Improving Automatic KCD Coding: Introducing the KoDAK and an Optimized Tokenization Method for Korean Clinical Documents

Geunyeong Jeong[1] **Juoh Sun**[1] **Seokwon Jeong**[2] **Hyunjin Shin**[3,4] **Harksoo Kim**[1]
[1]Konkuk University [2]Kangwon National University
[3]Konkuk University Medical Center [4]Konkuk University School of Medicine
{jyjg7218, qssz1326}@konkuk.ac.kr nlpsw@kangwon.ac.kr
shineye@kuh.ac.kr nlpdrkim@konkuk.ac.kr

Abstract

International Classification of Diseases (ICD) coding is the task of assigning a patient's electronic health records into standardized codes, which is crucial for enhancing medical services and reducing healthcare costs. In Korea, automatic Korean Standard Classification of Diseases (KCD) coding has been hindered by limited resources, differences in ICD systems, and language-specific characteristics. Therefore, we construct the **Ko**rean **D**ataset for **A**utomatic **K**CD coding (**KoDAK**) by collecting and preprocessing Korean clinical documents. In addition, we propose a tokenization method optimized for Korean clinical documents. Our experiments show that our proposed method outperforms Korean Medical BERT (KM-BERT) in Macro-F1 performance by 0.14%p while using fewer model parameters, demonstrating its effectiveness in Korean clinical documents.

1 Introduction

International Classification of Diseases (ICD) coding is the assigning of standardized codes from the ICD system to patients' electronic health records. This is essential for standardizing information across medical institutions, and it serves as the foundation for the analysis of medical statistics. Traditionally, professional coders performed this task; however, this approach was costly and error-prone(O'malley et al., 2005; Lang, 2007). Consequently, recent studies have been exploring the application of deep learning in automatic ICD coding(Xie and Xing, 2018; Mullenbach et al., 2018; Zhang et al., 2022).

However, there are noticeable gaps in the research on automatic ICD coding across countries for various reasons. One reason is the difference in ICD systems between nations. Each country has a unique medical environment and requirements, with the result that many countries use modified ICD systems(Alharbi et al., 2019; Harrison et al., 2021; Yan et al., 2022). For instance, in Korea, the Korean Standard Classification of Diseases (KCD) was developed. Therefore, a specific model or algorithm may not always be effective and the development of models that consider each nation's ICD system is essential. Another reason for these disparities is differences in linguistic characteristics. Each language possesses its own grammar, syntax, and semantic structure, which affect medical terms and their expression. Hence, models that do not consider these characteristics may exhibit limited performance. Lastly, resource availability also plays an important role. Resource-rich countries use large-scale clinical data and systematic medical knowledge resources(Bodenreider, 2004) to accelerate automatic ICD coding research(Yuan et al., 2022). In contrast, in countries with limited data and resources, research progress may be slower or more constrained.

To bridge the disparities in research levels across countries and enhance Korea's constrained clinical natural language processing research environment, we propose a comprehensive approach: the development of a new dataset for automatic KCD coding and the implementation of a tokenization method tailored for Korean clinical documents. First, we addressed the lack of Korean datasets for automatic KCD coding by collecting and preprocessing clinical documents with assigned KCD codes from Korean clinical environments. Subsequently, we employed an optimized tokenization method to ensure that the automatic KCD coding model accurately captured the linguistic characteristics in the Korean clinical documents. Our contributions are as follows:

- To the best of our knowledge, our study is the first to focus on automatic KCD coding and to

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 96–101

examine the important factors that need to be considered for its improvement.

- We construct a **Ko**rean **D**ataset for **A**utomatic **K**CD coding (**KoDAK**) using initial diagnostic records collected from Korean clinical environments, and we conduct a thorough statistical analysis of the data.
- We propose an optimized tokenization method that effectively captures the linguistic characteristics of Korean clinical documents.
- Through comparative experiments, we confirm that the proposed method shows significant improvements over the existing approach.

2 KoDAK

The most renowned dataset for automatic ICD coding is MIMIC-III(Johnson et al., 2016), an English dataset collected from the intensive care unit. Furthermore, datasets are available for countries with relatively abundant language resources, such as China(Cao et al., 2020) and Spain(Goeuriot et al., 2020). However, there is no such dataset in Korea, which makes conducting research challenging. To address this issue, we constructed the **Ko**rean **D**ataset for **A**utomated **K**CD coding (**KoDAK**) for automatic KCD coding. In this section, we describe the process of constructing the KoDAK.

2.1 Data Collection

In this study, we collected clinical records from Konkuk University Medical Center to create the KoDAK. The data collection process involved obtaining approval from the institutional review board and safeguarding patient confidentiality and anonymity. The collected clinical records consisted of text written by doctors describing the patient's initial diagnosis and symptoms. Similar to the CCHMC(Pestian et al., 2007) dataset, each record is labeled with the most appropriate single KCD code. The dataset encompasses 1,196,739 documents collected from 23 departments over 17 years, from 2005 to 2021. Records containing sensitive information from departments such as psychiatry, obstetrics, gynecology, and urology were excluded from the dataset.

2.2 Data Preprocessing

We preprocessed the collected clinical records to enhance data quality and the accuracy of the analysis. In the preprocessing step, we first corrected for spacing. Because documents are in a free-form style, spacing errors frequently occur. We used KoSpacing(Jeon), a Korean spacing correction library, to rectify the spacing errors. Second, we eliminated any embedded image links in the documents that did not offer meaningful information. Finally, we sorted the entire dataset by text length and removed the shortest 5% of samples because they likely contained insufficient information to determine the KCD codes accurately.

Clinical Note	persistent **sinus tarsi pain**. painful when walking, very painful even at night while sleeping. in the morning, when getting up and stepping on it, the pain is so intense it feels like collapsing. --- 지속적으로 **sinus tarsi pain**이 있음. 걸을 때 아프고, 밤에 잘 때도 너무 아프다. 아침에 일어나서 딛으면 주저앉을 것처럼 아프다.
KCD	S9200 (Fracture of calcaneus, closed)

Table 1: Example of clinical note and corresponding KCD code in the KoDAK, with the English translation of clinical note separated by a horizontal line. Bold text indicates English text that appears in the original clinical note.

2.3 Data Example and Statistics

Table 1 presents an example of clinical note and their corresponding KCD code from the completed dataset. As observed in the clinical notes, doctors often use a combination of Korean and English when describing symptoms because of the prevalence of English medical terminology. Consequently, the KoDAK, written by Korean medical professionals, contains many English words (Korean: 81%; English: 17%; Other: 2%).

The KoDAK comprises 8,862 KCD codes, which account for 49% of all the KCD codes. The dataset shows a long-tail distribution, with the top 20% of the most frequent KCD codes covering approximately 94% of the dataset. Moreover, the least frequent 1,894 KCD codes appeared only once in the dataset.

3 Automatic KCD Coding Approach

3.1 Tokenization

As illustrated in Table 1, English medical terms such as "sinus tarsi pain" are crucial for KCD

coding and must be carefully considered during tokenization. However, the existing Korean medical language model, KM-BERT(Kim et al., 2022), does not specifically account for English medical terms, which results in a low proportion of English tokens in the vocabulary (Korean: 71%; English: 2%; Other: 27%). This is likely because the training data for KM-BERT are mostly in Korean, unlike the KoDAK, which has a higher proportion of English. Table 2 presents the tokenization results of the samples in KoDAK using both the proposed method and the KM-BERT tokenizer.

Tokenizer	Tokens
KM-BERT	2007, 코, 다, 친, 후, **a**, **n**, **os**, **m**, **ia**, **s**, **n**, **or**, **ing**, +, **a**, **p**, **ne**, **a**, 가끔
Ours	2, 0, 0, 7, 코, 다친, 후, **anosmia**, **snoring**, +, **apnea**, 가끔

Table 2: Comparison of tokenization results for clinical notes from the KoDAK using KM-BERT and our proposed method. Bold tokens indicate English medical terms.

As shown in Table 2, KM-BERT excessively splits crucial medical terms into smaller units. Such tokenization can impair a model's language comprehension and make it challenging to train the model effectively.

To address this issue, we propose an optimized tokenization strategy that considers the characteristics of the KoDAK. We applied morpheme-aware subword tokenization(Park et al., 2020) for Korean, whereas we tokenized English at the word level to preserve the meanings of essential medical terms and abbreviations. Other text types, such as numbers and special characters, were tokenized at the character level. Through this tokenization process, we built a vocabulary of 73,241 tokens that comprised 32,390 Korean, 38,659 English, and 45 other text tokens. Specifically, we determined the number of English tokens required to preserve as many crucial medical terms as possible to maintain coverage of over 93%.

3.2 Model

For automatic KCD coding, we utilized a model based on the transformer encoder architecture(Vaswani et al., 2017), which has been proven to be effective in various natural language processing tasks. Instead of initializing the token embeddings with random values, we used a separate Word2Vec(Mikolov et al., 2013) to ensure richer representations for each token. Word2Vec was trained using the skip-gram approach, and the results were used as the initial embeddings for the transformer. Our model comprises six encoder layers. To classify the final KCD codes, we follow BERT(Devlin et al., 2019)'s classification model training framework by feeding the [CLS] token representation into a linear layer.

To train the model, we employed the cross-entropy loss function(Good, 1952), which encourages the model to assign higher probabilities to correct KCD codes.

4 Experiments

4.1 Experimental Settings

The dataset was divided into training and evaluation datasets. For the KCD codes with only one instance (1,894 labels), the sample was assigned exclusively to the evaluation data, whereas for the KCD codes with more than one instance (5,275 labels), the sample was distributed between the training and evaluation. Consequently, the training data consisted of 1,130,942 samples, and the evaluation data consisted of 65,797 samples.

In this study, we used KM-BERT as a comparative model. KM-BERT is a language model for Korean medical natural language processing designed to alleviate the challenges in text analysis due to the agglutinative nature of the Korean language and complex medical terminology. KM-BERT was trained on a collection of Korean medical corpora using BERT's pre-training framework.

To evaluate the performance of the proposed method, we used Macro-F1 and Micro-F1, which are widely recognized metrics for evaluating classification models.

4.2 Experiment Results

Our proposed model for automatic KCD coding demonstrated remarkable performance, as shown in Table 4. Notably, our model outperforms the comparative model by achieving a 0.14%p improvement in Macro-F1, which showcases the strength of the proposed model in consistently enhancing performance across numerous labels, regardless of sample size. The results are especially noteworthy, considering our model has only 55% of the parameters of the comparative model, and it is not pre-trained.

	Clinical Notes	Ours	KM-BERT
Case 1	**accessory thumb left O) at birth V) congenital** patient visited for surgical treatment due to congenital polydactyly. **accessory thumb left O) at birth V) congenital** 상기 환아 선천적 다지증으로 수술적 치료 위하여 내원함	Q691	M2124
Case 2	2 days ago, after a fall, the patient experienced **retrograde amnesia** and visited the outpatient department of our hospital. 2일전 넘어진 후 **retrograde amnesia** 발생하여 본원 외래 내원함	R412	S0620

Table 3: Case study illustrating the improved KCD code predictions by our proposed model compared with KM-BERT, with the English translation of clinical notes separated by a horizontal line. Bold text indicates English text that appears in the original clinical note.

	Macro-F1	Micro-F1	Params
KM-BERT	8.59	**44.36**	105M
Ours	**8.73**	43.39	60M
w/o W	7.64	41.29	60M
w/o W & T	7.12	40.89	32M

Table 4: Performance comparison of our proposed model with KM-BERT and ablation study results (w/o W: without Word2Vec token embedding initialization; w/o T: using KM-BERT tokenizer instead of our proposed tokenizer).

Although our model showed a relatively lower performance in Micro-F1, it is important to consider that KM-BERT, the comparative model, was pre-trained on a large medical corpus. This allows it to leverage knowledge transfer to improve the performance of classes with a large number of samples. Despite this advantage, our proposed model remains highly competitive and offers a more efficient alternative, especially in terms of model size and the absence of pre-training requirements.

4.3 Ablation Study

We conducted an ablation analysis to assess the impact of Word2Vec and the proposed tokenization method on the performance of the model. Table 4 presents the results of the study.

When Word2Vec was not used (w/o W in Table 4), the performance declined across all evaluation metrics. This implies that incorporating Word2Vec enhances the model by offering richer token representations. In addition, we observed that using Word2Vec accelerated the model's convergence (best number of epochs: 18 with Word2Vec, 22 without). Removing both our proposed tokenization method and Word2Vec and employing the KM-BERT tokenizer instead (w/o W & T in Table 4), the performance deteriorates further across all evaluation metrics relative to solely removing Word2Vec. This finding underscores the proposed tokenization method positively influencing the model's performance.

4.4 Case study

We conducted a case study to understand better the improvements made to the proposed model. Table 3 lists the cases in which the proposed model accurately predicted the correct KCD codes.

In Table 3, Case 1 presents a situation where the correct KCD code is Q691 (accessory thumb). The proposed model precisely classified it as Q691 by considering the term "accessory thumb" in the clinical note. On the other hand, the comparison model misclassified it as M2124 (Flexion deformity, hand), a subclass of M21 (Other acquired deformities of limbs), despite the presence of words like "선천적 (congenital)", "congenital", and "at birth" in the note. In Case 2, the correct KCD code was R412 (retrograde amnesia). Our model accurately identified it using the phrase "retrograde amnesia" from the notes. However, the comparison model misclassified it as S0620 (diffuse brain injury, without open intracranial wound), a similar but different code. This demonstrates the effectiveness of the proposed tokenization method in capturing the meaning of English medical terms and helping the model better understand and interpret documents.

5 Conclusion and Future Work

In this study, we introduced the KoDAK as a resource for facilitating automatic KCD coding research in Korea, where the lack of suitable datasets

has hindered such research. Furthermore, we proposed a tokenization method that effectively reflects the linguistic features of Korean clinical documents, thereby ensuring an accurate representation of crucial medical terms. Our approach outperformed KM-BERT, achieving a 0.14%p improvement in Macro-F1 while utilizing fewer parameters and without pre-training. In future research, we aim to address the unbalanced label distribution in the KoDAK and develop an enhanced pre-trained language model specifically designed for the Korean clinical field.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques)

References

Musaed Ali Alharbi, Godfrey Isouard, and Barry Tolchard. 2019. The development of icd adaptations and modifications as background to a potential saudi arabia's national version. *Global Journal of Health Science*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Pengfei Cao, Chenwei Yan, Xiangling Fu, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 294–301, Online. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lorraine Goeuriot, Hanna Suominen, Liadh Kelly, Antonio Miranda-Escalada, Martin Krallinger, Zhengyang Liu, Gabriella Pasi, Gabriela Gonzalez Saez, Marco Viviani, and Chenchen Xu. 2020. Overview of the clef ehealth evaluation lab 2020. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 255–271. Springer.

Irving John Good. 1952. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114.

James E Harrison, Stefanie Weber, Robert Jakob, and Christopher G Chute. 2021. Icd-11: an international classification of diseases for the twenty-first century. *BMC medical informatics and decision making*, 21(6):1–10.

Heewon Jeon. Kospacing: Automatic korean word spacing. https://github.com/haven-jeon/KoSpacing.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Yoojoong Kim, Jong-Ho Kim, Jeong Moon Lee, Moon Joung Jang, Yun Jin Yum, Seongtae Kim, Unsub Shin, Young-Min Kim, Hyung Joon Joo, and Sanghoun Song. 2022. A pre-trained bert for korean medical natural language processing. *Scientific Reports*, 12(1):1–10.

Dee Lang. 2007. Consultant report-natural language processing in the health care industry. *Cincinnati Children's Hospital Medical Center, Winter*, 6.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China. Association for Computational Linguistics.

John P. Pestian, Christopher Brew, Paweł Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Włodzisław Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, page 97–104, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pengtao Xie and Eric Xing. 2018. A neural architecture for automated ICD coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, Melbourne, Australia. Association for Computational Linguistics.

Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, and Qiang Li. 2022. A survey of automated international classification of diseases coding: development, challenges, and applications. *Intelligent Medicine*, 2(3):161–173.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, Dublin, Ireland. Association for Computational Linguistics.

Shurui Zhang, Bozheng Zhang, Fuxin Zhang, Bo Sang, and Wanchun Yang. 2022. Automatic ICD coding exploiting discourse structure and reconciled code embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2883–2891, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.