A Comparative Study of N-gram and Skip-gram for Clinical Concepts Extraction

Susan Sabra
Department of Data Science, Eurecom
Biot, France
susansabra22@gmail.com

Vian Sabeeh
Department of Computer Science and Engineering
Oakland University
Rochester, MI, USA
viants81@gmail.com

***Abstract*—State-of-the-art technologies for clinical knowledge extraction are essential in a clinical decision support system (CDSS) to make a prediction of a diagnosis. Automatic analysis of a patient's health data is a requirement in such a process. The unstructured part of the data in electronic health records (EHR) is critical, as it may contain hidden risk factors. We present in this paper a comparative study of two well-known techniques N-gram and Skip-gram to enhance the extraction of risk factors concepts from the clinical narratives after applying initial natural language processing (NLP) techniques. We evaluate the use of both techniques using a case study dataset of patients' records with venous thromboembolism (VTE). Results of the techniques' comparative study yielded an advancement of N-gram precision while Skip-gram produced a better performance in terms of the recall measure.**

Keywords— Clinical Decision Support Systems, MetaMap NLP, N-gram, Skip-gram, Knowledge Extraction.

I. INTRODUCTION

Natural Language Processing is at the heart of many researches in today's data analysis as it focuses on the users' contributions to Big Data through opinions and comments on social media platforms. Natural Language Processing (NLP) technique used in Stanford NLP APIs [1] is an approach that works out the grammatical structure of sentences by using a parser for tokenization, segmentation, stemming, stop words removal, and part-of-speech tagging (POS) [2]. Tokenization [3] is the process that splits the artifacts into tokens, whereas segmentation is the task of recognizing the boundaries of sentences. On the other hand, stemming [4] is the process of converting or removing inflected form to a common word form. In natural language processing (NLP), word embedding is a representation of high-dimensional dense vocabulary. It attracted attention in the last years in research that focused on information retrieval (IR) where the semantic aspect of data became the state-of-the-art for Big Data analysis. Word correlation techniques like N-gram [5] and Skip-gram [6] focus on the proper and context-relevant composition of multi-word expressions or concepts that can enhance the precision of extraction in certain results. In [7] they used N-gram technique after applying MetaMap NLP for extraction of medical concepts which proved successful by improving the extraction results [8].

Word embedding is a very popular technique that has been extensively used particularly with semantic-based studies [9, 10, 11, 12] as it can dramatically enhance the precision of the results. It transforms the learning method of any semantic model and reduces the distance measurement for the concepts in the context. The meaning of a word is insufficient anymore to identify specialized concepts such as medical and clinical concepts. The meaning of a word is influenced by the words that accompany it [13]. The repetitive co-occurrence of and distance between such words is transformed into a low-dimensional and continuous vector space. This technique provides a more robust codependence of these words.

Preventive Medicine focusing on analyzing the clinical narratives of a patient health record utilizes word codependence and embedding to ensure completeness and precision in concepts detection and identification.

An efficient word embedding approach proposed in [14] where two log-linear models: continuous bag of words (CBOW) and skip-gram are proposed to learn the neighboring relation of words in context. Later, the same authors introduced some modifications that largely improve the efficiency of model training in [15]. In this paper, we use both N-gram and Skip-gram techniques to compare the results of clinical concepts extraction from cardiovascular-focused clinical narratives. Our

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00151

conducted experiments produced extracted medical concepts as features for machine learning prediction through classification.

This paper is organized as follows: section I is an introduction to natural language processing techniques, section II is the background and literature review we prepared for this work, section III is the full description of our methodology and comparison between N-gram and Skip-gram, section IV is the evaluation and analysis of the experiments and section V is the conclusion of this paper.

II. RELATED WORK

As the focus of this work is on word embedding techniques and their impact on the studies' results, we will focus our literature review on a narrow scope of the techniques rather than their applications. N-gram, Skip-gram and CBOW are used to identify neighboring words in a context that enhance the identification of a more precise concept. Authors in [16] use N-gram on Turkish documents by using n-gram features. They apply different preprocessing techniques, namely, n-gram choice (character level or word level, bigram or trigram models), stemming, and use of punctuation to determine the Turkish document's author, genre, and the gender of the author. We use N-gram to extract word sequences to match medical concepts in medical ontologies in [17]. However, our previous results in [7] and analysis show that MetaMap NLP is not satisfactory to individually extract expressions of concepts that are composed of multiple words. This omission negatively impacted the whole framework as all other tasks depend on finding the risk factors concepts first. To enhance such an extraction process, an additional technique like N-gram is needed to deliver better results. N-gram is a text categorization technique based on using rank order statistics to compare text document and category document profiles of most frequent character n-grams. Experiments confirmed that this technique works very well for language categorization [5]. In [18] they proposed a novel machine learning based method for cytokine-receptor interaction prediction where a protein sequence is first transformed by incorporating the sequence evolutional information and then formulated by using the k-skip-n-gram model as one of their methods. One of the aspects they used was a combined N-gram and Skip-Gram model resulting in a better performance of prediction accuracy than existing methods. In [19] the authors used the continuous skip-gram model for entity mapping by creating a list of relevant features during feature extraction. The skip-gram model predicts words that can appear in the neighborhood of a selected word by computing the most similar and related keywords through the cosine similarity method discussed later in this paper. Term proximity, syntactic, or even semantic similarity are not always considered as the primary objective in various IR tasks depending on the context and the goal. Hence, terms *relevance* becomes the focus instead of all the before mentioned techniques. This is the motivation for developing unsupervised document relevance information in [20] where they proposed two learning models with different objective functions; one learns a relevance distribution over the vocabulary set for each query, and the other classifies each term as belonging to the relevant or non-relevant class for each query.

III. RISK FACTORS EXTRACTION SEMANTIC TECHNIQUES COMPARISON

In this section we present both techniques N-gram and Skip-gram for extracting medical concepts from clinical narratives.

A. N-gram Technique

When combined with MetaMap NLP, N-gram provides an enhanced extraction of medical concepts as proved in [7]. An N-gram is an N-character slice of a longer string. Although in the literature the term can include the notion of any co-occurring set of characters in a string [21]. Typically, it slices the string into a set of overlapping N-grams. We use N-grams of several different lengths simultaneously. For example, the expressions "oral contraceptive therapy" and "congestive heart failure" are 3-gram words that may not be matched with a concept in disease ontology if only the expressions "contraceptive therapy" and "heart failure" respectively were present in the text. Similarly, the example of "Chronic Obstructive Pulmonary Disease" will not be matched unless a 4-gram technique is used. Sometimes, abbreviations are used, and a 1-gram technique is sufficient to match it if it is present in

the ontology in such a form. Figure 1 illustrates the steps of N-gram process stemming after MetaMap NLP segmentation. N-gram module takes each sentence in a clinical narrative and sets window size initial value equal to the length of the sentence (considering the whole sentence as one long pattern). The module then compares the pattern to the concepts in UMLS [22] ontologies to find a match. If no matches were found, then the window size is decremented by 1. All possible patterns from the sentence are generated and stored in N-gram table for each new window size. Each N-gram table contains all the possible sequences of N words where N = window size - 1. Once a match is found then the pattern is added to the list of identified concepts. This updated list of concepts includes all risk factors concepts identified in the sentence.

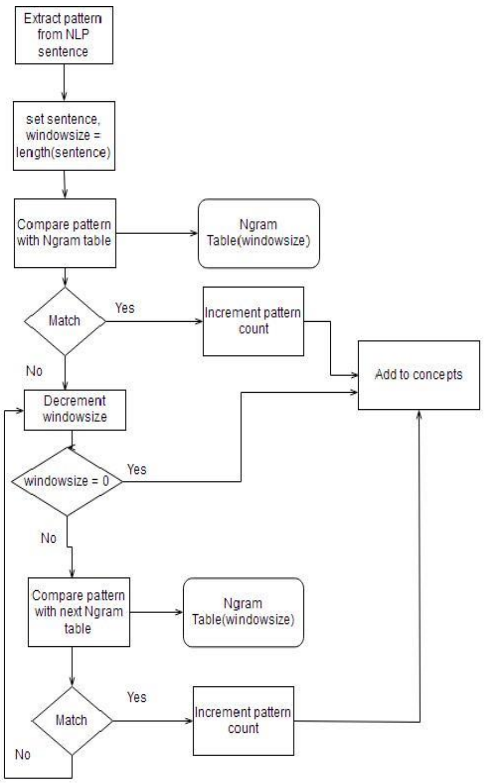

Fig 1: N-gram Diagram

B. Skip-gram Technique

Skip-gram is one of the unsupervised learning techniques used to find the most related words for a given word. It induces word embeddings by exploiting the signal from word-context co-occurrence [6]. Skip-gram is used to predict the context word for a given target word. It is the reverse of CBOW algorithm [23]. Skip-grams are considered as another means of discovering N-grams, or variants of N-grams, rather than a standalone method or an end in themselves.

The author in [24] use Skip-gram technique that employs the continuous skip-gram model along with the knowledge of information sources such as PubMed publicly available for collecting sequences from medical literature abstracts in the field of cardiovascular disease Venous Thromboembolism (VTE) for the purpose of this paper. These are used to help extract the contextually relevant keywords about VTE risk factors through the entities of the relevant medical concepts. Skip-gram employs the continuous skip-gram model to capture the surrounding keywords for the extracted keywords from PubMed sources with the reference of the extracted key sentences of the clinical narratives by mapping PubMed keywords with the key sentences of the clinical narratives. As a result, the continuous skip-gram model-based approach generates a set of keywords for all the medical concepts identified, which results in an enriched feature set. In [25] they defined Word2vec as a popular word-embedding approach that represents words on a fix-sized vector space model through either the skip-gram or continuous bag-of-words (CBOW) model. Word2vec is more effective in capturing semantic and syntactic word similarities from a huge corpus of text. The authors in [25] used Word2vec to construct a context sentence vector, and sense definition vectors then give each word sense a score using cosine similarity to compute the similarity between those sentence vectors. In our proposed model for words enrichment we use Skip-gram where we also apply the cosine similarity to compute the similarity between the extracted keywords in the sentences, which is described in Equation (1).

$$similarity = \cos(\theta) = \frac{A.B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i\,B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Figure 2 illustrates an example of Skip-gram with window size n = 2 where it finds the keywords neighboring the selected keyword and then going through the skipping method to find the correlation with other keywords depending on the window

size. The proposed approach computes the similarity for an enriched set of keywords using the cosine similarity measurement. Equation (1) calculates the similarity between the keywords

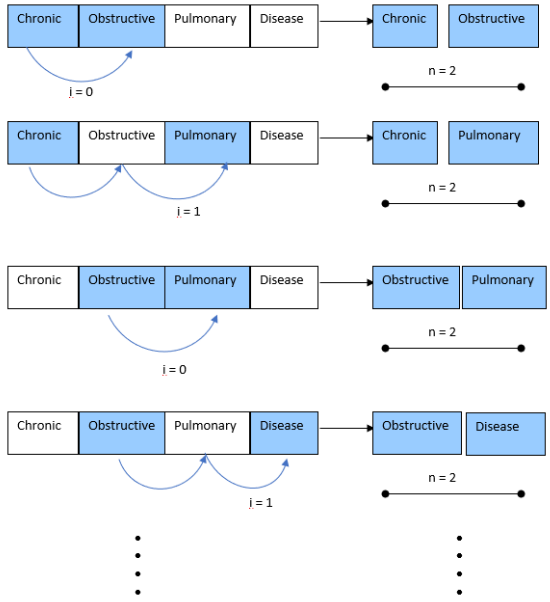

Fig 2: Skip-gram Example

generated by the continuous skip-gram model and other relevant words. In equation (1), 'A_i' and 'B_i' represent the two different features or keywords of a particular concept. 'i' refers to each keyword and 'n' denotes the total number of keywords in a particular medical concept. In accordance, the approach extracts the relevant keywords for each keyword based on the highest similarity.

IV. EVALUATION AND DISCUSSION

We implement our NLP to be followed by N-gram in the first experiment and by Skip-gram in the second one. For NLP we use UMLS/MetaMap. We experiment with the sequences of NLP with each of N-gram and Skip-gram techniques over 100 clinical narratives from patients' records that contain VTE risk factors. We use VTE as our case study for evaluating our model in this paper. Our data sources are two datasets: TREC CDS (Text Retrieval Conference – Clinical Decision Support) and I2B2 heart failure challenge [26, 27]. TREC CDS Track 2014 is a medical external resource which comprises of biomedical documents and clinical narratives, specifically a subset of PubMed. It is mainly used for the retrieval of biomedical articles answering generic clinical medical records. This dataset is the most VTE-relevant unstructured dataset that we use to analyze as clinical narratives portion from an EHR. We use both datasets with minimal preprocessing: segmentation only to separate sentences.

Considering TP as true positive, FP as false positive, FN as false negative, TN as true negative, P as Precision, and R as Recall as defined in [28], we calculate the F1-measure such as:

$$\text{F1 measure} = 2PR/(P+R) \quad (2)$$

where

$$P = TP / (TP+FP) \quad (3)$$

$$R = TP / (TP+FN) \quad (4)$$

TABLE I. N-GRAM VS. SKIP-GRAM PERFORMANCE

Sequences	P	R	F1
MetaMap NLP	63%	47%	54%
Seq 1 (NLP, N-gram)	81%	60%	69%
Seq 2 (NLP, N-gram, NLP)	85%	84%	85%
Seq 3 (NLP, Skip-gram)	60%	89%	72%
Seq 4 (NLP, Skip-gram, NLP)	63%	89%	74%

The sequences in Table 1 are:
Seq 1: MetaMap NLP, N-gram
Seq 2: MetaMap NLP, N-gram, MetaMap NLP
Seq 3: MetaMap NLP, Skip-gram
Seq 4: MetaMap NLP, Skip-gram, MetaMap NLP

These sequences show the order of applying the different techniques of NLP. The first use of MetaMap NLP applies the different POS, stemming, tokenization etc. except the negation while the second time of using MetaMap NLP after each N-gram and Skip-gram is simply for identifying the negation in the sentences for the purpose of eliminating the negated expressions or concepts from showing in the results. We can notice that each technique has an advantage over the other one in either precision or recall. The precision of N-gram shows a big improvement after applying the initial steps of MetaMap NLP in terms of both precision and recall in seq 1. The negation adds to that improvement showing in the results for seq 2. While the experiment with Skip-gram introduces an unexpected drop in the precision with an elevated recall performance. We can deduce that the sequence with Skip-gram yielded a better recall in seq 3 because of the good rate of FN that was close to null. We believe skip-gram precision's drop can be related to the cosine similarity evaluation as it is based on the cooccurrence and frequency of the full medical concepts as they show in the clinical narratives or

vice versa. A misrepresentation of the concept in the clinical narrative is based on the PubMed findings and not an expert's input so we believe depending on the selection of articles in PubMed, it can lead to low similarity impacting the final results. Skip-gram requires more memory to perform better. Another limitation on Skip-gram is the window size which might not include the anticipated associated words, so they remain undiscovered based on the cosine similarity formula. In addition, existing skip-grams searches may require a somewhat adjusted formulation which can be cumbersome. Hence, we can confirm under the current settings that N-gram can guarantee a better overall F1 measure under the current circumstances with the used dataset. This confirms the statement that Skip-grams are considered as another means of discovering N-grams, or variants of N-grams, rather than a standalone method or an end in themselves.

V. CONCLUSION

EHRs today contain critical and valuable information that is required to build a precise diagnosis for a patient in a personalized medicine focused paradigm. However, the presence of the clinical knowledge is not enough as it is presented in various forms in these electronic health records. Retrieval of such knowledge is key for the improvement of the aforementioned diagnosis paradigm. The more precise the extraction, the better are the results of the diagnosis. In this paper, we present a comparative study of two sequences of language processing techniques, namely N-gram and Skip-gram, combined with semantic web technologies to extract the risk factors concepts of VTE from clinical narratives. Our results confirm that Skip-grams are considered as another means of discovering N-grams, or variants of N-grams, rather than a standalone method. The proposed sequence of "MetaMap NLP, N-gram, and MetaMap NLP" techniques yields better results for precision of 85% and recall of 84%. This trio sequence proves to be better than other sequences or use of individual techniques in a standalone mode for extracting risk factors concepts necessary for CDSS to make a diagnosis or a treatment plan.

VI. FUTURE WORK

We envision an extension and a variation of this work to experiment with the proposed approaches in a combinatory form to evaluate as well whether the used NLP techniques can when combined produce better results than their individual use. On the other hand, it would be also interesting to evaluate the same approaches proposed here on other diseases ontologies especially with Skip-gram depending on PubMed content as a first step which might impact the results based on the selection of PubMed articles used.

REFERENCES

[1] https://stanfordnlp.github.io/CoreNLP/

[2] Lovins, Julie B. "Development of a stemming algorithm." (1968): 22-31.

[3]https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html .

[4]https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html .

[5] J. Graovac, "Text Categorization Using n-Gram Based Language Independent Technique," Intelligent Data Analysis, vol. 18, no. 4, pp. 677–695, 2014.

[6] J. Eisner, B. V. Durme, R. Cotterell, and A. Poliak, "Explaining and Generalizing Skip-Gram through Exponential Family Principal Component Analysis," 2017, doi: 10.18653/V1/E17-2028.

[7] S Sabra, K Mahmood, M Alobaidi "A Semantic Extraction and Sentimental Assessment of Risk Factors (SESARF): An NLP Approach for Precision Medicine: A Medical Decision Support Tool for Early Diagnosis from Clinical Notes" *Computer Software and Applications Conference (COMPSAC), 2017 IEEE*, Vol. 2, pp. 131-136. 2017

[8] MetaMap. https://metamap.nlm.nih.gov/. Accessed on 7 October 2020.

[9] E. Huang, R. Socher, C. Manning, and A. Ng, "Improving Word Representations via Global Context and Multiple Word Prototypes," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jeju Island, Korea, Jul. 2012, pp. 873–882, Accessed: Oct. 07, 2020. [Online]. Available: https://www.aclweb.org/anthology/P12-1092.

[10] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, Feb. 2011.

[11] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[12] J. Turian, L.-A. Ratinov, and Y. Bengio, *Word Representations: A Simple and General Method for Semi-Supervised Learning.*, vol. 2010. 2010, p. 394.

[13] M. Baroni and R. Zamparelli, *Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space.* 2010, p. 1193.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Jan. 2013, Accessed: Oct. 19, 2017. [Online]. Available: https://arxiv.org/abs/1301.3781.

[15] T. Mikolov, Q. Le, and I. Sutskever, "Exploiting Similarities among Languages for Machine Translation," Sep. 2013.

[16] A. Deniz and H. E. Kiziloz, "Effects of various preprocessing techniques to Turkish text categorization using n-gram features," in Computer Science and Engineering (UBMK), 2017 International Conference on, 2017, pp. 655–66.

[17] S. Sabra, M. Alobaidi, K. M. Malik and V. Sabeeh, "Performance evaluation for semantic-based risk factors extraction from clinical narratives," 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, 2018, pp. 695-701, doi: 10.1109/CCWC.2018.8301742.

[18] L. Wei, Q. Zou, M. Liao, H. Lu, and Y. Zhao, "A novel machine learning method for cytokine-receptor interaction prediction," *Combinatorial chemistry & high throughput screening*, vol. 19, Nov. 2015, doi: 10.2174/1386207319666151110122621.

[19] V. Sabeeh, M. Zohdy, and R. A. Bashaireh, "Enhancing the Fake News Detection by Applying Effective Feature Selection Based on Semantic Sources," *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, doi: 10.1109/CSCI49370.2019.00255.

[20] H. Zamani and W. B. Croft, “Relevance-based Word Embedding,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, Aug. 2017, pp. 505–514, doi: 10.1145/3077136.3080831.

[21] W. B. Cavnar and John M. Trenkle, “N-Gram-Based Text Categorization.” [Online]. Available: http://odur.let.rug.nl/~vannoord/TextCat/textcat.pdf. [Accessed: 25-Nov-2017].

[22] Bodenreider, O., 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl 1), pp. D267-D270.

[23] K. Orkphol and W. Yang, “Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet,” *Future Internet*, vol. 11, no. 5, p. 114, May 2019

[24] https://pubmed.ncbi.nlm.nih.gov/

[25] K. Orkphol and W. Yang, “Word Sense Disambiguation Using Cosine Similarity Collaborates with Word2vec and WordNet,” *Future Internet*, vol. 11, no. 5, Art. no. 5, May 2019, doi: 10.3390/fi11050114.

[26] Sungbin, S Choi, and Choi, J. (2014).. *SNUMedinfo at TREC CDS track 2014: Medical case-based retrieval task. SEOUL NATIONAL UNIV (REPUBLIC OF KOREA), 2014.*

[27] Stubbs, A., Kotfila, C., Xu, H., and Uzuner, O. (2015). "Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2". *J Biomed Inform*. 2015 Jul 22. pii: S1532-0464(15)00140-9. doi: 10.1016/j.jbi.2015.07.001.

[28] J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–24.