Automated Orthodontic Diagnosis from a Summary of Medical Findings

Takumi Ohtsuka[†], **Tomoyuki Kajiwara**[†], **Chihiro Tanikawa**[‡]
Yuujin Shimizu[‡], **Hajime Nagahara**[‡], **Takashi Ninomiya**[†]
[†]Ehime University [‡]Osaka University
{ohtsuka@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp
{ctanika@dent, yjnshimizu@dent, nagahara@ids}.osaka-u.ac.jp

Abstract

We propose a method to automate orthodontic diagnosis with natural language processing. It is worthwhile to assist dentists with such technology to prevent errors by inexperienced dentists and to reduce the workload of experienced ones. However, text length and style inconsistencies in medical findings make an automated orthodontic diagnosis with deep-learning models difficult. In this study, we improve the performance of automatic diagnosis utilizing short summaries of medical findings written in a consistent style by experienced dentists. Experimental results on 970 Japanese medical findings show that summarization consistently improves the performance of various machine learning models for automated orthodontic diagnosis. Although BERT is the model that gains the most performance with the proposed method, the convolutional neural network achieved the best performance.

1 Introduction

To make a proper orthodontic diagnosis, dentists need a wealth of knowledge and experience. Therefore, inexperienced dentists may overlook patient problems. Artificial intelligence technologies, such as automatic diagnosis, are promising for preventing such errors by dentists (Shimizu et al., 2022). Even for experienced dentists, automatic diagnosis technology can contribute in terms of workload reduction and improved efficiency. Therefore, this study focuses on automatic diagnosis from medical findings texts written by dentists.

While computer vision technologies for orthodontic applications, such as landmark identification from cephalometric X-rays (Kunz et al., 2020) and tooth segmentation on 3D dental surfaces captured by intraoral scanners (Lian et al., 2019), have been actively studied, research on natural language processing (NLP) technologies in this area has been limited. The only previous work (Shimizu et al., 2022) applying NLP to automatic diagnosis from medical findings relied on bag-of-words (BoW) feature extraction and support vector machine (SVM) classification without the benefit of deep-learning (DL), which has been successfully applied to a variety of tasks in recent years. We hypothesize that this is due to the frequent use of technical terms not covered by even powerful pre-trained models, as well as the long documents in which the medical findings are written in an inconsistent style. Specifically, the medical findings in this study average 1,886 tokens, with a maximum of 6,379 tokens, and contain many incomplete sentences, such as bullet points.

To solve the problems of document length and style inconsistency, we utilize a short summary of medical findings written by dental specialists for automatic diagnosis. In contrast to the original medical findings, our summary is about 90% shorter (179 tokens on average) and consists only of complete sentences. With these advantages, summary text facilitates feature extraction from documents by encoders in DL models.

To evaluate the effectiveness of automatic diagnosis from a summary of medical findings, we experimented with 970 Japanese medical findings. Experimental results on DL models of recurrent neural networks (RNN), convolutional neural networks (CNN), self-attention networks (SAN), and BERT (Devlin et al., 2019), a pre-trained SAN with masked language modeling objectives, showed that the CNN model achieved the best performance. Furthermore, the performance of the SVM and DL models was consistently improved when utilizing the summaries compared to the original medical findings. In particular, BERT was the best performance improvement with the proposed method.

2 Related Work

With the development of DL technologies, many medical applications including the field of orthodontics are being addressed.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 156–160

2.1 Medical Applications of NLP

One of the medical applications of NLP is automated question-answering (QA) (Nguyen, 2019). emrQA[1] (Pampari et al., 2018) and emrKBQA[2] (Raghavan et al., 2021) are large-scale corpora automatically created from English electronic medical records for QA in the clinical domain.

The MedWeb task in NTCIR-13[3] (Wakamiya et al., 2017) targeted user-generated text on social media, which is more accessible than electronic medical records. This competition addressed disease classification in three languages: English, Chinese, and Japanese.

In recent years, medical language processing related to COVID, a worldwide epidemic, has also been actively studied. Examples include COVID-QA[4] (Möller et al., 2020) for question answering, COVID-Q[5] (Wei et al., 2020) for question classification, and COVID-19 Real World Worry Dataset[6] (Kleinberg et al., 2020) for emotional analysis.

As described above, medical applications of NLP are being studied in a variety of languages and tasks. However, there are few efforts to apply NLP in the field of orthodontics. In particular, there is no application of DL-based NLP other than this study.

2.2 Orthodontic Applications of DL Models

Many applications of deep learning models in the field of orthodontics are computer vision technologies. Kunz et al. (2020) utilized CNN models to identify landmarks in cephalometric X-rays. They reported that training with 1792 images resulted in CNN models achieving nearly the same quality as experienced examiners. Lian et al. (2019) proposed MeshSNet, which performs tooth segmentation on 3D dental surfaces captured by intraoral scanners.

While DL-based computer vision models have been actively applied to the field of orthodontics, there is no previous study of DL-based NLP. The only previous work applying NLP to the field of orthodontics (Shimizu et al., 2022) has addressed automatic diagnosis from medical findings. They perform automatic diagnosis by feature extraction with BoW and classification with SVM, and do not benefit from recent DL technologies.

[1] https://github.com/panushri25/emrQA/

[2] https://github.com/emrQA/emrKBQA

[3] http://research.nii.ac.jp/ntcir/permission/ntcir-13/perm-en-MedWeb.html

[4] https://github.com/deepset-ai/COVID-QA

[5] https://github.com/JerryWei03/COVID-Q

[6] https://github.com/ben-aaron188/covid19worry

3 Proposed Method

We utilize deep learning-based document classification models for automatic diagnosis from medical findings in the field of orthodontics. We first describe these models in Section 3.1. Since medical findings are difficult to classify as they are, our proposed method instead utilizes a short summary of them, written in a consistent style. This is explained in Section 3.2.

3.1 DL-based Document Classification Models

In recent years, DL-based models have been widely used in NLP tasks, including document classification. In this study, we apply four types of DL models, including RNN, CNN, SAN, and BERT (Devlin et al., 2019), for automatic diagnosis from medical findings in the field of orthodontics.

Recurrent Neural Network (RNN) is one of the neural networks that deal with time-series data by recursively processing input data. In NLP, sentences are segmented into tokens as a preprocessing step, and the RNN processes the tokens in order from the beginning of the sentence. Since the original RNN is not good at long-term memory, extensions to BiLSTM, which uses LSTM cells and receives sentences in both directions, have improved performance on many tasks such as pronoun prediction (Stymne et al., 2017) and dependency parsing (Falenska and Kuhn, 2019). Nevertheless, it is difficult to achieve high performance for a very long series exceeding 1,000 tokens (Li et al., 2018). This study also employs the BiLSTM model as an RNN.

Convolutional Neural Network (CNN) is one of the neural networks that utilizes convolutional filters and pooling layers to extract features from input data as regions rather than points. While this is a model typically used for computer vision, it is also known to be effective in NLP, such as text classification (Kim, 2014). Instead of convolving n neighboring pixels as a region in computer vision, CNN acquires an n-gram representation by convolving n continuous tokens in NLP.

Self Attention Network (SAN) is another neural network that deals with series by learning contextualized token embeddings instead of aggregating

information like RNN and CNN. It was originally proposed as an encoder-decoder neural network for machine translation (Vaswani et al., 2017), but is also used for text classification with an encoder only. BERT (Devlin et al., 2019), which pre-trained SAN model for the objective of masked language modeling with a large-scale corpus, has remarkable performance on a number of NLP tasks through fine-tuning on the target task. Furthermore, models such as SciBERT (Beltagy et al., 2019) in the scientific domain and ClinicalBERT (Alsentzer et al., 2019) in the medical domain, which are pre-trained to focus on the desired domain, achieve even higher performance on specific tasks. Unfortunately, there are no pre-trained masked language models specific to the orthodontic domain, thus we utilize a SAN model that is trained from scratch and a general-purpose BERT that is pre-trained on Wikipedia. Note that BERT is limited to a maximum of 512 input tokens to balance memory usage and performance.

3.2 Utilizing a Summary of Medical Findings

Although the DL models described in the previous section are widely used in recent NLP tasks, it is difficult to handle texts longer than 1,000 tokens due to the difficulty of learning extreme long-term dependencies and limitations in memory usage. The medical findings in the field of orthodontics that we deal with in this study are very long documents, with an average of 1,886 tokens and a maximum of 6,379 tokens, as shown in Table 1. This is too long a text to be handled by BiLSTM or BERT.

Not only do DL models suffer from the text length, but also from inconsistencies in the writing style of the medical findings. These writing styles vary for each dentist who writes. Writing style issues include the use of incomplete sentences with bullets and indentation with spaces and tabs. While they improve visual clarity for human readers, they are noise to NLP models because such information is removed in pre-processing steps in many cases. Especially for pre-trained models such as BERT, these incomplete sentences, with different characteristics from the pre-training corpus, may seriously impair performance.

To address these problems, we propose to utilize its short summary with a consistent writing style instead of the original medical findings. These summaries are manually written by experienced dentists and are written in complete sentences without the use of bullets, indentations, or other decorations. As shown in Table 1, these summaries consist of an average of 179 tokens, that is, about 10% of the length of the original medical findings. Furthermore, even the longest summaries are not affected by the limit on the maximum length of input tokens in BERT. We assigned these summaries to all medical findings in our dataset. Compared to the noisy and lengthy original medical findings, these summaries are expected to improve the performance of DL-based document classification models.

	Original	Summary
Avg.	1,886	179
Max.	6,379	467
Min.	312	61

Table 1: Number of tokens for each document.

4 Evaluation

To evaluate the effectiveness of the proposed method, automatic diagnosis is performed from the original medical findings or a summary of them, and their performance is compared. We treat this task as a multi-label document classification.

4.1 Setting

In our experiment, we use documents of medical findings in Japanese for 970 patients who visited for orthodontic treatment. This dataset includes the text of the medical findings written by the dentist in charge, as well as the patient's facial and X-ray images. However, utilizing these images remains our future work, and we only use text in this study. We have assigned a short summary, described in Section 3.2, to every medical finding in this dataset. Each medical finding is also assigned multiple labels corresponding to the patient's medical condition. There are a total of 322 labels, with each patient having an average of 12 labels.

As text preprocessing, line feed characters were removed and full-width alphanumeric characters were normalized to half-width. We used Sudachi[7] (Takaoka et al., 2018) for word segmentation, except for the BERT model, for which a specific subword segmenter is provided. For evaluation, we used 5-fold cross-validation. The evalua-

[7] https://github.com/WorksApplications/SudachiPy

	RNN	CNN	SAN	BERT
Number of dimensions of embedding layer	256	256	256	768
Number of dimensions of hidden layers	256	256	512	768
Number of hidden layers	1	1	2	12
Dropout rate	0.2	0.2	0.1	0.1
Batch size	64	64	16	32

Table 2: Hyperparameters of deep learning models.

	Type of medical findings	
Models	Original	Summary
BoW+SVM	0.41	0.46
RNN	0.20	0.31
CNN	**0.44**	**0.48**
SAN	0.29	0.38
BERT	0.27	0.43

Table 3: Experimental results (F1-score).

tion metric used was the F1-score.

For the document classification model, we evaluate four DL models, including RNN, CNN, SAN, and BERT[8] (Devlin et al., 2019), described in Section 3.1, as well as SVM used in the previous work (Shimizu et al., 2022). The baseline model, denoted as BoW+SVM, employs the Binary Relevance method (Tsoumakas and Katakis, 2007) to train a binary classification for each label and utilizes the RBF kernel for SVM.[9] Our DL models use Adam (Kingma and Ba, 2015) as an optimizer. Other hyperparameters are listed in Table 2.

4.2 Result

Table 3 shows the experimental results. Deep learning models other than CNN suffer from document length and style inconsistencies, resulting in significantly poorer performance than the existing model of BoW+SVM.

When a short document summarized by the dentist is used in place of the original medical findings, the F1-scores for all models consistently improve. Notably, the performance of the pre-trained BERT has improved the most substantially. We believe this is due to the use of complete sentences that are consistent with the pre-training corpus and the elimination of information lost owing to the constraint of the maximum sentence length. These experimental results show that short summaries written in a consistent style are effective in improving the performance of automatic diagnosis in the field of orthodontics.

We found that the CNN model achieved the best performance for both the original document and summary inputs. Since medical findings often contain technical terms consisting of multiple tokens, we believe that a CNN model capable of capturing n-gram features through convolution would be suitable for this task.

5 Conclusion

In this study, we improved the performance of automatic diagnosis in orthodontic treatment by utilizing a short document that was manually summarized from medical findings by dentists. Experimental results on Japanese datasets show that the proposed method consistently improves the performance of various DL models. Among them, our CNN model outperformed the existing model and updated the state-of-the-art performance.

Our future work includes the automatic generation of summaries and the development of multimodal automatic diagnosis taking into account image information. Although this study utilized summaries of medical findings manually generated by experienced dentists, there is a substantial cost to creating such a dataset. It is desirable to develop an automatic diagnostic system that reduces the workload on dentists by automatically generating summaries. In addition, our dataset includes both facial and X-ray images. This allows us to develop multimodal models that incorporate findings from the field of computer vision, which are actively studied. Multimodal automatic diagnostic systems that combine both image and linguistic information in a complementary manner are expected to have higher performance.

[8] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

[9] https://scikit-learn.org/

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Agnieszka Falenska and Jonas Kuhn. 2019. The (Non-)Utility of Structural Features in BiLSTM-based Dependency Parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 117–128.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring Emotions in the COVID-19 Real World Worry Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Felix Kunz, Angelika Stellzig-Eisenhauer, Florian Zeman, and Julian Boldt. 2020. Artificial Intelligence in Orthodontics. *Journal of Orofacial Orthopedics / Fortschritte der Kieferorthopädie*, 81(1):52–68.

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. 2018. Independently Recurrent Neural Network (IndRNN): Building a Longer and Deeper RNN. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*.

Chunfeng Lian, Li Wang, Tai-Hsien Wu, Mingxia Liu, Francisca Durán, Ching-Chang Ko, and Dinggang Shen. 2019. MeshSNet: Deep Multi-scale Mesh Feature Learning for End-to-end Tooth Labeling on 3d Dental Surfaces. In *Proceedings of the 22th International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 837–845.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A Question Answering Dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Vincent Nguyen. 2019. Question Answering in the Biomedical Domain. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 54–63.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368.

Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. 2021. emrKBQA: A Clinical Knowledge-base Question Answering Dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73.

Yuujin Shimizu, Chihiro Tanikawa, Tomoyuki Kajiwara, Hajime Nagahara, and Takashi Yamashiro. 2022. The Validation of Orthodontic Artificial Intelligence Systems That Perform Orthodontic Diagnoses and Treatment Planning. *European Journal of Orthodontics*, 44(4):436–444.

Sara Stymne, Sharid Loáiciga, and Fabienne Cap. 2017. A BiLSTM-based System for Cross-lingual Pronoun Prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 47–53.

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. 2018. Sudachi: A Japanese Tokenizer for Business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 2246–2249.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000—-6010.

Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2017. Overview of the NTCIR-13 MedWeb Task. In *In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, pages 40–49.

Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. 2020. What Are People Asking about COVID-19? A Question Classification Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.