (true positive). It was inappropriately predicted when the top topic was analyzed with incompatible diseases (false positive) and vice versa (false negative); otherwise, it is regarded as true negative. These evaluations were performed for each disease to create 2×2 contingency tables (►**Supplemental Fig. S3** [available in the online version]). In each table, the precision ratio, recall rate, and the harmonic mean of the F1 score were calculated to compare the performances of IRIS NLP and MeCab. The

Wilcoxon rank-sum test was used for analyses, which were performed using R version 3.4.1 (The R Foundation for Statistical Computing, Vienna, Austria), with two-sided tests, and the statistical significance was set to p less than 0.05.

Results

When NLP was performed without filtration in clinical texts of subjective and objective findings, both of which originated from 10,520 study patients, OD-NLP (i.e., IRIS NLP Japanese) and WD-NLP (i.e., MeCab) segmented documents into 169,913 entities and 44,758 words, respectively (►Table 1). The number of entities segmented by OD-NLP was approximately four times higher than that of words by WD-NLP, while the number of all documents and corresponding diseases was similar to the two NLP methods. To compare the two NLP methods in equivalent number of entities/words, six datasets were created for the respective NLPs depending on the threshold of filtration with TF-IDF. In word segmentation by OD-NLP, six datasets were further generated to ascertain the importance of another DMV filtration according to the number of entities by TF-IDF (►Fig. 1).

Because the number of documents and corresponding diseases were similar between OD-NLP and WD-NLP without filtration (i.e., datasets of iK and Me), we first compared the accuracy of prediction in disease names from medical records between two NLPs by generating appropriate topics using topic model, which corresponded to respective diseases (►Table 2, ►Supplemental Fig. S2 [available in the online version]). ►Table 2 shows that generated topics in each document revealed high accuracy and specificity in estimating diseases. However, we found that both NLPs resulted in low values in precision, recall, and F-measure. There were no differences in precision ($p = 0.874$) and F-measure ($p = 0.957$) except for recall ($p = 0.002$) between the two NLPs, indicating a comparable disease prediction from topics between OD-NLP and WD-NLP (►Fig. 2A). On the contrary, when medical doctors assessed their topic terms, including extracted entities/words with respective diseases, five out of six them fairly evaluated without discriminating among NLP methods. The top 30 topic terms by OD-NLP contained more meaningful topics for respective diseases than those by WD-NLP (►Fig. 2B). The superiority of topic terms in OD-NLP was associated with the disease compared with those in WD-NLP, which substantially increased by 66 diseases on average in a total of 114 diseases.

Next, we evaluated the relationship between the number of entities/words in the respective diseases, depending on the threshold of TF-IDF and F-measure of the topics

Table 1 Overview of datasets

Dataset name	Segmentation-method	Filtered no.	Filtered-method	Filtered-threshold	Documents	Entities/words	Document frequency 95%	Diseases
iK	IRIS NLP	Filtered-0	No filtered	None	8,421	1,69,913	7	114
iKTDF260	IRIS NLP	Filtered-1	TF-IDF	≥0.26	3,744	34,556	3	87
iKTDF330	IRIS NLP	Filtered-2	TF-IDF	≥0.33	2,411	21,270	2	68
iKTDF425	IRIS NLP	Filtered-3	TF-IDF	≥0.425	978	8,663	2	34
iKTDF510	IRIS NLP	Filtered-4	TF-IDF	≥0.51	308	2,577	1	15
iKTDF580	IRIS NLP	Filtered-5	TF-IDF	≥0.58	76	508	1	5
iKTDF620	IRIS NLP	Filtered-6	TF-IDF	≥0.62	22	122	1	2
Me	MeCab	Filtered-0	No Filtered	None	9,647	44,758	79	118
MeTDF033	MeCab	Filtered-1	TF-IDF	≥0.033	9,647	34,875	39	118
MeTDF084	MeCab	Filtered-2	TF-IDF	≥0.084	7,911	20,474	17	109
MeTDF167	MeCab	Filtered-3	TF-IDF	≥0.167	2,807	8,239	8	74
MeTDF245	MeCab	Filtered-4	TF-IDF	≥0.245	717	2,544	4	30
MeTDF306	MeCab	Filtered-5	TF-IDF	≥0.306	162	481	3	11
MeTDF346	MeCab	Filtered-6	TF-IDF	≥0.346	30	77	3	3
iKDMV400	IRIS NLP	Filtered-1	DMV	≥400	3,923	34,716	3	81
iKDMV500	IRIS NLP	Filtered-2	DMV	≥500	2,583	20,570	3	63
iKDMV600	IRIS NLP	Filtered-3	DMV	≥600	1,163	8,297	2	35
iKDMV700	IRIS NLP	Filtered-4	DMV	≥700	425	2,565	2	17
iKDMV800	IRIS NLP	Filtered-5	DMV	≥800	109	486	3	5
iKDMV900	IRIS NLP	Filtered-6	DMV	≥900	23	59	2	2

Abbreviations: DMV, dominance value; iK, iKnow; Me, MeCab, ; NLP, natural language processing; TF-IDF, term frequency-inverse document frequency.

Table 2 Comparison of IRIS NLP with MeCab in no filtered datasets

	Dataset	Min	25%	Median	75%	Max	*p*-Value
Precision	iK	0.002	0.008	0.015	0.035	0.354	0.874
	Me	0.003	0.008	0.015	0.037	0.370	
Recall	iK	0.067	0.167	0.233	0.343	0.636	0.002
	Me	0.043	0.200	0.295	0.455	0.933	
F-measure	iK	0.005	0.015	0.029	0.064	0.299	0.957
	Me	0.006	0.016	0.028	0.066	0.356	
Accuracy	iK	0.668	0.901	0.931	0.959	0.990	0.077
	Me	0.735	0.880	0.923	0.945	0.988	
Specificity	iK	0.672	0.910	0.938	0.964	0.991	0.066
	Me	0.738	0.883	0.927	0.950	0.991	

Abbreviations: iK, iKnow; Me, MeCab; NLP, natural language processing.

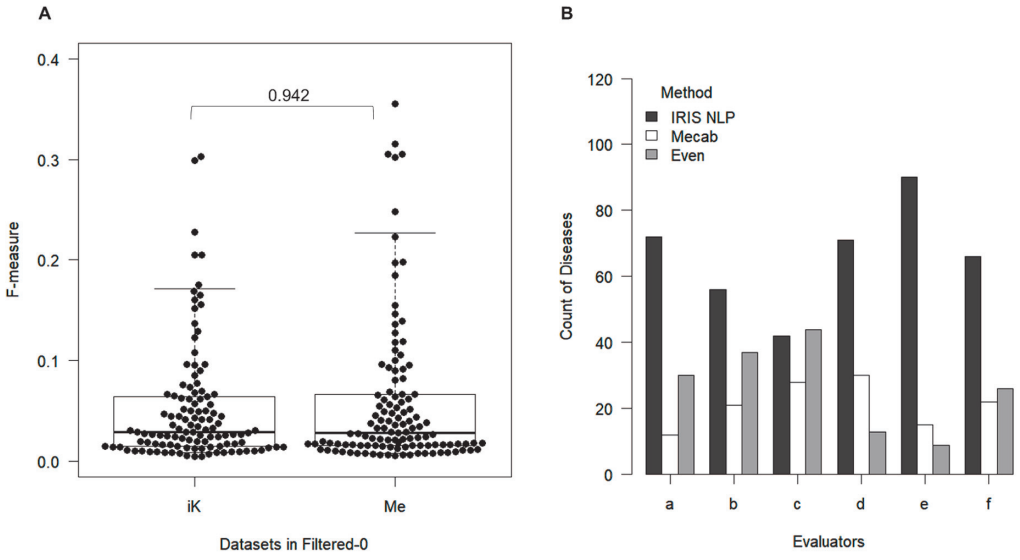

Fig. 2 Comparisons in no filtered datasets between IRIS natural language processing (NLP) and MeCab. (**A**) Box plots of F-measure with median in Filtered-0. *p*-Value showed no differences by Wilcoxon rank-sum test. (**B**) The number of diseases that were the most counted in each physician who blindly evaluated the number of topic terms expressing respective diseases. When the number of counted terms was the same between NLPs, it was regarded as "Even." iK, iKnow.

(►**Fig. 3A**). Each dataset showed an equivalent number of entities/words according to the respective thresholds of TF-IDF in both NLPs. According to the equalization, we created six different datasets, namely Filtered-1, -2, -3, -4, -5, and -6, as summarized in ►**Table 1**. In the first three datasets with lower thresholds of TF-IDF, the F-measure in OD-NLP showed significantly higher accuracy than that in WD-NLP; $p = 0.027$ in Filtered-1, p less than 0.001 in Filtered-2, and p less than 0.001 in Filtered-3. The higher the threshold, the less the number of entities/words extracted, which resulted in a decrease in corresponding diseases, thus, increasing values of F-measure. However, we confirmed that statistical differences in F-measure between OD-NLP and WD-NLP disappeared at high F-measure values ($p = 0.112$ in Filtered-4, $p = 0.090$ in Filtered-5, and $p = 0.200$ in Filtered-6, respectively).

Then, in two datasets near the maximum threshold of TF-IDF filtration that showed significant differences in F-measure between NLPs (i.e., Filtered-3 and -4), medical doctors examined whether corresponding topic terms including extracted entities/words were associated with disease (►**Fig. 3B**). In comparison between iKTDF425 and MeTDF167 in Filtered-3, 25 common diseases were evaluated, and we confirmed that there were a large number of diseases in OD-NLP, whose topic terms were clearly associated with respective diseases ($p = 0.035$). Therefore, this was recapitulated in the absence of filtration. Alternatively, in nine diseases from comparison datasets between iKTDF510 and MeTDF245 in Filtered-4, the rate of diseases was similar to that in Filtered-0 and -3. However, there were no differences in the number of diseases in which the topic terms were associated with the respective diseases ($p = 0.149$).

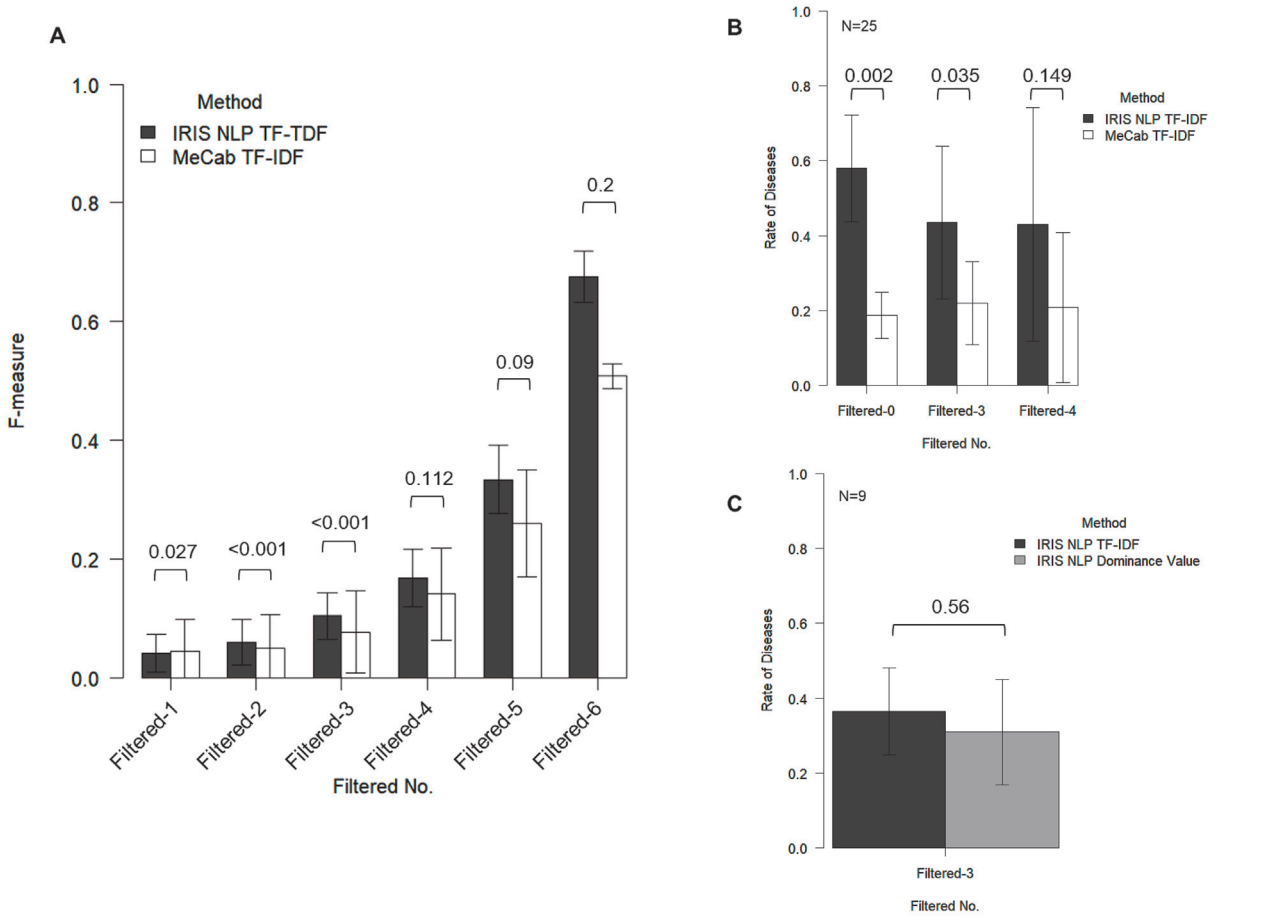

Fig. 3 Evaluations of filtered topic models. (**A**) Comparisons of F-measure in equalized number of entities/words with term frequency and inverse document frequency (TF-IDF) filtration between IRIS natural language processing (NLP) and MeCab. (**B**) The rate of diseases shows which NLP method had the most related topic terms to respective diseases by the datasets. "Even" is not shown. (**C**) Differences of filtration in IRIS NLP between TF-IDF and dominance value in the rate of diseases as evaluated in 3B. Data are shown in means ± standard deviation.

Lastly, the disparity of filtration in disease prediction between TF-IDF and DMV was assessed in OD-NLP (►**Table 3**). As the DMV threshold increased, the F-measure increased, similar to TF-IDF. F-measure in DMV was significantly higher than that in TF-IDF only at the lowest threshold; $p = 0.015$ between iKTDF260 and iKDMV400, but there were no differences in other thresholds between TF-IDF and DMV (►**Supplemental Fig. S4**, available in the online version).This ascertains the superiority of OD-NLP over WD-NLP, even when the filtration was changed from TF-IDF to DMV. In iKDMV600, corresponding to iKTDF425 in Filtered-3, the topic terms were examined in their common 25 diseases by medical doctors. The results substantially confirmed that topic terms in DMV were equally associated with diseases to those in TF-IDF (►**Fig. 3C**).

Discussion

We compared the two different NLP methods in clinical texts from subjective and objective findings, one with ordinary dictionary type: WD-NLP (MeCab) and the other without a dictionary, called OD-NLP (IRIS NLP Japanese). Morphological analyses with dictionary techniques have been conducted, to find the sequence of three orthographical characters in written Japanese, because of the linguistic character without obvious word boundaries.[9–11] To the best of our knowledge, this is the first study comparing OD-NLP with usual WD-NLP in Japanese medical records. We demonstrated that word segmentation into entities potentiates the extraction of characteristic terms of respective diseases.

We utilized topic models as an informatic classification to generate the description of documents to perform the comparison, although there are existing methods such as conditional random fields (CRFs), recurrent neural networks (RNNs), and clustering methods; K-Means.[28–30] Machine learning requires training datasets with largely defined annotations that require human processing in advance to obtain higher accuracy, while clustering methods classify information depending on the shallow information of the text. On the contrary, the topic model does not require annotated training data; thus, it classifies documents relating to the latent semantics.[31] Significantly, no studies have compared IRIS NLP Japanese with other NLPs much less to say nothing of the usefulness in the state-of-the-art models including CRF and RNN. Our topic models were created from identical information sources for comparison of

Table 3 Differentiation of filtering in IRIS NLP between TF-IDF and DMV

	Dataset	Min	0.25	Median	0.75	Max	*p*-Value
F-measure	iKTDF260	0.008	0.024	0.033	0.054	0.218	0.015
	iKDMV400	0.011	0.028	0.042	0.068	0.305	
	iKTDF330	0.019	0.036	0.048	0.078	0.196	0.214
	iKDMV500	0.018	0.041	0.056	0.089	0.229	
	iKTDF425	0.054	0.086	0.096	0.121	0.261	0.404
	iKDMV600	0.035	0.076	0.097	0.114	0.356	
	iKTDF510	0.111	0.130	0.154	0.214	0.250	0.317
	iKDMV700	0.061	0.125	0.138	0.178	0.457	
	iKTDF580	0.247	0.306	0.364	0.372	0.386	0.841
	iKDMV800	0.205	0.309	0.313	0.328	0.478	
	iKTDF620	0.645	0.660	0.676	0.691	0.706	1.000
	iKDMV900	0.647	0.657	0.666	0.676	0.686	

Abbreviations: DMV, dominance value; iK, iKnow; Me, MeCab; NLP, natural language processing; TF-IDF, term frequency-inverse document frequency.

extracted terms regardless of the type of NLP method. The approach enables comparison without prejudice, in which the NLP method extracts a characteristic sequence of entities/words from the existing medical records in the EHR systems.

In NLP comparison without any filtering, the number of entities in OD-NLP was fourfold higher than that of WD-NLP in common document sources of clinical texts. Moreover, 95 percentile value of document frequency and the maximum were everywhere in OD-NLP than in WD-NLP. These results suggest that OD-NLP segments a sentence into varied terms that were taken as various meaningful units, which enabled the generation of characteristic expressions from Japanese medical records. This capability allows entities to reveal features and, thus, potentially estimate diseases. On the contrary, WD-NLP becomes poor in variety because it extracts words that are basically in the term dictionary, and leads in dealing with a combination of simple words, when the dictionary does not have appropriate character sequences.

When the filtering was conducted with TF-IDF in each NLP step, the medians of F-measure in OD-NLP were all higher than those in WD-NLP, especially in lower thresholds of TF-IDF. These results are similar to other filtrations of DMV. Medians of F-measure in OD-NLP were higher in DMV, especially in the lower thresholds, while the differences disappeared when the threshold increased. These patterns suggest that OD-NLP includes a wide variety of entities, and it potentiates a higher classification accuracy. This is because the topic terms describing the characteristics of diseases remain when topics frequently appearing in common throughout the documents are mainly excluded by the method of lower filtrations of TF-IDF. In contrast to OD-NLP, WD-NLP recognizes a combination of determined words because the dictionary does not always have characteristic expressions of disease: a series of meaningful words are occasionally segmented into several words that are in the term dictionary. This is one of the biggest fundamental problems in WD-NLP, especially in written Japanese, because it is hard to follow many new words generated successfully. At higher thresholds of filtration, the frequency of characteristic expressions of respective diseases appears in a few documents in OD-NLP, thus, categorized as an appropriate topic, which results in no differences between OD-NLP and WD-NLP. We found that OD-NLP outperformed WD-NLP even when the filtering methods were changed from TF-IDF to DMV. The values of DMV were determined by the relationships between each entity in documents by calculating the frequency and the syntax analysis. Therefore, DMV decreases when sentences have many important words and a complicated syntax, by which entities with characteristic descriptions of disease are excluded in higher thresholds of DMV.

We confirmed that OD-NLP was able to extract important terms in documents because most evaluators believed that the topic terms were practically related to diseases, even with low classification accuracy. The levels of precision were relatively low for the following reasons: First, we set only default values as tuning hyperparameters in the LDA library (e.g., α, η, γ, the number of iterations) in addition to topic number.[32,33] The improvement in processing accuracy is achieved by setting additional parameters. However, we did not adjust this time because the improvement in prediction accuracy was out of our current scope. Second, we did not process the standardization of synonyms for the same reason and addressed them as different entities/words. The separated words or entities could have been replaced by unified terms by a published thesaurus or synonym dictionaries.[34,35] Third, generated topics from topic models were required to adopt the correspondence table depending on the characteristics, and were bound to the respective diseases. We performed the optimization by replacing

the position of the topic with the diagonal lines to maximize the sum of each score in the cross-tabulation. However, there were some diseases associated with several topics simultaneously, suggesting that topics predicted different diseases with similar probability. The above-mentioned circumstances were considerable reason of the low precision and some diseases may actually be one of the candidates for a differential diagnosis. The improvement of accuracy for estimating diseases by pretraining (e.g., RNN, CRF) would be further required after establishing the importance of IRIS NLP Japanese. Moreover, it would be necessary to gather longitudinal medical records for future validation because this research was a single-center study.

Notably, OD-NLP by IRIS NLP is not about cutting up sentences in the smallest possible tokens that carry meaning (e.g., words), but rather to split sentences into the tokens that carry an indivisible meaning for their use in a sentence.[36] Accordingly, even in Western languages that are ready for use IRIS NLP application (English, French, German and Dutch), marking word boundaries with clear separators such as spaces and markers, noun phrases like compound nouns, and concatenated nouns are not originally needed splitting and thus those pieces carry an indivisible meaning. Meanwhile, neither Japanese nor Chinese have any obvious word boundaries between characters, but actually they have some different features; Japanese is composed of a mixture of ideogram and phonogram, whereas Chinese is represented only with ideogram. That's reason why the character unit itself can be treated with a token that carries indivisible meaning and pretrained character embedding (e.g., word2-vec and BERT) is often applied before deep learning in Chinese.[37,38] Moreover, IRIS NLP Chinese has not been developed yet, and our current approach to the relating problems in Japanese NLP holds a peculiar aspect that cannot be simply applied to Chinese.

Conclusion

We investigated the performance of OD-NLP without a dictionary in medical records written in Japanese using a topic model, compared with WD-NLP. We proved that an alternative application of OD-NLP extracts more characteristic information for disease prediction than the conventional method. The current results suggest that entity recognition can be useful in Japanese EHR systems for document summarization, similar document retrieval, and clinical application, although further improvement in accuracy and the enlargement of application would be required.

Authors' Contribution

Y.Y. designed the study and retrieved a series of data under the supervision by T.N. and K.U. Y.Y. and T.I. performed a statistical analysis of collected data. T.N. and K.U. helped make the overall study design and conceptualized interpretations and criticisms. Y.Y. and T.N. drafted the manuscript and it was critically revised by T.N., T.I., and K.U. The final manuscript was approved by all authors.

Funding

This study was supported by Bristol-Myers Squibb Foundation Grants Number 41762123 (to TN) and an endowment fund of the department – Project Number 147100001k.

Conflict of Interest

None declared.

Acknowledgment

IRIS NLP demonstration environment was generously provided by InterSystems Japan Co. We received technical supports from Datacube Co., Ltd. for the use of IRIS NLP and the system development. We would like to thank Editage (www.editage.com) for English language editing.

References

1 Industry JAHIS. Research on the implementation of medical information systems (order entry and electronic medical record systems). Accessed March 6, 2023 at: https://www.jahis.-jp/action/id=57?contents_type=23
2 Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3
3 Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016;23(05):1007–1015
4 Pham AD, Névéol A, Lavergne T, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics 2014;15(01):266
5 Wilke RA, Xu H, Denny JC, et al. The emerging role of electronic medical records in pharmacogenomics. Clin Pharmacol Ther 2011;89(03):379–386
6 Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;17(01):128–144
7 Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009;42(05):760–772
8 Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011;18(05):544–551
9 Nishimoto N, Terae S, Uesugi M, Ogasawara K, Sakurai T. Development of a medical-text parsing algorithm based on character adjacent probability distribution for Japanese radiology reports. Methods Inf Med 2008;47(06):513–521
10 Ahltorp M, Skeppstedt M, Kitajima S, Henriksson A, Rzepka R, Araki K. Expansion of medical vocabularies using distributional semantics on Japanese patient blogs. J Biomed Semantics 2016;7(01):58
11 Usui M, Aramaki E, Iwao T, Wakamiya S, Sakamoto T, Mochizuki M. Extraction and standardization of patient complaints from electronic medication histories for pharmacovigilance: natural language processing analysis in Japanese. JMIR Med Inform 2018;6(03):e11021
12 Aramaki E, Yano K, Wakamiya S. MedEx/J: A one-scan simple and fast NLP tool for Japanese clinical texts. Stud Health Technol Inform 2017;245:285–288
13 Jiang G, Ogasawara K, Endoh A, Sakurai T. Context-based ontology building support in clinical domains using formal concept analysis. Int J Med Inform 2003;71(01):71–81

14 Suzuki T, Yokoi H, Fujita S, Takabayashi K. Automatic DPC code selection from electronic medical records: text mining trial of discharge summary. Methods Inf Med 2008;47(06):541–548
15 Li Y, Wang X, Hui L, et al. Chinese clinical named entity recognition in electronic medical records: development of a lattice long short-term memory model with contextualized character representations. JMIR Med Inform 2020;8(09):e19848
16 Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. J Am Med Inform Assoc 2014;21(05):808–814
17 Hu D, Huang Z, Chan TM, Dong W, Lu X, Duan H. Utilizing Chinese admission records for MACE prediction of acute coronary syndrome. Int J Environ Res Public Health 2016;13(09):912
18 Fushimi K, Hashimoto H, Imanaka Y, et al. Functional mapping of hospitals by diagnosis-dominant case-mix analysis. BMC Health Serv Res 2007;7:50
19 Bronselaer A, Tré GD. Concept-relational text clustering. Int J Intell Syst 2012;27(11):970–993
20 Corporation I. iKnow. Accessed March 6, 2023 at: https://github.com/intersystems/iknow
21 MeCab. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. Accessed March 6, 2023 at: http://taku910.github.io/mecab/
22 Liu J, Shindo H, Matsumoto Y. Development of a computer-assisted Japanese functional expression learning system for Chinese-speaking learners. Educ Technol Res Dev 2019;67(05):1307–1331
23 Ujiie S, Yada S, Wakamiya S, Aramaki E. Identification of adverse drug event-related Japanese articles: natural language processing analysis. JMIR Med Inform 2020;8(11):e22661
24 Aoki M, Yokota S, Kagawa R, Shinohara E, Imai T, Ohe K. Automatic classification of electronic nursing narrative records based on Japanese standard terminology for nursing. Comput Inform Nurs 2021;39(11):828–834
25 Sagara K. ComeJisyo. Accessed March 6, 2023 at: https://ja.osdn.net/projects/comedic/
26 Aizawa A. An information-theoretic perspective of TF-IDF measures. Inf Process Manage 2003;39(01):45–65
27 Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research 2003;3(04–05):993–1022
28 Luo G. MLBCD: a machine learning tool for big clinical data. Health Inf Sci Syst 2015;3:3
29 Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85–117
30 Haider MM, Hossin MA, Mahl HR, Arif H. Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm. Ieee Region 10 Symp 2020:283–286
31 Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. Springerplus 2016;5(01):1608
32 Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T. Public discourse and sentiment during the COVID 19 pandemic: using latent Dirichlet Allocation for topic modeling on Twitter. PLoS One 2020;15(09):e0239441
33 Wang H, Wu F, Lu W, et al. Identifying objective and subjective words via topic modeling. IEEE Trans Neural Netw Learn Syst 2018;29(03):718–730
34 Torii M, Yang EW, Doan S. A Preliminary Study of Clinical Concept Detection Using Syntactic Relations. AMIA Annu Symp Proc 2018;2018:1028–1035
35 Henriksson A, Moen H, Skeppstedt M, Daudaravičius V, Duneld M. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. J Biomed Semantics 2014;5(01):6
36 Hazewinkel MC, de Winter RFP, van Est RW, et al. Text analysis of electronic medical records to predict seclusion in psychiatric wards: proof of concept. Front Psychiatry 2019;10:188
37 Wang Q, Zhou YM, Ruan T, Gao DQ, Xia YH, He P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. Journal of Biomedical Informatics 2019;92
38 Li LQ, Zhao J, Hou L, Zhai YK, Shi JM, Cui FF. An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. Bmc Med Inform Decis 2019;19(Suppl 5):235