Lecture Notes in Electrical Engineering 996

Yue Wang
Yuyang Liu
Jiaqi Zou
Mengyao Huo *Editors*

Signal and Information Processing, Networking and Computers

Proceedings of the 10th International Conference on Signal and Information Processing, Networking and Computers (ICSINC)

ICSINC

Springer

Lecture Notes in Electrical Engineering 996

Series Editors

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

Yue Wang · Yuyang Liu · Jiaqi Zou ·
Mengyao Huo
Editors

Signal and Information Processing, Networking and Computers

Proceedings of the 10th International Conference on Signal and Information Processing, Networking and Computers (ICSINC)

Springer

Editors
Yue Wang
China Academy of Space Technology
Beijing, China

Yuyang Liu
Beijing University of Posts
and Telecommunications
Beijing, China

Jiaqi Zou
Beijing University of Posts
and Telecommunications
Beijing, China

Mengyao Huo
Beijing University of Posts
and Telecommunications
Beijing, China

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-19-9967-3 ISBN 978-981-19-9968-0 (eBook)
https://doi.org/10.1007/978-981-19-9968-0

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

It is our great honor to welcome you to the 10th International Conference on Signal and Information Processing, Network and Computers (ICSINC 2022). The ICSINC 2022 Committee has been monitoring the evolving COVID-19 pandemic. We have decided to delay the 2022 edition of this conference from July to September.

ICSINC 2022 provides a forum for researchers, engineers, and industry experts to discuss recent development, new ideas and breakthrough in signal and information processing schemes, computer theory, space technologies, big data, and so on.

ICSINC 2022 received 314 papers submitted by authors, in which 150 papers were accepted and included in the final conference proceedings. The accepted papers were presented and discussed in 27 regular technical sessions and 2 workshops.

On behalf of the ICSINC 2022 committee, we would like to express our sincere appreciation to the TPC members and reviewers for their tremendous efforts. Especially, we appreciate all the sponsors for their generous support and advice, including Springer, China Unicom, HuaCeXinTong company. Finally, we would also like to thank all the authors for their excellent work and cooperation.

Songlin Sun
Yue Wang
Changbo Zhu
ICSINC 2022 General Chairs

Committee Members

International Steering Committee

Songlin Sun	Beijing University of Posts and Telecommunications, China
Ju Liu	Shandong University, China
Chenwei Wang	DOCOMO Innovations (DoCoMo USA Labs), USA
Jiaxun Zhang	China Academy of Space Technology, China
Takeo Fujii	The University of Electro-Communications, Japan
Hongwu Li	China Unicom Research Institute, China
Xiaoyu Ye	China Unicom Research Institute, China
Changbo Zhu	China Unicom QingHai Branch, China
Fang Wang	China Unicom QingHai Branch, China

General Co-chairs

Songlin Sun	Beijing University of Posts and Telecommunications, China
Yue Wang	China Academy of Space Technology, China
Changbo Zhu	China Unicom QingHai Branch, China

Technical Program Committee Chairs

Xinzhou Cheng	China Unicom Research Institute, China
Qiang Zhang	China Unicom QingHai Branch, China

Publication Chair

Yuyang Liu	Beijing University of Posts and Telecommunications, China

Exhibits Chair

Lianxin Xin	China Unicom QingHai Branch, China

Local Service Committee

Zhijuan Tao	China Unicom QingHai Branch, China
Wei Shang	China Unicom QingHai Branch, China

Sponsors

Springer

China Unicom

HuaCeXinTong

Contents

Big Data Workshop Session 1

Information Technology Session 1

A New Multi-instance Learning Algorithm Integrated with TF-IDF for Entity Relation Extraction in Electronic Medical Records

Yiyang Xiong(✉), Yajuan Qiao, Shilei Dong, Xuezhi Zhang, and Hua Tan

China Telecom Corporation Limited Research Institute, Beijing, China
xiongyy1@chinatelecom.cn

Abstract. Electronic medical record (EMR) is the core data which generated in the process of clinical treatment. By extracting the record text's entity relationship, we can catch a large amount of information closely related to the patient's condition. Deep learning is widely used in the field of entity relationship extraction in EMR. Among them, seq-to-seq is the most popular network structure for text sequence, but it can't distinguish the influence between different words. Therefore, multi-instance learning which are improved by attention mechanism can adjust the weight of words is proposed and achieved outstanding results. However, the reason for the creation of the attention mechanism parameter is always puzzling: it lacks interpretability and it is difficult for us to understand the origin of the numbers. Term frequency–inverse document frequency (TF-IDF) is an algorithm to calculate the importance of words according to their occurrence frequency, so we think integrating it into the construction of neural network may increase model's interpretability. In this paper, we propose a new multi-instance learning algorithm, which integrates the TF-IDF into the construction of deep learning model. Compared with the network without TF-IDF, our model achieves an improvement of nearly 2% and the convergence rate is reduced by 5% on average.

Keywords: Multi-instance learning · TF-IDF · Relation extraction

1 Introduction

1.1 Background

In daily medical activities, doctors, nurses and other medical personnel usually use the medical information system to generate multiple types of medical records which is called electronic medical records (EMR). EMR records the diagnosis and treatment process and medication suggestions in detail, which has high research value [1]. With gradual improvement of medical informatization, EMR is becoming more and more popular, and gradually plays a more important role in clinical diagnosis and treatment.

In EMR natural language processing, entity relation extraction is the core tasks. It refers to obtain the relationship between known entities. For example, here is a text: doctors treat patients. This text has a pair of entities of doctors and patients, and the entity relationship is 'treat' (the relationship of 'treat' needs to be defined in advance).

Y. Wang et al. (Eds.): ICSINC 2022, LNEE 996, pp. 3–10, 2023.
https://doi.org/10.1007/978-981-19-9968-0_1

Deep learning not only has strong ability, but also can solve problems end-to-end, so it become mainstream method of EMR entity relation extraction. Deep learning applied to EMR can be roughly divided into supervised learning algorithm and unsupervised learning algorithm. Supervised learning algorithm usually adopts convolutional neural network which represented by textCNN [2] and PCNN [3], it can obtain excellent performance [4], but these methods need a large number of positive samples, so the cost of large-scale learning is high. Semi supervised learning can make up for this problem because it can quickly obtain a large number of annotation data. The commonly used semi supervised learning include remote supervision learning [5] and multi-instance learning [6]. Among them, multi-instance learning introduces attention mechanism, so the weight of context can be adjusted, and it achieves better results and become the most popular algorithm.

1.2 Our Innovation

The introduction of attention mechanism greatly increases the processing ability of neural network for long text [7]. Transformer, which abandons neural network and adopts attention mechanism completely [8], even leads the research direction of natural language processing. Because the text length of EMR is generally long, it is particularly effective. However, the initial parameters of attention mechanism are often set randomly, which brings two problems: 1. The algorithm has great randomness in convergence. 2. Insufficient interpretability, so it is difficult to realize improvement in the future. In addition, the embedding layer of neural network lacks meaningful annotation of data: it usually only has location information. So, if we can set the initial parameters of neural network and attention mechanism through the understanding of the text, can we create a better model?

TF-IDF is a very mature technology in the field of text search [9]. It can judge the importance of a word in the text through word frequency. In this paper, we introduce its calculation results into the initial parameters of model, and propose a new multi-instance learning algorithm for EMR. By doing this, we can deliver neural network the importance of each word before training starts. Compared with the network without integration, the convergence speed of our proposed algorithm has been improved by more than 5%, and the accuracy has also been improved.

2 Model Architecture

2.1 Overview

When encountering deep learning problems, the ideal state is that the data is either black or white, such as whether a picture is a face. However, in practice, the samples we have are difficult to meet the situation above, especially for the problem of entity relationship extraction: when any two entities appear at the same time, the relationship is usually not unique. For example, there are two sentences: doctors treat nurses and nurses cooperate with doctors. Nurses and doctors have different relationships in these two sentences: treatment and cooperation. Remote supervision method and the multi-instance learning belong to semi supervised learning which can solve the problem above.

Since multi-instance learning can significantly reduce the impact of false tagging in Distant Supervision method [10], we adopt its framework.

The multi-instance learning algorithm makes a hypothesis: if an entity pair has a certain relationship in the knowledge base. Then, in the data containing the entity pair, at least one sentence expresses this relationship. Unlike ordinary supervised learning, which aims at a single sample, multi-instance learning is bag oriented, and a bag contains multiple samples. It has two definitions: 1. If one sample in a bag is a positive sample, the whole bag is marked as a positive bag. 2. If a bag is all negative samples, then the bag is negative bag. Let me give you an example to understand this. For example, we can judge whether there is gold ore from the excavated ore. If there is gold in the trial excavated ore, it is considered that there is gold ore here. If none of the trial excavated ore contains gold, it is considered that there is no gold ore here.

Therefore, for the problem of electronic medical record, we adopt multi-instance learning algorithm. Our multi-instance learning algorithm consists of two parts: neural network part and attention mechanism part. Our innovation integrates the results of TF-IDF calculation into the construction of two modules. Through our innovation, our model has the ability to transmit the importance of each word to the neural network from the perspective of word frequency, which quickly improves the convergence speed and slightly improves the accuracy. Next, we will introduce the construction of the model in detail.

2.2 Neural Network

In terms of neural network, we adopt Piecewise Convolutional Neural Network (PCNN) as basic framework. The TF-IDF PCNN model we proposed has following framework. Different colors in the figure represent the results of different convolution kernels. In addition, we use blue to represent the starting position of the entity.

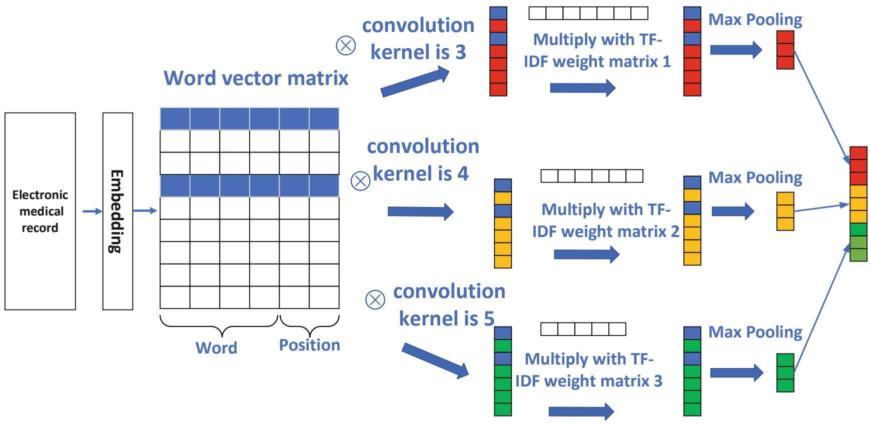

Fig. 1. The figure shows the structure of the neural network. It consists of five parts: word embedding, convolution, TF-IDF weight adjustment, Max pooling and splicing.

As shown in Fig. 1, We use a common sentence in the electronic medical record to describe how to deal with this example and convert it into sentence vector. Next, we will elaborate the processing process of each step through the example.

Word embedding. We first need to convert the text into a word vector that can be understood by computer and neural network. We take a common Chinese text in electronic medical record as an example. Here is a Chinese sentence: "医生和护士相互协作". Since it consists of 9 rows (each row represents a word), and the number of columns is set to 6. This text is transformed into a 9 * 6 matrix.

There are two kinds of word embedding methods. One is that word vector can be obtained in other corpora in advance, and the other is obtained by network training as unknown parameters [11]. These two methods have their own advantages. The first method can use other corpora to get more a priori knowledge. The second method can grasp the features associated with the current task better. In order to make good use of the advantages of both, our model adopts a dual channel method: one of them is embedded in pre trained words and will not change; The other is initialized in the same way, but will be changed with the training process. Please note that since we use the method of piecewise convolution, we need to mark the position of the entity and record it in the word vector (the position part of the vector). The starting position of doctors and nurses is marked here and highlighted in blue in Fig. 1. It can be described as the following formula.

$$S = R^{r*c} \tag{1}$$

R is the length of the sentence (number of words), C is the width of the convolution kernel, and the size of C is determined by position coding and word embedding.

Convolution. The biggest feature of convolution in natural language processing is the width of convolution kernel is the same as that of sentence matrix, and it can only move in the height direction, which reflects words are the smallest constituent unit of sentences. Therefore, in the example given in this paper, the width of convolution kernel is 6. In terms of length, since convolution kernels of different sizes can learn complementary features from each other [12], we set up three convolution kernels with lengths of 3, 4 and 5, and obtained convolution results respectively. The contents can be described by the following formula.

$$C \in R^{r+w-1}, C_j = wq_{j-w+1:j} \quad s + w - 1 \geq j \geq 1, \tag{2}$$

W is the size of convolution. Since we use three convolution kernels, so j is equal to three and the convolution output is:

$$C = \{C_1, C_2, C_3\}, C_{i3} = w_i q_{3-w+1:3} \tag{3}$$

TF-IDF adjustment. Here is an important embodiment of our TF-IDF in innovation. In the construction of word vector, each word is embedded according to the way of pre training or neural network convergence, and the neural network treats them equally. However, we should understand that the information behind each entity is actually different. If a word appears infrequently in the corpus and suddenly appears in our sentences, it may bring more information and deserve our more attention. TF-IDF reflects this fact. Its formula consists of three parts:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4}$$

TF (Term Frequency) indicates the frequency of the entry in the text, where $n_{i,j}$ indicates the number of times the entry t_i appears in the document d_j, and $TF_{i,j}$ indicates the frequency of the entry t_i in the document d_j.

$$IDF_i = log \frac{|D|}{1 + \left|j : t_i \in d_j\right|} \tag{5}$$

IDF (inverse document frequency) indicates the prevalence of keywords, where |D| indicates the number of all documents and |j: $t_i \in d_j$| indicates the number of documents containing the term t_i.

$$TF - IDF = TF * IDF \tag{6}$$

TF-IDF is obtained by multiplying formulas (1) and formulas (2). We can infer that TF-IDF tends to filter out common words and pay attention to the words that appear frequently in the text and are rare in the corpus.

We calculate TF-IDF value of each word according to word frequency and combine them to form TF-IDF vector, which can reflect the importance of each word in the sentence. We multiply it with the result of convolution:

$$C_t = TF * IDF * C \tag{7}$$

Now, we can tell the neural network which words need more attention.

Max pooling and splicing. Finally, we use pooling layer to reduce the complexity of the system, and splice the results of different convolution kernel calculations. A complete sentence vector is now complete.

2.3 Attention Mechanism

Multi-Instance Learning usually produces a lot of false data, we need to reduce the impact of noise. In addition, sentences deserve different attention due to its word's frequency. Based on the problems above, we propose a new attention mechanism improved by TF-IDF. The specific framework is shown in Fig. 2. Our attention mechanism is sentence-level.

We creatively calculate the contribution of each sentence to the text through TF-IDF, which means that the model has an initial parameter with physical meaning. In addition, the data oriented by multi-instance learning has many wrong labels, so it is also necessary to reduce the impact of these wrong labels through the attention mechanism. The formula of the attention mechanism is as follows:

$$\beta = \frac{exp(e_i)}{\sum_k exp(e_k)} \tag{8}$$

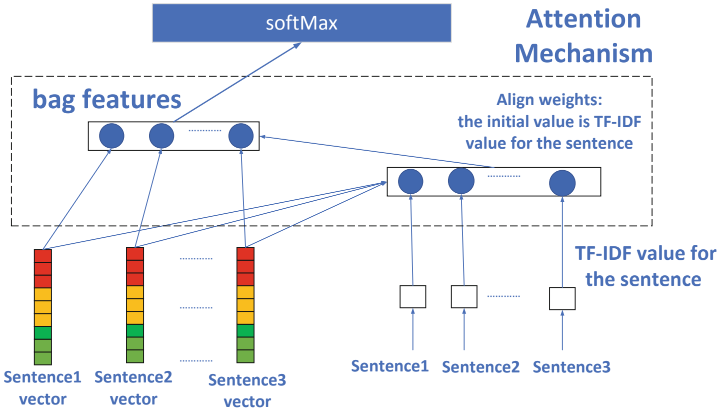

Fig. 2. The figure describes our sentence-level attention mechanism. Its initial parameters are calculated by TF-IDF. Its calculation will get the final classification result through SoftMax.

Among them, e_i is called a query function, which scores the matching degree between the input sentence x_i and the prediction relationship r and its specific expression is as follows, TF-IDF determines its initial value, where A is a weighted diagonal matrix:

$$e_i = x_i \boldsymbol{r} * TF - IDF \ * \boldsymbol{A} \tag{9}$$

Finally, we define the output of SoftMax according to [13]:

$$p(r|S) = \frac{exp(o_r)}{\sum_{k=1}^{n_r} exp(o_k)} \tag{10}$$

Among them, n_r is the total number of relations in the corpus and o is the output of neural network.

3 Experiment and Analysis

We are committed to the informatization of the medical industry. Therefore, the experiment of this paper is oriented to electronic medical records. Related work is usually based on the English data set. The difference of this paper is that the experiment is carried out for the Chinese text. The data set we use is from Ruijin Hospital in Shanghai. The content of the dataset is diabetes clinic information and related academic papers provided by Ruijin hospital. It has the entity information that has been tagged, so we can extract entity relationship directly. There are ten kinds of entity relations defined in the dataset. Five of them are related to diseases: Test Disease, Symptom_Disease, Treatment_Disease, Drug_Disease, Anatomy_Disease. The remaining categories are related to drugs: Frequency_Drug, Duration_Drug, Amount_Drug, Method_Drug, SideEff_Drug.

Our experiment evaluates the model ability by using Precision, Recall and F1, we compare our proposed algorithm with multi-instance learning algorithm (PCNN + ATT),