Semi-automatic Construction of a Word Complexity Lexicon for Japanese Medical Terminology

Soichiro Sugihara[1] **Tomoyuki Kajiwara**[1] **Takashi Ninomiya**[1]
Shoko Wakamiya[2] **Eiji Aramaki**[2]
[1]Ehime University {sugihara@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp
[2]Nara Institute of Science and Technology {wakamiya, aramaki}@is.naist.jp

Abstract

We construct a word complexity lexicon for medical terms in Japanese. To facilitate communication between medical practitioners and patients, medical text simplification is being studied. Medical text simplification is a natural language processing task that paraphrases complex technical terms into expressions that patients can understand. However, in contrast to English, where this task is being actively studied, there are insufficient language resources in Japanese. As a first step in advancing research on medical text simplification in Japanese, we annotate the 370,000 words from a large-scale medical terminology lexicon with a five-point scale of complexity for patients.

1 Introduction

Communication between medical practitioners and patients is important to facilitate understanding of the diagnosis and agreement on a treatment plan (Ha and Longnecker, 2010). One of the factors that make communication difficult in the medical field is the difference in expertise between medical practitioners and patients. In particular, since many medical terms are difficult for patients to understand, medical practitioners are expected to paraphrase them into simple expressions to make them easier to understand.

To solve this problem, medical text simplification (Leroy and Endicott, 2012; Joseph et al., 2023; Yang et al., 2023) has been studied, mainly in English. However, there is a lack of available lexicons and corpora for medical text simplification in Japanese. In this study, as a first step to tackle Japanese medical text simplification, we construct a complexity lexicon for medical terms.

We first recruited 40 annotators, who were not medical practitioners via crowdsourcing to survey word complexity for 10,000 medical terms. As a result, we found that the number of unknown medical terms decreased with age and that men tended to be unaware of medical terms related to pregnancy and childbirth, among other characteristics observed for each of the attributes of the annotators. Furthermore, we trained a complexity estimation model for medical terms using machine learning with features such as character types, word frequencies, and word embeddings, and achieved higher performance than existing methods. Finally, as shown in Table 1, we estimated the word complexity for 370,000 disease names and symptom expressions from a large-scale medical terminology lexicon in Japanese[1] (Ito et al., 2018). Our word complexity lexicon will be available[2] upon publication of this paper.

Complexity	Medical Terminology
1 (Simple)	めまい (Dizzy)
2	感電死 (Electrocution)
3	若年性脱毛症 (Premature Alopecia)
4	後天性てんかん (Acquired Epilepsy)
5 (Complex)	掌蹠膿疱症性骨関節炎 (Pustulotic Arthro-Osteitis)

Table 1: Examples of Japanese medical terminology.

2 Related Work

Large-scale word complexity lexicons in English have been constructed using two approaches. One is to estimate word complexity using the log ratio of the probability of word occurrence in the normal and simple corpora (Pavlick and Nenkova, 2015). The other is to manually annotate word complexity for a subset of the vocabulary and train a word complexity estimation model using these annotations (Pavlick and Callison-Burch, 2016;

[1] https://sociocom.naist.jp/manbyou-dic/

[2] https://github.com/EhimeNLP/J-MeDic-Complexity

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 329–333
June 21, 2024 ©2024 Association for Computational Linguistics

Maddela and Xu, 2018). In Japanese, the former approach cannot be applied because of the unavailability of a large-scale corpus written in simple language. Therefore, this study takes the latter approach to construct a word complexity lexicon.

In Japanese, a domain-independent word complexity estimation model has been proposed that employs character types, word frequencies, and word embeddings as features (Kajiwara et al., 2020). For word complexity estimation specific to the medical domain, a method that takes into account the number of characters and morphemes has been proposed (Yamamoto et al., 2019). Similar to these previous studies, we train a machine learning-based word complexity estimator.

3 Word Complexity Annotation

3.1 Crowdsourcing

To train the word complexity estimation model, we asked non-medical practitioners to annotate the complexity of medical terms. These medical terms are 10,000 terms randomly selected from the top 30,000 terms with the most reliable terminology in a large-scale lexicon of disease names in Japanese[1] (Ito et al., 2018).

For diversity of annotators, eight groups were formed based on a combination of age (20s, 30s, 40s, and 50s) and gender (male and female), with five annotators per group, for a total of 40 annotators recruited. For the crowdsourcing service, we used Lancers[3] and paid the annotators 1 JPY per word (1,000 JPY per hour).

The annotators assigned each word the following a five-point scale of complexity.

1. I use this term in my daily conversation.
2. I have used this terminology.
3. I can understand what this term means.
4. I have seen or heard this term but do not know what it means.
5. I do not know what this term means and have never seen or heard of it.

To improve quality, two levels of filtering were applied to the annotators. First, we requested a small annotation of 300 words. We reviewed the responses and asked only those who had no problems to annotate the remaining 9,700 words. In addition, after all 10,000 words were annotated, inter-annotator agreement was calculated for each group of age and gender. Annotators with a Quadratic Weighted Kappa (QWK) (Cohen, 1968) of less than 0.3 with someone in the group were excluded and new annotators were recruited.

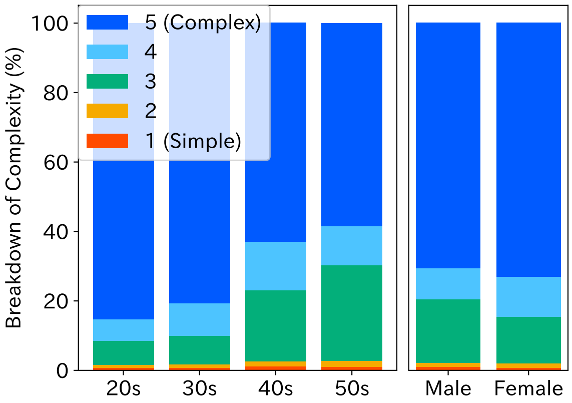

Figure 1: Distribution of complexity by age and gender.

3.2 Analysis

We analyze characteristics by age and gender based on our complexity annotations. Figure 1 shows the distribution of complexity labels by age and gender. In their 20s and 30s, only about 10% of medical terms are understood. As they get older, the number of medical terms they don't know decreases. However, even in their 50s, more than 70% of medical terms cannot be understood.

Next, we observe examples of medical terms that are known above a certain age. All annotators know "しゃっくり" (hiccups) and "かぜ" (cold) used in daily conversation, while only annotators in their 40s or older or 50s know "食道ポリープ" (esophageal polyp) and "大腿骨骨折" (femur fracture) which tend to increase in patients as they get older. These imply that our complexity annotations reflect age-specific characteristics.

Finally, we observe examples of medical terms that certain groups do not know. Young men in their 30s and younger seem to be unfamiliar with some of the medical terms related to pregnancy and childbirth, such as "異常胎位" (abnormal fetal presentation) and "早発卵巣不全" (premature ovarian failure). These imply that our complexity annotations reflect gender-specific characteristics.

4 Word Complexity Estimation

We train a machine learning-based word complexity estimation model in addition to the three ba-

[3] https://www.lancers.jp

sic features used in the previous study (Yamamoto et al., 2019), with three proposed features. As in previous studies (Yamamoto et al., 2019; Kajiwara et al., 2020), we use the support vector machine (SVM) model[4] for machine learning.[5]

4.1 Basic Features

Character Types These features represent the types of characters (hiragana, katakana, kanji, numbers, and alphabetic characters) that make up a medical term. It consists of the following 15 dimensions: binary features (5 dimensions) that represent the presence or absence of each character type, integer features (5 dimensions) that represent the number of characters for each character type, and integer features (5 dimensions) that represent the maximum number of consecutive characters for each character type.

Number of Morphemes This is one-dimensional integer feature that represents how many morphemes a medical term is composed of. Medical terms are tokenized with MeCab[6] (IPADIC) (Kudo et al., 2004) and the number of morphemes is counted.

Character/Morpheme Frequencies These features are the frequencies of the letters and morphemes that make up the medical term in the corpus. Six types of frequency information are used as the features: the total, average, maximum, and minimum frequencies of morphemes in the medical term, as well as the frequency of the first morpheme and the frequency of the last morpheme. Japanese Wikipedia was used as the corpus, and MeCab was used as the morphological analyzer. Note that frequencies are used logarithmically, but as in previous study (Yamamoto et al., 2019), when the frequency is 0, 0 is used instead of log 0. These features are obtained not only in morpheme units but also in character units, for a total of a 12-dimensional real number of features.

4.2 Proposed Features

PF1: Frequencies on Web Corpus We count frequencies of characters and morphemes similar to basic features on the CC-100[7] (Conneau et al., 2020), a large-scale Web corpus. These are 12-dimensional real number of features, same as the basic features. Counting frequencies on multiple corpora is known to contribute to the word complexity estimation (Kajiwara and Komachi, 2018). However, as mentioned earlier, this study does not use the Twittr and the BCCWJ corpora used in previous study (Yamamoto et al., 2019), so a large-scale Web corpus is employed instead.

PF2: Word Frequencies In contrast to previous study (Yamamoto et al., 2019), we also count the frequency of medical terms in word units without segmentation. This is implemented by extending MeCab's morphological analysis with a Japanese disease lexicon[8] (Ito et al., 2018). We count word frequencies in each of the Wikipedia and CC-100 corpora, logarithmize them, and use them as two-dimensional real number features.

PF3: Word Embeddings We also employ word embeddings, which has been used in previous study (Kajiwara et al., 2020). We use pre-trained fastText[9] (Bojanowski et al., 2017). If a medical term consists of multiple morphemes, each of those vectors is averaged and used as a 300-dimensional real number of features.

5 Experiments and Results

We train and evaluate word complexity estimation models using complexity annotations for 10,000 medical terms.

5.1 Experiments

Dataset We average the complexity labels obtained from 40 annotators and round them to integers to define a five-point scale of gold complexity labels for 10,000 medical terms. Since this task is an ordinal classification, we use accuracy and QWK (Cohen, 1968) as evaluation metrics. As shown in Table 2, the training and evaluation dataset were randomly split at a ratio of 9:1 for our experiments. Since our dataset is unbalanced, we

[4] We also experimented with neural networks, but the SVM model achieved higher performance.

[5] As one of the features, previous study (Yamamoto et al., 2019) employed word frequencies counted on Twitter. However, we do not use this feature because changes in Twitter's API restrictions have made this counting difficult. Furthermore, word frequencies from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2010) are not used in this study, since previous study (Yamamoto et al., 2019) reported that these word frequencies were not effective.

[6] https://taku910.github.io/mecab/

[7] https://data.statmt.org/cc-100/

[8] https://sociocom.naist.jp/j-meddic-for-mecab/

[9] https://fasttext.cc/docs/en/crawl-vectors.html

Labels	1	2	3	4	5	Total
Train	33	100	341	2,650	5,876	9,000
Test	4	11	38	294	653	1,000
Total	37	111	379	2,944	6,529	10,000

Table 2: Number of terms per complexity.

adjusted the label ratios in both training and evaluation datasets to be equal by stratified splitting.[10]

Model For word complexity estimation model, a multi-class classification model was implemented using SVM (RBF kernel) in scikit-learn (1.3.2)[11] (Pedregosa et al., 2011). The hyperparameters C and gamma were selected from $\{1, 5, 10, 50, 100\}$ and $\{0.0001, 0.0005, 0.001, 0.05, 0.1\}$, respectively, and the combination with the highest QWK was selected by grid search with a five-fold cross-validation.[12] The features were standardized.[13]

Comparative Methods We compare the proposed method to two types of baselines. One is a simple baseline that always outputs the most frequent class, label 5. The other is a baseline that uses only the basic features of Section 4.1, which replicates the previous study (Yamamoto et al., 2019). Our method uses the proposed features of Section 4.2 in addition to the basic features.

5.2 Results

Table 3 shows the experimental results. Existing method using only basic features does not perform well enough, as it is equivalent in accuracy to a baseline that always outputs the most frequent labels. The proposed method significantly improved performance over these baselines by 14 points in accuracy and 28 points in QWK.

To clarify the effectiveness of each of the proposed features, an ablation analysis was performed to remove one of the proposed features from the proposed method. The fact that both accuracy and QWK decrease when any of the features are excluded shows that all of our proposed features are useful. Note that the performance decreases significantly when PF3 is excluded, suggesting that word embeddings are a particularly important feature. When each of the proposed features was used alone, PF1 alone outperformed the baselines, revealing that frequency features on a large-scale Web corpus are also useful for estimating the complexity of medical terminology.

	Accuracy	QWK
Baseline	0.653	-
Basic features	0.653	0.456
Proposed method	**0.793**	**0.732**
Proposed method w/o PF1	0.782	0.729
Proposed method w/o PF2	0.785	0.695
Proposed method w/o PF3	0.718	0.612
Only PF1	0.658	0.483
Only PF2	0.612	0.444
Only PF3	0.768	0.660

Table 3: Experimental results of word complexity estimation.

6 Conclusion

In this study, we trained a word complexity estimation model based on word complexity annotations of 10,000 Japanese medical terms by 40 non-medical practitioners. Our word complexity annotations revealed that even though the number of unknown medical terms decreases with increasing age, more than 70% of medical terms are difficult to understand, even for those in their 50s. Experiments on word complexity estimation revealed that features of word frequencies and word embeddings obtained from a large-scale Web corpus are useful. Finally, we developed a word complexity estimator for Japanese medical terms that can classify five levels of complexity with about 80% accuracy, and released a word complexity lexicon[2] covering about 370,000 Japanese medical terms.

Although this study focused on disease and symptom names in Japanese, our future work includes the application of complexity estimation to more diverse medical terminology, such as drug names and names of human body parts. Note that the "word complexity" in this study was judged by the patients themselves. Even if the patients themselves consider it to be simple, it is possible that medical misunderstandings may have occurred.

[10] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[11] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[12] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[13] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Acknowledgements

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *TACL*, 5:135–146.

Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jennifer Fong Ha and Nancy Longnecker. 2010. Doctor-Patient Communication: A Review. *Ochsner Journal*, 10(1):38–43.

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, and Eiji Aramaki. 2018. J-MeDic: A Japanese Disease Name Dictionary Based on Real Clinical Usage. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 2365–2369.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. Multilingual Simplification of Medical Texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Complex Word Identification Based on Frequency in a Learner Corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199.

Tomoyuki Kajiwara, Daiki Nisihara, Tomonori Kodaira, and Mamoru Komachi. 2020. Language Resources for Japanese Lexical Simplification. *Journal of natural language processing*, 27(4):189–210.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Gondy Leroy and James E. Endicott. 2012. Combining NLP with Evidence-Based Methods to Find Text Metrics Related to Perceived and Actual Text Difficulty. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 749–754.

Mounica Maddela and Wei Xu. 2018. A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 1483–1486.

Ellie Pavlick and Chris Callison-Burch. 2016. Simple PPDB: A Paraphrase Database for Simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 143–148.

Ellie Pavlick and Ani Nenkova. 2015. Inducing Lexical Style Properties for Paraphrase and Genre Differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Hideya Yamamoto, Kaoru Ito, and Eiji Aramaki. 2019. Fukugougo no Kouseiso Jouhou wo Kouryo Shita Byoumei Nannido no Suitei (Estimation Methods for Medical Term's Difficulty Utilizing Information on Constituents). In *Proceedings of the 25th Association for Natural Language Processing*, pages 1495–1498. (in Japanese).

Ziyu Yang, Santhosh Cherian, and Slobodan Vucetic. 2023. Data Augmentation for Radiology Report Simplification. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1922–1932.