applied sciences

MDPI

Article

Keyword Extraction Algorithm for Classifying Smoking Status from Unstructured Bilingual Electronic Health Records Based on Natural Language Processing

Ye Seul Bae [1], Kyung Hwan Kim [2,3,*], Han Kyul Kim [1], Sae Won Choi [1], Taehoon Ko [4], Hee Hwa Seo [1], Hae-Young Lee [5] and Hyojin Jeon [1]

1 Office of Hospital Information, Seoul National University Hospital, Seoul 03080, Korea; byeye1313@gmail.com (Y.S.B.); hank110@snu.ac.kr (H.K.K.); swc1@snu.ac.kr (S.W.C.); heehwaseo@gmail.com (H.H.S.); tarahjjeon@naver.com (H.J.)
2 Department of Thoracic & Cardiovascular Surgery, Seoul National University Hospital, Seoul 03080, Korea
3 Department of Thoracic & Cardiovascular Surgery, College of Medicine, Seoul National University, Seoul 03080, Korea
4 Department of Medical Informatics, The Catholic University of Korea, Seoul 06591, Korea; taehoonko@snu.ac.kr
5 Department of Internal Medicine, Seoul National University Hospital, Seoul 03080, Korea; hylee612@snu.ac.kr
* Correspondence: kkh726@snu.ac.kr

Featured Application: The study presents an improved and easily obtainable method in terms of automatic smoking classification from unstructured bilingual electronic health records.

Abstract: Smoking is an important variable for clinical research, but there are few studies regarding automatic obtainment of smoking classification from unstructured bilingual electronic health records (EHR). We aim to develop an algorithm to classify smoking status based on unstructured EHRs using natural language processing (NLP). With acronym replacement and Python package Soynlp, we normalize 4711 bilingual clinical notes. Each EHR notes was classified into 4 categories: current smokers, past smokers, never smokers, and unknown. Subsequently, SPPMI (Shifted Positive Point Mutual Information) is used to vectorize words in the notes. By calculating cosine similarity between these word vectors, keywords denoting the same smoking status are identified. Compared to other keyword extraction methods (word co-occurrence-, PMI-, and NPMI-based methods), our proposed approach improves keyword extraction precision by as much as 20.0%. These extracted keywords are used in classifying 4 smoking statuses from our bilingual EHRs. Given an identical SVM classifier, the F1 score is improved by as much as 1.8% compared to those of the unigram and bigram Bag of Words. Our study shows the potential of SPPMI in classifying smoking status from bilingual, unstructured EHRs. Our current findings show how smoking information can be easily acquired for clinical practice and research.

Keywords: smoking; natural language processing; electronic health records; document classification; lifestyle modification

Citation: Bae, Y.S.; Kim, K.H.; Kim, H.K.; Choi, S.W.; Ko, T.; Seo, H.H.; Lee, H.-Y.; Jeon, H. Keyword Extraction Algorithm for Classifying Smoking Status from Unstructured Bilingual Electronic Health Records Based on Natural Language Processing. *Appl. Sci.* **2021**, *11*, 8812. https://doi.org/10.3390/app11198812

Academic Editor: Keun Ho Ryu

Received: 23 August 2021
Accepted: 17 September 2021
Published: 22 September 2021

1. Introduction

Smoking is a major risk factor in developing coronary artery disease, chronic kidney disease, cancer, and cardiovascular disease (CVD) [1,2]. It is also considered as a modifiable risk factor for CVDs and other conditions associated with premature death worldwide [3–6]. Consequently, smoking status can be used to assess the risk of certain diseases and to suggest first-line interventions based on clinical guidelines.

Despite the effectiveness and importance of smoking cessation for disease prevention, smoking information is under-utilized and not easily measured. It is often buried in a narrative text rather than in a consistent coded form. The rapid adoption of electronic health

records (EHRs) has also scattered patient information across various systems as both structured and unstructured data [7]. However, most information stored in EHRs takes the form of free text in clinical narratives. Therefore, applying natural language processing (NLP) methods is essential in automatically transforming the clinical free text into structured clinical data, which can be further utilized by machine learning algorithms [8–10].

Several previous studies successfully extracted smoking information using NLP in English. Jonnagaddala et al. [7] extracted smoking information and converted it into a structured format. Shoenbill et al. [11] retrieved assessments and advice about lifestyle modifications using an open-source NLP tool, cTAKES. However, unlike languages such as English, in which whitespaces define the boundaries between words, word tokenization is not a trivial task in Korean that has no clear word delimiting rules. Furthermore, in most non-English speaking countries, their medical records are often bilingual, containing terms expressed in both English and their native languages. Therefore, extracting smoking information from Korean EHR requires a different set of NLP methods to overcome this issue of bilingual free text.

This paper presents a novel smoking-related keyword extraction algorithm from unstructured bilingual EHRs. Medical records often contain diverse information about the patient's medical and health status. However, our NLP-based keyword extraction algorithm filters out and identifies keywords solely relevant to smoking status. Furthermore, the identified smoking-related keywords can also be utilized to improve the accuracy of the existing NLP-based smoking status classification algorithms. Our proposed method is purely unsupervised and does not require a huge training dataset, which is especially critical in the medical domain. Therefore, it serves as a practical solution to any medical institutions, including those in non-English speaking countries, planning to transform their free text EHR into a structured keyword format.

2. Materials and Methods

2.1. Data

We applied our keyword extraction method to 4711 clinical notes collected from Seoul National University Hospital (SNUH) from 1 January to 31 December 2017, through the clinical data warehouse (CDW) of SNUH, SUPREME (Seoul National University Hospital Patients Research EnvironMEnt). Of those, 3512 notes were collected from the department of family medicine (including the. Patients with diabetes), and the rest were from the department of pulmonary and critical care medicine (including the patients. with chronic obstructive pulmonary disorders). Each clinical note contains a patient's overall medical history as recorded by the doctors from each department. Furthermore, each note contains, on average, 157.04 tokens. However, the range of the token length varies between 1 and 589.

As the notes are written as free text, different doctors express identical terms or concepts differently. The notes contain both English and Korean words, which is very common practice in Korea and further complicates the keyword extraction. Although several researchers have worked with bilingual or multilingual EHRs [12–14], our paper is the first to focus on extracting bilingual keywords from EHRs. Based on patient smoking status, each of the notes were manually labeled into one of these four categories: current smokers, past smokers, never smokers, and unknown. Clinical notes were manually labeled by 3 medical students and 1 nurse, and errors were reviewed by 1 physician. Despite these class labels, sentences or words that suggest a patient's smoking status were not annotated. Consequently, the labels are not used in the keyword extraction process. However, including all notes regardless of their class labels introduces additional difficulty in extracting meaningful smoking-related keywords that can test the robustness of the proposed algorithm. Table 1 shows the overall statistics for our data. This study was approved by the Institutional Review Board of Seoul National University Hospital (Institutional Review Board number: N-1906-076-1040).

Table 1. Overall summary of SNUH clinical notes data.

	Family Medicine	Pulmonary and Critical Care Medicine	Total
Current smokers	1046	84	1130
Past smokers	547	431	978
Never smokers	399	144	543
Unknown	1520	540	2060
Total	3512	1199	4711

2.2. SPPMI-Based Keyword Extraction

In this work, we introduce SPPMI (Shifted Positive Point Mutual Information) [15]-based keyword expansion to extract smoking status-related keywords from bilingual EHRs. It is an end-to-end unsupervised method, thus not requiring any annotated data or model training. Therefore, it is easily applicable in biomedical practice as manual annotation is both time-consuming and especially costly in the medical field. SPPMI-based keyword extraction uses three main steps: text preprocessing, seed word preparation, and keyword extraction (Figure 1).

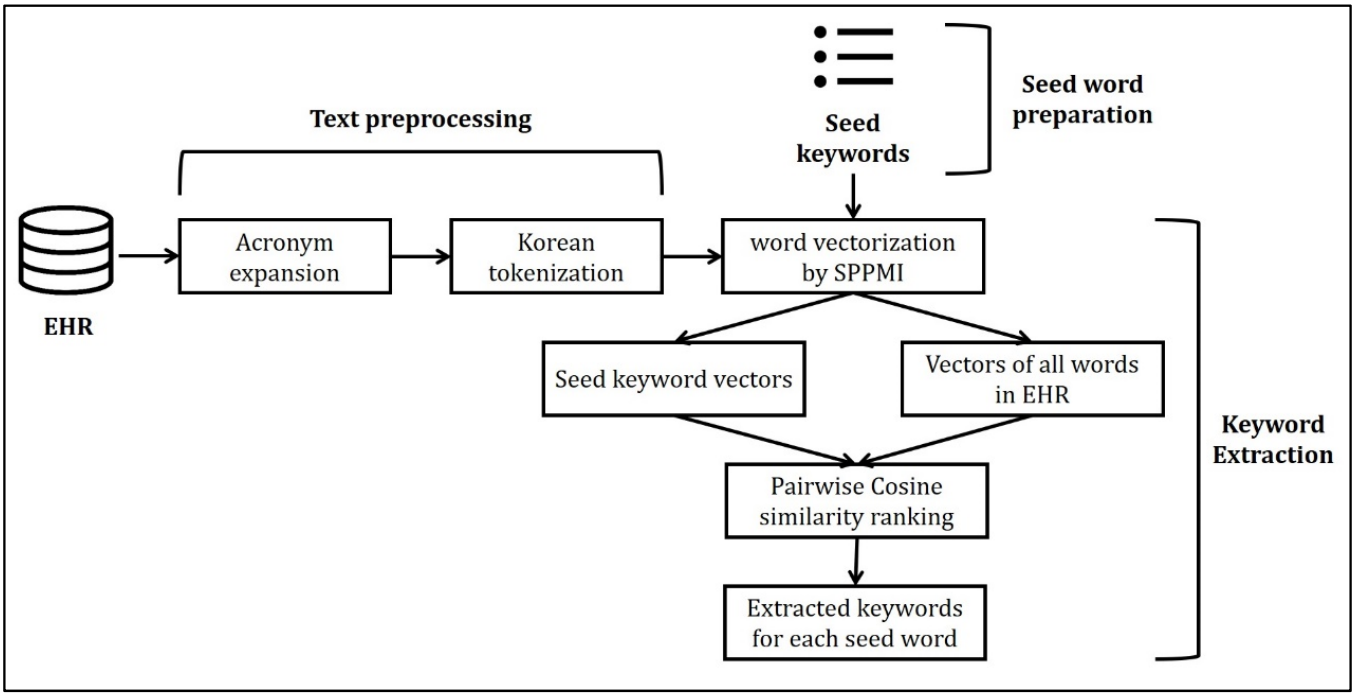

Figure 1. Overall summary of our SPPMI-based keyword extraction method.

During the text preprocessing step, we identified 170 commonly used medical acronyms and replaced them with their full expressions. Unlike in English, not all words in Korean are delimited by spaces. As an agglutinating language, these words are often delimited by a set of special words or particles [16]. To tokenize Korean phrases (eojeol) into the most semantically relevant words, we applied the Python package soynlp (https://github.com/lovit/soynlp, accessed on 8 September 2021) to the texts written in Korean. Without relying on any predefined dictionaries, soynlp finds the boundaries between Korean words by estimating the probability of those boundaries at the character level.

After text preprocessing, we prepared a list of known smoking status-related seed keywords. They served as a basis for finding other keywords that describe a patient's smoking status. Two medical professionals analyzed frequently occurring words in our data and generated 50 smoking status-related keywords: 11 never-smoker-related keywords, 16 past-smoker-related keywords, and 13 current-smoker-related keywords. We limited

our seed words to unigrams or bigrams for clarity and computational efficiency. For Korean keywords, their unigrams and bigrams are defined in terms of words identified by soynlp. A few examples of seed keywords from each smoking status are provided in Table 2.

Table 2. Examples of our seed keywords. Both English and Korean words were selected as seed words. The two Korean keywords superscripted 1 and 2 translate to "non-smoking" and "stopped smoking", respectively.

Never Smoker	Past Smoker	Current Smoker
smk never	smk ex	current smoker
smk negative	smoker ya	smk yr
비흡연 [1]	금연 중 [2]	smoking

During the keyword extraction step, we identified semantically similar words to each of our seed words. To calculate the semantic similarity, we applied SPPMI to represent both our seed words and all the words in our EHR data as numerical vectors. We then calculated and ranked the pairwise cosine similarity between each seed word vector and other word vectors. Words with the highest cosine similarity to the seed words are identified as the extracted keywords.

In SPPMI, each word is initially represented as a vector of its pointwise mutual information (PMI) scores [17] with every other word in the dataset. As described in equation 1, PMI provides a probabilistic measure of association between two words by comparing their jointly occurring probability with their individual probabilities. Because they can effectively capture word similarity, PMI and its variants, such as normalized PMI (NPMI) [18], are frequently used in NLP [19–21].

$$\mathrm{PMI}(\mathrm{word}_1,\ \mathrm{word}_2) = \ln\left(\frac{p(word_1,\ word_2)}{p(word_1) \times p(word_2)}\right) \tag{1}$$

Equation (1). Pointwise mutual information (PMI).

As shown in Equation (2), SPPMI shifts the PMI values of each word vector by a global constant ($\log k$). If lower-dimensional word vectors are desired, matrix factorization, such as singular value decomposition (SVD), is additionally applied. Depending on the value of k or the number of singular values used in the SVD, SPPMI can capture word similarity better than other neural-network-inspired word representation methods [22]. For example, Levy et al. [15,22] showed that word2vec is implicitly factorizing a word co-occurrence matrix, in which each co-occurrence is calculated as PMI. As the underlying mechanism of word2vec [23] and SPPMI is identical, they experimentally showed that SPPMI could achieve a similar level of performance as word2vec. In this work, we chose the same values of k (1, 5, 15) and the number of singular values (100, 500, 1000) as used in the original paper.

$$\mathrm{SPPMI}(\mathrm{word}_1,\ \mathrm{word}_2) = \max(PMI(word_1,\ word_2) - \log k,\ 0) \tag{2}$$

Equation (2). Shifted Positive Point Mutual Information (SPPMI).

3. Results

3.1. Experiment Setting

To objectively measure the performance of SPPMI-based keyword expansion, we compare its extracted keywords with those from six other models. Although they share identical text preprocessing and seed word preparation steps, they all use different keyword extraction steps. The word co-occurrence, PMI vector, and NPMI vector models represent each word as a vector of its co-occurrence counts, PMI scores, and NPMI scores, respectively, with every other word in the dataset. Based on those vector representations, the baseline models rank pairwise cosine similarity to extract keywords that belong to the same smoking

status as their seed words. For the PMI and NPMI score-based keyword extraction models, we did not create any word vectors. Instead, we calculated the pairwise PMI and NPMI scores, respectively, between each seed word and all other words in the dataset. By ranking the scores, these two models similarly extract relevant keywords for each seed word. In word2vec models, each word vector is represented by the weights of a neural network that is trained to predict a word given its neighboring words. As one of the most basic word embedding methods, it is widely applied in solving various word-level NLP tasks [24–27]. In our experiment, we trained nine word2vec models with different hyperparameters. For the dimension of the word vectors, we set it to be 100, 200, or 300. For the context size, we designated it to be 2, 4, or 6. To extract relevant keywords with word2vec, we once again used a pairwise cosine similarity measure.

3.2. Keyword Extraction Precision

We extracted the top 1, 5, 10, and 20 keywords from each of these unsupervised keyword extraction models for each of our 50 initial seed words. As the complete set of smoking-related keywords are not available in the dataset, we used precision to measure the performance of the keyword extraction. To compare the precision of the models, two human annotators independently assessed the extracted keywords. Each extracted keyword was deemed correct only when both annotators agreed that it described the same smoking status as its input seed word.

As shown in Table 3, our SPPMI-based keyword expansion method showed superior performance in extracting smoking status-related keywords from bilingual EHRs. It is interesting to note that the basic SPPMI model does not significantly improve keyword extraction precision. Instead, it was the lower-dimensional word vectors created from the SVD that exhibited superior and robust performance. As the number of generated keywords increased, the precision inevitably decreased because the number of words related to smoking status in our dataset is limited.

Table 3. Comparison of precision between SPPMI-based keyword expansion and five baseline models (word co-occurrence, PMI vector, NPMI vector, PMI score, and NPMI score). The values of d represent the number of singular values used in the SVD or the dimension of word vectors in word2vec. The values of c indicate the context size used in training word2vec.

Methods \ # of Keywords	Top 1	Top 5	Top 10	Top 20
Word co-occurrence	38.00%	35.60%	30.40%	29.50%
PMI vector	42.00%	36.40%	31.20%	27.40%
NPMI vector	40.00%	35.60%	32.00%	27.70%
PMI score	38.00%	36.80%	37.00%	**34.50%**
NPMI score	36.00%	38.40%	37.40%	34.20%
SPPMI (k = 1)	42.00%	36.40%	31.20%	27.40%
SPPMI (k = 5)	38.00%	34.80%	31.80%	27.40%
SPPMI (k = 15)	36.00%	30.40%	27.00%	27.20%
SPPMI-SVD (k = 1, d = 100)	46.00%	40.80%	**38.40%**	31.50%
SPPMI-SVD (k = 1, d = 500)	46.00%	**41.20%**	35.40%	29.80%
SPPMI SVD (k = 1, d = 1000)	50.00%	38.80%	31.20%	29.50%
SPPMI-SVD (k = 5, d = 100)	**56.00%**	37.60%	32.00%	29.30%
SPPMI-SVD (k = 5, d = 500)	42.00%	36.00%	32.80%	30.00%
SPPMI- SVD (k = 5, d = 1000)	42.00%	34.40%	32.80%	30.10%
SPPMI SVD (k = 15, d = 100)	**56.00%**	40.40%	33.40%	30.60%
SPPMI- SVD (k = 15, d = 500)	44.00%	32.40%	29.80%	29.30%
SPPMI-SVD (k = 15, d = 1000)	44.00%	33.20%	31.00%	28.50%

Table 3. *Cont.*

Methods \ # of Keywords	Top 1	Top 5	Top 10	Top 20
word2vec (c = 2, d = 100)	10.00%	6.80%	7.60%	7.20%
word2vec (c = 4, d = 100)	11.11%	8.00%	6.20%	5.20%
word2vec (c = 6, d = 100)	8.82%	5.20%	4.60%	4.00%
word2vec (c = 2, d = 200)	10.00%	9.60%	8.40%	7.20%
word2vec (c = 4, d = 200)	9.09%	7.20%	6.40%	4.90%
word2vec (c = 6, d = 200)	8.00%	4.80%	4.60%	4.20%
word2vec (c = 2, d = 300)	16.00%	9.60%	7.40%	5.90%
word2vec (c = 4, d = 300)	8.00%	6.80%	5.20%	3.80%
word2vec (c = 6, d = 300)	8.00%	5.20%	4.60%	4.60%

It is also interesting to note that word2vec shows poor precision. There are several reasons why word embedding methods generally do not work effectively in our data. One of the most critical issues is the huge number of unique words relative to the size of the data. When doctors are writing clinical notes, they are often simultaneously interacting with their patients. Due to this real-time nature, the expressions in the notes are often short and abbreviated. Consequently, they do not strictly adhere to standardized expressions, and their expression styles often differ between doctors. Due to this issue, semantically similar yet structurally different terms are prevalent in our dataset. Without normalizing these terms, word embedding methods fail to generalize. However, bilingual term normalization itself is another critical future research topic that is beyond the scope of this paper. The detailed precision for each of the three smoking classes (never smoker, past smoker, and current smoker) is included in Tables S1–S3 in Supplementary Materials. Additionally, Table 4 includes a few examples of extracted keywords from each smoking status. As a reference, all 50 of our seed keywords, the extracted keywords, and their statistics are publicly available at https://github.com/hank110/smoking_status_keywords (accessed on 8 September 2021).

Table 4. Examples of extracted keywords. Both English and Korean keywords were simultaneously extracted for each of our seed keyword. The two bilingual keywords superscripted 1 and 2 translate to "smoking negative" and "years ago ppd (packs per day)", respectively. The two Korean keywords superscripted 3 and 4 translate to "still cigarette" and "haven't quit", respectively.

Never Smoker	Past Smoker	Current Smoker
흡연 negative [1]	년전 ppd [2]	아직 담배 [3]
s negative	smoking ya	못 끊었어요 [4]
never smoker	quit since	still smoking

3.3. Smoking Status Classification

Our SPPMI-based keyword extraction method can also be applied to training a smoking status classifier from EHR data. Several previous works have applied machine learning algorithms or statistical analysis to classify smoking status from EHR [28–33]. Among these works, a linear support vector machine (SVM) trained from unigram and bigram bag of words has consistently shown the highest classification accuracy [7,34–36].

However, this smoking status classification accuracy can be further improved by preprocessing unigram and bigram bag of words by the keywords extracted from SPPMI. Instead of representing each EHR record by the frequencies of every word in the dataset, we represent it as a bag of keywords. For all non-keywords within the record, we simply treat them as an identical word. For example, if we decide to represent each record with five keywords extracted from SPPMI, each record becomes a vector with six dimensions (five for keywords and one for other non-keywords).

To test the impact of our SPPMI-based extracted keywords on the smoking status classification, we have fixed the classifier to be a linear SVM. Subsequently, we compare

the classification accuracy resulting from different clinical note vector representations. For all classification methods, we have trained the classifiers from 80% of our SNUH clinical notes data and used the rest as test data. As the entire EHR records were initially annotated with each patient's actual smoking status, accuracy can be measured in terms of F1 score to compare the impact of different vector representations on the classifier's performance (Table 5).

Table 5. Comparison of overall smoking status classification accuracy. All accuracies are reported in F1 scores.

Methods \ # of Extracted Keyword	1	5	10	20
Bag of Words		90.35%		
LSA	49.63%	55.04%	57.79%	61.93%
LDA	43.69%	43.69%	43.69%	43.69%
SPPMI (k = 1)	90.67%	90.99%	91.30%	91.30%
SPPMI (k = 5)	**91.20%**	89.93%	91.52%	91.83%
SPPMI (k = 15)	90.67%	91.20%	90.99%	91.41%
SPPMI-SVD (k = 1, d = 100)	90.88%	90.35%	90.56%	91.09%
SPPMI-SVD (k = 1, d = 500)	90.88%	90.88%	91.09%	**91.94%**
SPPMI SVD (k = 1, d = 1000)	90.46%	91.09%	**92.15%**	92.79%
SPPMI-SVD (k = 5, d = 100)	90.56%	**91.41%**	90.46%	91.41%
SPPMI-SVD (k = 5, d = 500)	90.56%	91.30%	91.41%	91.83%
SPPMI- SVD (k = 5, d = 1000)	90.56%	90.88%	91.94%	91.52%
SPPMI SVD (k = 15, d = 100)	91.09%	90.56%	90.35%	91.73%
SPPMI- SVD (k = 15, d = 500)	90.24%	90.56%	91.30%	91.09%
SPPMI-SVD (k = 15, d = 1000)	90.77%	90.77%	91.30%	91.30%

Compared to the unigram and bigram-based Bag of Words approach used in [7,34–36], classifying smoking status solely with the keywords extracted with our SPPMI-based approach improves the overall accuracy up to 1.8% (Table 4). This improvement in accuracy becomes more evident when we observe the classification accuracy of each smoking status (Tables S4–S7 in Supplementary Materials). For classifying smokers, our approach improves the F1 score by as much as 9.04%. Furthermore, the improved classification result also signifies that our method is capable of expanding meaningful keywords from our seed word.

In terms of machine learning, preprocessing clinical note vectors by our keywords serves as a form of dimension reduction or feature engineering. However, our approach outperforms other conventional dimension reduction techniques in document vectors, such as Latent Semantic Analysis (LSA) [37,38] and Latent Dirichlet Allocation (LDA) [39]. This superior smoking status classification result once again emphasizes the capability of our approach to expand keywords that are truly relevant to each smoking status.

3.4. Frequency Distribution of the Expanded Keywords

As our method solely relies on vector similarity, it is capable of extracting even infrequently occurring keywords. As shown in Figure 2, approximately 60% of our extracted keywords occur less than five times in the entire data. Keyword extractions that utilize statistical measures [40] or the feature importance of classifiers [41] will not be as effective as our method in capturing these infrequent keywords. As they have less impact on the overall classification accuracy, they will simply be disregarded or be replaced by more frequent features or keywords. However, these infrequent keywords provide equally meaningful insight into EHR records, especially when the amount of data or standardization is limited. They capture different ways of expressing smoking status and might even represent typos in the expressions.

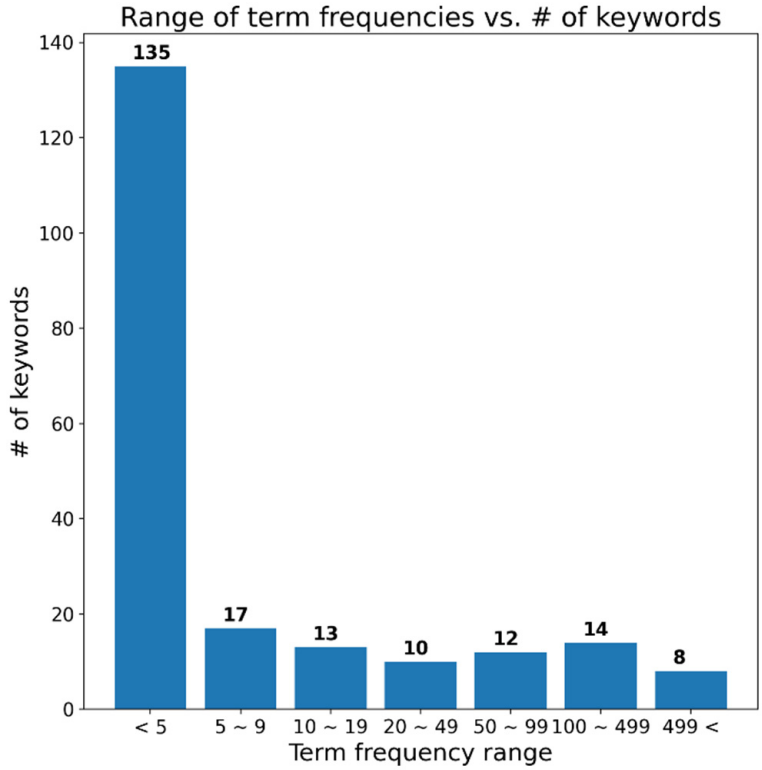

Figure 2. The number of extracted keywords with respect to their frequencies.

4. Discussion

4.1. Limitations of Pre-Trained Language Models

In this paper, we have excluded pre-trained language models such as BERT [42] and GPT [43] based methods from our experiment due to our data's domain difference and bilingual property. These pre-trained models are trained to learn effective language models from general-purpose datasets such as English Wikipedia and Bookscorpus. However, the word generating distribution in the medical domain cannot be simply assumed to be similar to those of Wikipedia or Bookscorpus. As medical notes are simultaneously generated while doctors are interviewing or examining their patients, the notes are often succinct, largely containing abbreviations, domain-specific jargon, and incomplete sentences. Therefore, without the fine-tuning pre-trained model, it cannot effectively capture the domain-specific word generating distribution present in our medical notes. Although domain-specific models such as BioBERT [44] offer fine-tuned language models for biomedical applications, the terms expressed in the clinical notes significantly differ from the normalized and pre-processed terms used in training these models. Due to the small number of notes relative to the huge number of unique words, the benefits of additional fine-tuning these existing models with our data are also limited.

Most importantly, there are no bilingual or multilingual language models in the biomedical domain at the moment. Without models that are simultaneously trained from multilingual medical text data, aligning embedding space of different language models in the medical domain is an ongoing research topic that we hope to address in the future. Furthermore, the biggest bottleneck in this multilingual language model approach is the limited medical corpus available in Korean. Consequently, to the best of the authors' knowledge, there is no large-scale pre-trained language model trained from Korean medical text data. This paper's experiment with word2vec emphasizes not only the need but also the difficulty in creating a large-scale language model for Korean medical data. Despite various hyperparameter settings, the low precision from our word2vec models indicates that they have failed to capture the semantic similarity between terms. A longer and larger set of medical notes will provide additional contextual cues to improve the performance of

word2vec and other neural network-based language models. However, creating a large medical corpus is extremely costly as it is not an easily available data source and requires medical professionals' input in processing the data. We hope that this paper will serve as a starting point for this expensive yet necessary process.

Only with sufficient medical data, language models, or deep learning algorithms are effective in the medical domain. For example. Arnaud et al. [45] uses 260,000 Emergency Department records to train its CNN text classifier, while Yao et al. [46] fine-tunes its BERT model from a Chinese clinical corpus that contains over 18 million tokens. As more Korean medical notes are currently being collected in a structural format, we also plan to improve our bilingual keyword extraction in the future. Once a large-scale Korean medical corpus is ready, training a more sophisticated language model or aligning word embedding space based on transfer learning will be interesting research topics to pursue.

4.2. Implications to Bilingual EHRs

The main purpose of EHR is to support patient care and administrative tasks related to treatment as a repository of clinical data. Therefore, EHR is not optimized for accurate retrieval of many data. Eventually, in the process of using EHR for research purposes, problems such as low accessibility, poor performance, and lack of data analysis functions arise. In particular, clinicians often use free text when recording clinical findings in EHR. Natural language processing is a representative method of extracting meaningful data from documents recorded as free statements. Attempts are being made to extract drug prescriptions, problem lists, and comprehensive clinical information from clinical documents recorded in the form of free statements using natural language processing or to use them for document classification and retrieval [47–49]. However, although current research on medical images or bio-signals has progressed considerably, research on analyzing textual medical data is insignificant. In particular, studies composed of multiple languages are not common [50]. Korean medical institutes face additional difficulties in natural language processing, as EHR contains both Korean and English. This study provides a meaningful basis for extracting insights from the clinical data warehouse, mapping documents to a standardized terminology system, and classifying bilingual EHR documents.

4.3. Strengths and Limitations

In the previous studies, smoking data were mainly collected through two sources: a structured questionnaire or a manual review of the clinical notes by researchers. Using a structured questionnaire may require a lot of effort to increase the response rate. Similarly, manual review needs significant amount of time and human resources with the possibility of human error. Our proposed method allows researchers and practitioners to easily obtain smoking information from EHR. It is also the first work to extract smoking information from the clinical free text that contains two different languages, Korean and English. Previous works on smoking status classification focused on binary classification. Their algorithms classified either smokers and non-smokers or past smokers and non-smokers. However, our classification algorithm based on keyword extraction uses multilabel classification that distinguishes between the current, past, non-smokers, and unknowns. Various treatment guidelines for noncommunicable diseases [51,52] recommend examining smoking history during every outpatient visit and educating lifestyle modification. If there is no smoking-related information in the patient's previous EHR, the medical staff can perceive it as a cue to collect new information. Moreover, it was confirmed that smoking-related keywords were expressed in various ways even if the previous chart contents were copied and pasted. Our proposed algorithm will have a practical application in automatically mapping and preprocessing unstructured clinical notes composed of various keywords that occur under the copy and paste practice.

One limitation of this study is the possibility that some clinicians did receive patients' smoking history but failed to enter it into EHR. Therefore, simply labeling patients with unknown smoking status as the unknowns may be insufficient. Second, our study was

conducted only with free-text clinical notes obtained from two departments in one tertiary hospital. The clinical notes of diabetes patients in the family medicine and COPD patients in the respiratory medicine often include the patients' smoking history, allowing us to build and validate our approach. Further validation study using data from other centers is required to test our proposed method's robustness and applicability.

5. Conclusions

Our study showed the great potential in classify smoking status from bilingual unstructured EHRs. To the best of our knowledge, this paper is one of the first works that confirmed the possibility of extracting meaningful keywords from bilingual unstructured EHRs. Instead of medical staff manually perusing through the notes, our proposed algorithm explored the possibility of replacing this time-consuming and expensive approach with an automated methodology leveraged by NLP. Due to the limitations on the amount of data available and the relatively small portion of EHRs dedicated to patients' smoking history, we could not train or apply sophisticated bilingual language models for our keyword extraction task from scratch. However, as the size of EHR would continue to increase, we plan to apply recent advancements in NLP to improve the accuracy of keyword extraction in the near future. Smoking information can be easy to acquire and use for clinical practice or research with our current findings.

Supplementary Materials: Python implementation of the study is available online at: https://www.mdpi.com/article/10.3390/app11198812/s1 or https://github.com/hank110/smoking_status_keywords, Table S1: Comparison of precision between SPPMI-based keyword expansion and five other baseline models on extracting never smoker-related keywords (Word co-occurrence, PMI vector, NPMI vector, PMI score, and NPMI score models). The values of d represent a number of singular values used in SVD, Table S2: Comparison of precision between SPPMI-based keyword expansion and five other baseline models on extracting past smoker-related keywords (Word co-occurrence, PMI vector, NPMI vector, PMI score, and NPMI score models). The values of d represent a number of singular values used in SVD, Table S3: Comparison of precision between SPPMI-based keyword expansion and five other baseline models on extracting current smoker-related keywords (Word co-occurrence, PMI vector, NPMI vector, PMI score, and NPMI score models). The values of d represent a number of singular values used in SVD, Table S4: Comparison of never smoker classification accuracy. All accuracies are reported in F1 scores, Table S5: Comparison of past smoker classification accuracy. All accuracies are reported in F1 scores, Table S6: Comparison of current smoker classification accuracy. All accuracies are reported in F1 scores, Table S7: Comparison of unknown smoking status classification accuracy. All accuracies are reported in F1 scores.

Author Contributions: K.H.K. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Y.S.B., S.W.C., H.K.K. and T.K. conceived and designed the study. Y.S.B. acquired the data. Y.S.B., H.K.K., H.H.S., H.-Y.L., H.J. and T.K. analyzed and interpreted the data. Y.S.B. and H.K.K. drafted the manuscript. Critical revision of the manuscript was provided by Y.S.B. and H.K.K. Y.S.B. and K.H.K. provided administrative, technical, or material support. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant no 04-2019-0250 from the Seoul National University Hospital Research Fund.

Institutional Review Board Statement: This study was approved by the Institutional Review Board of Seoul National University Hospital (Institutional Review Board number: N-1906-076-1040). Date of approval was 27 June 2019.

Informed Consent Statement: The informed consent was waived.

Data Availability Statement: Data is not available.

Acknowledgments: We would like to gratefully acknowledge three medical students, Yongjun Jang, Soyoung Kong, and Seunghyun Yoon, who participated in SNUH mentorship program. They helped with the initial labeling of clinical notes.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baker, F.; Ainsworth, S.R.; Dye, J.T.; Crammer, C.; Thun, M.J.; Hoffmann, D.; Repace, J.L.; Henningfield, J.E.; Slade, J.; Pinney, J. Health risks associated with cigar smoking. *Jama* **2000**, *284*, 735–740. [CrossRef]
2. Freund, K.M.; Belanger, A.J.; D'Agostino, R.B.; Kannel, W.B. The health risks of smoking the framingham study: 34 years of follow-up. *Ann. Epidemiol.* **1993**, *3*, 417–424. [CrossRef]
3. Jha, P.; Ramasundarahettige, C.; Landsman, V.; Rostron, B.; Thun, M.; Anderson, R.N.; McAfee, T.; Peto, R. 21st-century hazards of smoking and benefits of cessation in the United States. *N. Engl. J. Med.* **2013**, *368*, 341–350. [CrossRef] [PubMed]
4. Jha, P. Avoidable global cancer deaths and total deaths from smoking. *Nat. Rev. Cancer* **2009**, *9*, 655–664. [CrossRef] [PubMed]
5. Godtfredsen, N.S.; Holst, C.; Prescott, E.; Vestbo, J.; Osler, M. Smoking reduction, smoking cessation, and mortality: A 16-year follow-up of 19,732 men and women from The Copenhagen Centre for Prospective Population Studies. *Am. J. Epidemiol.* **2002**, *156*, 994–1001. [CrossRef] [PubMed]
6. Mons, U.; Müezzinler, A.; Gellert, C.; Schöttker, B.; Abnet, C.C.; Bobak, M.; de Groot, L.; Freedman, N.D.; Jansen, E.; Kee, F.; et al. Impact of smoking and smoking cessation on cardiovascular events and mortality among older adults: Meta-analysis of individual participant data from prospective cohort studies of the CHANCES consortium. *BMJ* **2015**, *350*, h1551. [CrossRef]
7. Jonnagaddala, J.; Dai, H.-J.; Ray, P.; Liaw, S.-T. A preliminary study on automatic identification of patient smoking status in unstructured electronic health records. In Proceedings of the BioNLP 15, Beijing, China, 30 July 2015; pp. 147–151.
8. Kim, H.K.; Choi, S.W.; Bae, Y.S.; Choi, J.; Kwon, H.; Lee, C.P.; Lee, H.-Y.; Ko, T. MARIE: A Context-Aware Term Mapping with String Matching and Embedding Vectors. *Appl. Sci.* **2020**, *10*, 7831. [CrossRef]
9. Elbattah, M.; Arnaud, É.; Gignon, M.; Dequen, G. The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications. In Proceedings of the HEALTHINF, Vienna, Austria, 11–13 February 2021; pp. 825–832.
10. Golmaei, S.N.; Luo, X. DeepNote-GNN: Predicting hospital readmission using clinical notes and patient network. In Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, Virtual Conference, 1–4 August 2021; pp. 1–9.
11. Shoenbill, K.; Song, Y.; Gress, L.; Johnson, H.; Smith, M.; Mendonca, E.A. Natural language processing of lifestyle modification documentation. *Health Inform. J.* **2020**, *26*, 388–405. [CrossRef]
12. Miñarro-Giménez, J.A.; Cornet, R.; Jaulent, M.-C.; Dewenter, H.; Thun, S.; Gøeg, K.R.; Karlsson, D.; Schulz, S. Quantitative analysis of manual annotation of clinical text samples. *Int. J. Med. Inform.* **2019**, *123*, 37–48. [CrossRef]
13. Pilán, I.; Brekke, P.H.; Øvrelid, L. Building a Norwegian Lexical Resource for Medical Entity Recognition. *arXiv* **2004**, arXiv:2004.02509.
14. Leslie, H. openEHR archetype use and reuse within multilingual clinical data sets: Case study. *J. Med. Internet Res.* **2020**, *22*, e23361. [CrossRef]
15. Levy, O.; Goldberg, Y. Neural word embedding as implicit matrix factorization. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2177–2185.
16. Kang, M.-Y. Topics in Korean Syntax: Phrase Structure, Variable Binding and Movement. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1988.
17. Church, K.; Hanks, P. Word association norms, mutual information, and lexicography. *Comput. Linguist.* **1990**, *16*, 22–29.
18. Bouma, G. Normalized (pointwise) mutual information in collocation extraction. *Proc. GSCL* **2009**, *30*, 31–40.
19. Ravichandran, D.; Pantel, P.; Hovy, E. Randomized algorithms and NLP: Using locality sensitive hash functions for high speed noun clustering. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, MI, USA, 23–25 June 2005; pp. 622–629.
20. Han, L.; Finin, T.; McNamee, P.; Joshi, A.; Yesha, Y. Improving word similarity by augmenting PMI with estimates of word polysemy. *IEEE Trans. Knowl. Data Eng.* **2012**, *25*, 1307–1322. [CrossRef]
21. Arora, S.; Li, Y.; Liang, Y.; Ma, T.; Risteski, A. A latent variable model approach to pmi-based word embeddings. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 385–399. [CrossRef]
22. Levy, O.; Goldberg, Y.; Dagan, I. Improving distributional similarity with lessons learned from word embeddings. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 211–225. [CrossRef]
23. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
24. Wang, P.; Xu, B.; Xu, J.; Tian, G.; Liu, C.-L.; Hao, H. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing* **2016**, *174*, 806–814. [CrossRef]
25. Kim, H.K.; Kim, H.; Cho, S. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing* **2017**, *266*, 336–352. [CrossRef]
26. Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; Qin, B. Learning sentiment-specific word embedding for twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 23–25 June 2014; Volume 1, pp. 1555–1565.
27. Nikfarjam, A.; Sarker, A.; O'connor, K.; Ginn, R.; Gonzalez, G. Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 671–681. [CrossRef] [PubMed]
28. Uzuner, Ö.; Goldstein, I.; Luo, Y.; Kohane, I. Identifying patient smoking status from medical discharge records. *J. Am. Med. Inform. Assoc.* **2008**, *15*, 14–24. [CrossRef]

29. Cohen, A.M. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. *J. Am. Med. Inform. Assoc.* **2008**, *15*, 32–35. [CrossRef] [PubMed]
30. Clark, C.; Good, K.; Jezierny, L.; Macpherson, M.; Wilson, B.; Chajewska, U. Identifying smokers with a medical extraction system. *J. Am. Med. Inform. Assoc.* **2008**, *15*, 36–39. [CrossRef] [PubMed]
31. Golden, S.E.; Hooker, E.R.; Shull, S.; Howard, M.; Crothers, K.; Thompson, R.F.; Slatore, C.G. Validity of Veterans Health Administration structured data to determine accurate smoking status. *Health Inform. J.* **2020**, *26*, 1507–1515. [CrossRef]
32. Groenhof, T.K.J.; Koers, L.R.; Blasse, E.; de Groot, M.; Grobbee, D.E.; Bots, M.L.; Asselbergs, F.W.; Lely, A.T.; Haitjema, S.; van Solinge, W. Data mining information from electronic health records produced high yield and accuracy for current smoking status. *J. Clin. Epidemiol.* **2020**, *118*, 100–106. [CrossRef] [PubMed]
33. De Silva, L.; Ginter, T.; Forbush, T.; Nokes, N.; Fay, B.; Mikuls, T.; Cannon, G.; DuVall, S. Extraction and quantification of pack-years and classification of smoker information in semi-structured Medical Records. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011.
34. Figueroa, R.L.; Soto, D.A.; Pino, E.J. Identifying and extracting patient smoking status information from clinical narrative texts in Spanish. In Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 2710–2713.
35. Patel, J.; Siddiqui, Z.; Krishnan, A.; Thyvalikakath, T.P. Leveraging electronic dental record data to classify patients based on their smoking intensity. *Methods Inf. Med.* **2018**, *57*, 253–260. [CrossRef] [PubMed]
36. Caccamisi, A.; Jørgensen, L.; Dalianis, H.; Rosenlund, M. Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records. *Upsala J. Med Sci.* **2020**, *125*, 316–324. [CrossRef]
37. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
38. Hofmann, T. Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999; pp. 50–57.
39. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
40. Matsuo, Y.; Ishizuka, M. Keyword extraction from a single document using word co-occurrence statistical information. *Int. J. Artif. Intell. Tools* **2004**, *13*, 157–169. [CrossRef]
41. HaCohen-Kerner, Y.; Gross, Z.; Masa, A. Automatic extraction and learning of keyphrases from scientific articles. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico, 13–19 February 2005; pp. 657–669.
42. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
43. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
44. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
45. Arnaud, É.; Elbattah, M.; Gignon, M.; Dequen, G. Deep learning to predict hospitalization at triage: Integration of structured data and unstructured text. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 4836–4841.
46. Yao, L.; Jin, Z.; Mao, C.; Zhang, Y.; Luo, Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1632–1636. [CrossRef] [PubMed]
47. Xu, H.; Stenner, S.P.; Doan, S.; Johnson, K.B.; Waitman, L.R.; Denny, J.C. MedEx: A medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 19–24. [CrossRef]
48. Haerian, K.; Varn, D.; Vaidya, S.; Ena, L.; Chase, H.; Friedman, C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin. Pharmacol. Ther.* **2012**, *92*, 228–234. [CrossRef]
49. Park, R.W. A clinical research strategy using longitudinal observational data in the post-electronic health records era. *J. Korean Med. Assoc.* **2012**, *55*, 711–719. [CrossRef]
50. Névéol, A.; Dalianis, H.; Velupillai, S.; Savova, G.; Zweigenbaum, P. Clinical natural language processing in languages other than english: Opportunities and challenges. *J. Biomed. Semant.* **2018**, *9*, 1–13. [CrossRef]
51. American Diabetes Association. 5. Facilitating behavior change and well-being to improve health outcomes: Standards of medical care in diabetes—2021. *Diabetes Care* **2021**, *44* (Suppl. 1), S53–S72. [CrossRef]
52. Unger, T.; Borghi, C.; Charchar, F.; Khan, N.A.; Poulter, N.R.; Prabhakaran, D.; Ramirez, A.; Schlaich, M.; Stergiou, G.S.; Tomaszewski, M. 2020 International Society of Hypertension global hypertension practice guidelines. *Hypertension* **2020**, *75*, 1334–1357. [CrossRef]