SSACBKM: An Integration Model for Biomedical Relationship Extraction

Huiping Shi

Yunnan University

Xiaobing Zhou (✉ zhouxb@ynu.edu.cn)

School of Information Science and Engineering, Yunnan University, 650091 Kunming, China

Research article

Keywords: Biomedicine, Bert, DDI, Chemprot

Posted Date: October 8th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-86122/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

SSACBKM: an integration model for Biomedical relationship extraction

Huiping Shi^{*} and Xiaobing Zhou[†]

^{*}Correspondence:
heidi0kai@gmail.com
School of Information Science and
Engineering, Yunnan University,
650091 Kunming, China
Full list of author information is
available at the end of the article
[†]Equal contributor

Abstract

Background: The cause of the disease is one of the main contents of biomedical research. Extracting effective relational information from a large number of biomedical texts has important applications for biomedical research. At present, most of the work of biomedicine is to use manual screening or use rule-based or feature-based pipeline network models to obtain screening characteristics. These methods require a lot of time to design specific rules or features to complete specific tasks, resulting in some features that non-compliant features cannot be filtered out.

Results: The model gets micro-F1 scores of 0.802 and 0.876 on the Chemprot data set and DDI data set, respectively. The resources that can be used in this project can be found in <https://github.com/HunterHeidy/DDICPI->.

Conclusions: Experiments have proved that without Bert, you can get good results by learning from Bert's core ideas.

Keywords: Biomedicine; Bert; DDI; Chemprot

Background

Background

Currently, the number of biomedical texts is growing rapidly, and biomedical text mining is becoming more and more important. On average, more than 3,000 new articles are published daily in peer-reviewed journals, excluding preprints and technical reports in various archives, such as clinical trial reports [1]. Although the accuracy of manual screening of biomedical texts is high, traditional manual inspection methods consume a lot of labor, material resources and time, and are more susceptible to natural factors. Therefore, the reform of biomedical text mining becomes more and more important. Natural language processing technology continues to mature, giving biomedical scientists more possibilities. Relation extraction is an important research topic in natural language processing, and its main purpose is to extract the semantic relations between marked entity pairs in sentences. That is, on the basis of entity recognition, the study of relationship extraction determines the relationship category between entity pairs in unstructured text, and forms structured data for storage and access [2]. For example, in "diffuse alveolar hemorrhage easily combined with lung infection", "diffuse alveolar hemorrhage" and "pulmonary infection" are both diseases, and the relationship between them is "disease-complication". The

earliest prototype of relationship extraction was proposed by Hirst. His "automatically obtain synonyms" natural language model uses the rule of "is a" to complete the relationship extraction. This is the earliest relationship extraction model[3]. For biomedical relationship extraction research, In the past, the most common process was to use word segmentation tools to process the original text and then use SVM[4], etc. combined with a large number of manual rules to remove medical events and predict the attributes of medical events. For example, Roberts[5] used the SVM classifier in the clinical information extraction system to realize the recognition of the relationship. In this research, he tried to recognize the relationship across sentences, but the accuracy rate obtained was very low. There are many reasons. For example, the model cannot fully obtain the hidden information; SVM cannot capture the hidden information of the long-distance mountain below; and the creation of artificial rules has a great relationship with specific fields, and there is a limitation that the general type is not strong etc. In recent years, more and more researchers have applied RNN, RNN variants, CNN, and CNN variants to biomedical relationship extraction research, and achieved good results. In 2017, Zheng et al, used the combination of pos model and BiLSTM model to apply to the DDI(Drug Drug Interaction) dataset, and obtained a result of 0.77 [6]. These research results point out the direction of biomedical research. More and more researchers are applying NLP models to the biomedical field. A large number of biomedical texts and textual professionalism provide a more suitable platform for NLP. The current NLP models are usually developed based on word embedding[7], which allows them to learn similar representations of semantically similar words. Compared with traditional feature-based models, this is a big improvement. However, there are still shortcomings to be resolved. First, most biomedical texts are long sentences. In NLP, the effect of the model is closely related to the actual sentence length[8]. A typical model sequentially performs recursive calculations on the input and output sequences according to the relative order of the positions of the data to be processed and converts the relative positions of characters into the time step of the model calculation at a time. The method of sequential calculation following the time series limits the application of the parallel calculation of training data. Therefore, the calculation of long sequences will limit the batch input of training data due to memory. Secondly, biomedical texts are highly logical. Different positions of words in the text have obvious differences in the meaning of the text. Given the current status quo, this paper proposes to optimize SparseTransformer and combination capsule processing text, use BiLSTM to process keywords, and add them to the model as supplementary features. Finally, in the model screening, top three features in each window representative weights are selected as the expression of the word.

Related work

The research on biomedical information relationship extraction is divided into two categories, the traditional statistical method and the deep learning neural network method. For the traditional statistical method, a common process is to use word segmentation tools to process the original text and then use Conditional Random Fields (CRFs)[9], SVM etc. combined with a large number

of manual rules to extract medical events and predict the attributes of medical events. Recent year, an increasingly researchers have applied deep learning methods such as RNN, RNN variants, CNN, and CNN variants to the research of biomedical relation extraction, and achieved good results [10–12]. For example, Zhao et al[10], apply multi-layer CNN to extract PPI corpus and evaluate the effectiveness of multi-layer CNN in the PPI extraction task. The RNN model is considered more suitable for automatically learning features from long sentences. In 2016, Liu et al tried to use the CNN model in the DDI corpus and got an F1-score of 0.6795[11], which was 0.027 higher than the existed best model. In 2018, Zhang et al. [12] combined RNN with CNN, integrated them in a form of complementary advantages, applied them to PPI and DDI datasets, and obtained F1-scores 0.7020 and 0.7290, respectively. Andre Lamurias et al, proposed an RNN-LSTM model based on a variant of RNN [13], a long short-term memory network (LSTM) model. As a variant of RNN, the LSTM network mainly adds a gating unit to alleviate the gradient explosion phenomenon of RNN. Attention was first applied to the image field. In 2014, attention was modified by Bahdanau et al, and used in [14] and was applied to machine translation. Their work is regarded as the first to apply the Attention mechanism to the NLP field. Then, the Attention mechanism is widely used. Combined with neural network, deep learning pioneer Hinton proposed Capsules[15]. Hinton's understanding is: each capsule represents an attribute, and the vector of the capsule represents the "frame" of this attribute. In other words, we used to use a scalar to indicate the presence or absence of this feature (such as whether there is a feather), and now we use a vector to indicate that it not only indicates whether there is but also indicates "what kind of" (such as what color, what texture feathers). This opened a new path for later neural network development. In 2017, a research by Google [16] turned the research direction of natural language processing to the attention mechanism. The core idea: set the Q sequence, K sequence, and V sequence as the same text into multiple attention mechanisms for learning [17]. Google's discover revealed the superior performance of the attention mechanism in natural language processing. In 2018, Google released Bert[18] again, sensationalizing the entire natural language processing academia. Bert used the Transformer structure to construct a multi-layer two-way encoder network, and the Transformer structure of Bert directly used the structure in Transformer[16]. A multilayer bidirectional encoder network is constructed using the Transformer structure [16]. Lots of researcher find that the pre-training model of Bert has very good effects on various tasks.

In 2013, DDI corpus (drug-drug interactions) was proposed by Maria Herrero-Zaza, and the baseline was F1-mario 0.65 [19]. In 2016, Bio Creative used DDI corpus in Task V [20]. In 2019, Andre et al developed the Bo-lstm [21] algorithm based on BiLSTM, which increased the baseline by 0.004. In the same year, Xia et al got f1 0.75 based on CNN and pooling [22]. In 2019, Peng et al developed a pre-training model bio-Bert suitable for biomedicine based on Bert and trained on PubMed to get the best result of 0.79.9 [1]. Nguyen et al developed an R-Bert model based on Bert and got a score of 0.809 [23]. Chen et al proposed that this is concept embedding, focusing on the main biological concepts mentioned in the biomedical literature, F1-macro reached 0.8105.

Peng et al, based on the Bert multi-task learning method, applied the training weights from other similar data sets to the DDI corpus, and obtain an F1-macro result of 0.82 [24].

The ChemProt[25] (chemical-protein interaction) corpus, a disease chemical biology database, based on a variety of chemical-protein annotation resources and a compilation of disease-related protein-protein interactions (PPI)[26]. In 2019, Sun et al developed Bert-Att-Capsule based on Bert and got a score of F1 0.74 [27]. Peng et al. proposed a multi-task training Bert, which achieved an F1-macro score of 0.72 on ChemProt [24]. Lee et al. obtained a result of 0.779 on ChemProt based on pre-training Biobert[1].

Concerning extraction tasks, the data of most tasks are highly unbalanced. Data imbalance causes data sparseness problems. Due to the lack of sufficient data, the classifier's ability to describe sparse samples is insufficient, and it is difficult to effectively classify these sparse samples. In one task, if there is a serious imbalance in sample proportions, the model can use data enhancement methods to alleviate this phenomenon. In response to this phenomenon, the mitigation method is random sampling in preprocessing. The biggest advantage of random sampling is simplicity, but the disadvantages are also obvious. After upsampling, some samples will appear repeatedly in the data set, and the model will partially over fitted. The disadvantage of downsampling is that the training set loses part of the data, and the model only learns part of its features [28–31]. Some models improve learning algorithms, giving fewer data more weight, or using integrated learning to ease[28]. However, this will add parameters to the model and reduce the generalization of the model.

Methods

Overview of the SSCBKM model

In this paper, we propose an SSCBKM model, which combines with SquashSparseTransformer, Capsules, BiLSTM and K-max-Flatten, as shown in figure 11. This model uses SquashSparseTransformer block to process sentences in parallel with Capsule block. BiLSTM processes keywords and the results obtained are added to sentence as supplementary features. Finally, after K-max-Flatten feature screening, the top k features of each window are screened out, and the final result is obtained through Softmax layer constraint classification.

Embedding layer

The model introduces word2vec pre-training word vectors. The word2vec pre-training word vector is a bag of words obtained by processing Pubmed and PMC corpus (biomedical corpus) using the skip-gram algorithm in advance. The skip-gram algorithm and CBOW algorithm obtains the vector representation of the word through the context of the word (one-hot representation)[32]. Skip-gram generates background words based on the central word. CBOW generates the centre word based on the background word[33]. Why use word2vec? Words with similar semantics have similar contexts, allowing our neural network to produce similar output vectors for similar words during the training process. The network can generate similar word vectors for words with similar

semantics. Because the output layer at this time has been trained and will not change. This determines why the neural network model gets accurate classification results when training weights. Although word vectors are also generated in the middle, the word vectors are a by-product. Word2vec is fast in training and effective, and it occupies much less memory than GloVe. For our experimental equipment, word2vec is more suitable for us. In the DDI data set, there are a total of 5748 words, of which 401 is not included. In the Chemprot data set, there are a total of 10848 words, the unlisted words are 1325. Since there are relatively few unregistered words, the model ignores unregistered words.

Input layer

We introduce the mask mechanism to mark the meta data. In the process of data preprocessing, the previous practice is to use word embedding to process characters and convert them into digital forms that can be processed by the computer. Such as:

$S = \text{"this is an tihistamine"}$

Text vector representation

$E_s = [2, 0, 30]$

In the text processing process, the sentence length must be equal. To obtain all the information, the text is usually processed to a fixed maximum length, and "0" is used for sentences with insufficient length. For example, in sentence S, suppose the maximum text length is 5, padding with '0', $E_s = [2, 0, 30, 0, 0]$. Since the model cannot distinguish whether the data "0" seems to be a valid value, we calculate average $A_{E_s} = (2 + 0 + 30 + 0 + 0)/5 = 6.4$. However, the real thing in the data is the average value $A_{E_s} = (2 + 0 + 30)/3 = 10.3$, which has a serious deviation. In order to solve this difference, the algorithm introduces the MASK mechanism [18] to mark the valid value of the text. A bool vector $m = [1, 1, 1, 0, 0]$ is created to mark the valid values in S. The mean value is calculated $A = \text{Sum}(E_s \text{ bigotimes } m) / \text{Sum}(m)$.

Encoding layer

Inspired by the attention mechanism, research finds that most image processing neural network models are equally applicable to NLP and have great development potential. This paper introduces the SquashSparseTransformer mechanism and capsule model originally applied to images into natural language processing and transforms the original image processing mechanism into an algorithm suitable for natural language processing.

In this model, the SparseTransformer algorithm firstly applies the sparse self-attention used for image research to natural language processing. SquashsparseTransformer and K-max-flatten are developed based on sparse attention mechanism and maxpooling.

SquashSparseTransformers

SquashSparseTransformer is based on the SparseTransformer framework[34]. The difference lies in the realization of self attention[35]. Through top-k selection, attention is reduced to sparse

attention. In this way, the parts that are most helpful to attract attention are kept and other irrelevant information is deleted. This selective method is effective in preserving important information and eliminating noise. Attention can be focused more on the most contributing value factors. Next, we introduce the sparsity of self attention, and then extend it to context attention, as shown in figure2

In self attention, query vector $Q[l_Q, d]$, key vector $K[l_k, d]$ and value vector $V[l_v, d]$ are the linear transformations of the source context, that is, the input of each layer,

$Q = W_{Q_x}$, $K = W_{K_x}$, $V = W_{V_x}$ Sparse Transformer first generates the attention score P, as shown below:

$$P = \frac{QK^T}{\sqrt{d}} \quad (1)$$

Then, assuming that the larger the score, the higher the correlation, the model evaluates the value of the score P. The sparse attention masking operation function $M(p, k)$ is implemented on P to select top-k contribution elements. Specifically, we select the k largest elements in each row of P and record their positions in the position matrix $[i, j]$, where k is a hyper-parameter. Specifically, assuming that the k-th largest value of the i-th row is t_i , if the value of the j-th component is greater than t_i , record the position (i, j). We concatenate the thresholds of each row to form a vector $t = [t_1, t_2, t_3, \dots, t_{l_Q}]$. The masking function $M(p, k)$ is expressed as follows:

$$M(p, k)_{ij} = \begin{cases} P_{ij} & \text{if } P_{ij} \geq t_i(k - \text{thlargestvalueofrow } i) \\ -\infty & \text{if } P_{ij} < t_i(k - \text{thlargestvalueofrow } i) \end{cases} \quad (2)$$

In the top-k option, the k highest are selected explicitly. This is different from dropout where points are dropped randomly. Such an explicit choice can not only ensure the preservation of important components, but also simplify the model because k is usually a small value such as 5 or 10. The next step after top-k selection is normalization:

$$A = \text{softmax}(M(P, L)) \quad (3)$$

$$A = \text{squash}(A) = \frac{\|A\|^2 A}{0.5 + \|A\|^2 \|A\|} \quad (4)$$

Where A is the standardized score. Since the masking function $M(p, k)$ assigns scores smaller than the first k to negative infinity, the normalized score, that is, the probability is approximately zero. The output representation of self attention C can be calculated as follows:

$$C = AV \quad (5)$$

The output is the expected value after the sparse distribution A . According to the distribution of the selected components, the attention in the sparse transformer model can be more focused. In addition, this sparse attention can be extended to contextual attention. Similar to but different from the self-attention mechanism, Q is no longer a linear transformation of the source context, but a decoding state s . In the implementation, we replace Q with W_{Qs} , where W_Q is still a learnable matrix. In short, the attention of our proposed Sparse Transformer distracts attention weights. Attention can be focused on the most contributing factors, which is consistent with self-attention and contextual attention.

Capsules

Capsules[15] were also used to obtain image features at the beginning. To be applied into natural language processing, this model transforms image features into text features. A Dense structure with dynamic routers that can dynamically scan text windows is formed. The capsule model obtains the position of each feature and the association between the front and back features, as shown below Fig. 3. According to the input text feature space $S_1 = w_1, w_2$, the probabilities of $S_2 = v_1, v_2$ are derived respectively.

$$(p_{1|1}, p_{2|1}) = \frac{1}{Z_1} (e^{(w_1, v_1)}, e^{(w_2, v_1)}) \quad (6)$$

Feature w_1 alone is not enough, we need to calculate $(p_{1|1}, p_{2|1}), (p_{1|2}, p_{2|2})$. The probability distribution obtained by the feature i is $(p_{1|i}, p_{2|i})$, and this feature is cut into four parts[36], and then passed to v_1, v_2 and finally, v_1, v_2 are the accumulation of the incoming features of each bottom layer. The cumulative sum of each feature is obtained. And activate the function

$$v_1, v_2 = \sum_i^2 p_{1|i}, \sum_i^2 p_{2|i} \quad (7)$$

$$v_j = \text{squash}(\sum_i p_{j|i} u_i) \quad (8)$$

$$v_j = \text{squash}(\sum_i \frac{e(w_i, v_j)}{z_i} u_i) \quad (9)$$

Different from the activation function of the original capsule model, the algorithm in this paper adopts the calculation method of Squash. The feature of this function is that it can enlarge when

the modulus length is very close to 0, instead of compressing globally like the original function.

$$\text{squash}(x) = \frac{\|x\|^2 x}{0.5 + \|x\|^2 \|x\|} \quad (10)$$

Encoding keyword layer

In this layer, there are two other inputs, key phrase 1 and key phrase 2. These two key phrases describe the key phrases involved in the text, and the key phrases occupy a lot of weight. To solve the data preprocessing part in the process of data imbalance, we use synonym substitution, and other operations.

$$F1 = \text{Add}(\text{BiLSTM}(e_1), S_1) \quad (11)$$

$$F2 = \text{Add}(\text{BiLSTM}(e_2), S_2) \quad (12)$$

Based on RNN[37], LSTM[38] adds hidden state and cell control, that is, during the transmission process, there are two transmission states, one c_t (cell state) and one h_t (hidden state), and converted to a value between 0 and 1 through a sigmoid activation function, as a gated state, and combine forward and backward training LSTM to form BiLSTM[39]. The calculation formula of a time step is as follows:

$$z = \tanh(w \odot (x^t + h^{t-1})) \quad (13)$$

$$z^i = \sigma(w^i \odot (x^t + h^{t-1})) \quad (14)$$

$$z^f = \sigma(w^f \odot (x^t + h^{t-1})) \quad (15)$$

$$z^o = \sigma(w^o \odot (x^t + h^{t-1})) \quad (16)$$

$$h^t = z^o \odot \tanh(c^t) \quad (17)$$

$$y^t = \sigma(w^t h^t) \quad (18)$$

where σ is the matrix multiplication corresponding to the same shape, z^f, z^i, z^o is obtained by multiplying the splicing vector by the weight matrix. Each time step is divided into three stages, and the forgetting is extreme. By calculating z^f , the information is filtered in the previous state c^{t-1} , and the unimportant information is forgotten. x^t information is controlled to select memory. In the output stage, the state that can be output is controlled by z^o [29]. To deal with MASK, in BiLSTM, the Layer is rewritten into a class that can accept mask. That is, the forward and backward LSTM are made once and then spliced or added. The final output will contain 0 in the padding part (because the padding part is involved in the operation at the beginning)[18]. So there is no way to exclude it afterward, only exclude it beforehand

$$Output = reverse_sequence(LSTM(s, mask)) \quad (19)$$

then do a forward LSTM, and then reverse the result back. Note that only the non-padding part is reversed when inverting. In this way, it can be ensured that the padding part is always not involved in the recursive operation and that it is aligned with the result of the forward LSTM.

Modeling layer

Pooling is to simplify the weight composition and filter out the important weights. In Max-pooling and Average-pooling, the average or maximum value of the feature window is directly taken, ignoring other less important features. Therefore, in the sliding window process, the first k value is large each time.

$$x = top_select(w) \quad (20)$$

$$o = flatten(x) \quad (21)$$

Output layer

In this layer, we calculate the document representation O, and pass it to the relu layer and softmax layer.

$$O = Relu(o) \quad (22)$$

$$output = Softmax(O) \quad (23)$$

Results

The model experiment in this article is based on the wikipedia-pubmed-and-PMC-w2v vocabulary. One-hot processes the training set and test set, and uses generation to process the input data. The results of our model are micro-F1 scores of 0.802 and 0.8767 on chemprot and DDI, respectively.

Data

Chemprot[20]: Chemprot is a mission data set published on Bio Creative VI Track 5. Chemprot consists of 1820 people interacting with chemical-protein and using PubMed abstract Bio Creative VI text mining chemical-protein interactive sharing tasks. There are 5 types in total: CPR:3, CPR:4, CPR:5, CPR:6 and CPR:9. There are 4154 items in the training set and 2416 items in the test set.

DDI[19]: The 2013 DDI extraction corpus is a collection of 792 texts selected from the Drug Bank database and other 233 Medline abstracts in benchmark. There are a total of four categories, Effect, mechanism, advise, and int. The training set is 2397 and the test set is 1004. There are four categories in total: DDI-effect': 1212,'DDI-mechanism': 946,'DDI-advise': 633,'DDI-int': 146. The corpus distribution of DDI and ChemProt is shown in table 1

Data preprocessing

For processing data balance, we use word conversion (synonymous substitution, addition, and deletion of words, random insertion) to convert the text into text with different expressions. First, we introduce the NLTK [40] word segmentation toolkit to build phrase trees. Under normal circumstances, the algorithm can only split sentences into words based on spaces. However, in many cases, words and phrases have completely different meanings. For example, "Not at all." The content of the phrase has nothing to do with the word itself. Based on the NLTK word segmentation toolkit, we segment the text, delete the text stop words and extract the stem. This forms a new text that contains the core expression of each sentence in the original text. The statistical data distribution of the text in the DDI training set is shown in figure 4a and the statistical data distribution of the text of the Chemprot training set is shown in figure 4b. The way to solve the data imbalance is to replace synonymously, add or delete words, and insert randomly. Setting the parameter means that the label needs to increase the multiple of the synonymous text. For example, for DDI-advise, it needs to be doubled. The model in this paper uses the method in [31] for reference. In preprocessing, based on the NLTK library, four algorithms are used on the data set:

Synonym replacement (SR): Randomly select n non-stop words from the sentence. Replace each word with one of the synonyms selected at random.

Random Insertion (RI): Find random synonyms of random words that do not stop words in the sentence. Insert the synonym at random locations in the sentence. Do these n times.

Random Exchange (RS): Randomly select two words in a sentence and exchange their positions. Do these n times.

Random deletion (RD): For each word in the sentence, delete it randomly with probability p.

According to different data sets, a certain priority is given to the four algorithms. For example, in the Chemprot data set, the ratio of CPR: 9 categories to CPR:4 categories is approximately 3:1.

At this time, the priority of the algorithm can be set to RI, RD, RS, and SR.

Evaluation

The micro F1-Score is used as the evaluation index of the model. F1-Score calculates the accuracy, recall and F1 score for each category. Micro will calculate TP, FP, FN, TN for each category, and add them to form a new TP, FP, FN, TN, and then can calculate Precision and Recall to obtain micro-F1.

$$precision_i = \frac{true_positive}{true_positive + false_positive} \quad (24)$$

$$recall_i = \frac{true_positive}{true_positive + false_negative} \quad (25)$$

$$Percision = \sum_i^n percision \quad (26)$$

$$Recall = \sum_i^n recall \quad (27)$$

$$F1_Score = 2 \cdot \frac{Percision * Recall}{Percision + Recall} \quad (28)$$

Experimental comparison

It can be found from the comparative experiment that the effect of Bert is very good, and the effect of some Bert-based variants such as Biobert[1] is also very good. Common LSTM and CNN models such as PM-BLSTM [41], CNN[42] etc. are better than those of Bert and Bert. The effect of the variant is much worse. But Bert training requires a very strong equipment base, and it is difficult to train Bert[18] when the infrastructure is insufficient. From this, we can infer that reasonable use of the self-attention mechanism can get good results. The experiment uses micro F1 as the evaluation standard. The model comparison is shown in table 2.

Ablation experiment

In order to verify the function of each module of the model, we perform an ablation experiment, as shown in table 3

The Influence of capsule

Experiments of ChemProt and DDI are to eliminate the capsule algorithm from the entire experiment. It is found that although the F1 of the experiment has been reduced, the change is

not big. This is because, in the application of natural language processing, the application of the capsule model is like a can. The dense layer of multiple iterations. But in the model, the capsule plays a role in supplementing the sparse self-attention loss position information.

The Influence of SquashSparseTransformers

The sparse self-attention algorithm is eliminated from the entire experiment, and it is found that the F1 score of the experiment is reduced, and the reduction is 0.2 more than that of the capsule. It can be found that SquashSparseTransformers plays a core feature extraction role in the model.

The Influence of BiLSTM

If BiLSTM is eliminated, F1 score is greatly reduced. This is because the keywords processed by BiLSTM are the key relationship feature information in the text. In the text information, the keywords occupy the largest weight.

The Influence of K-max-Flatten

The K-max-Flatten algorithm is eliminated from the entire experiment. After SquashSparseTransformers, capsules, and BiLSTM are optimized, the salient features of the text can be filtered out, but not all salient features can play a decisive role in the classifier. So without K-max-flatten, F1 score drops drastically, directly down to 0.742. Although the experimental F1 has been reduced, it has not changed much. But for DDI corpus, the result is only 0.006 less than the complete model. This is related to the size of the data set. Large data sets contain more vocabulary and have a stronger ability to fit simple models. Therefore, the effect of increasing or reducing feature calculations is not significant.

Experimental parameter settings

The experimental parameter settings are as table 4

Discussion

At present, the research on relation extraction of biomedical texts mainly focuses on the learning of neural networks. In 2018, with Bert research appeared, NLP research set off a wave of research on Bert. But Bert has extremely high requirements for experimental equipment, which limits the research of some researchers. So we learn Bert's core ideas, and borrow part of Bert's research, and apply it to the study of relationship extraction of biomedical texts, which can achieve better prediction results with limited resources.

Conclusion

In the experiment of this article, when the hardware conditions required by Bert are too great, we can use a circuitous route. We try to combine Bert's core ideas with other models to obtain more possibilities with lower experimental equipment requirements. The improved version of the

attention mechanism, SquashSparseTransformers, implements the functions of the original Transformers, reduces the computational time complexity through the Top-select filtering mechanism, and obtains influential feature values through the modular function Squash, thereby obtaining higher feature impact. We use capsules to obtain the location information of the text and retain certain functions filtered by SquashSparseTransformers to prevent important functions that are ignored by SquashSparseTransformers. Finally, we supplement the keyword information through BiLSTM and select top-3 features as the feature representation of the text. Experiments have found that on the two multi-classification data sets of DDI and Chemprot, this circuitous path is feasible and is better than the Bert model.

Abbreviations

SSACBKM :SquashSparseTransformer, Capsules, BiLSTM, K-max-Flatten,
SVM:Support vector machine
RNN:Recurrent Netural Networks
CNN: Convolution neural networks
BiLSTM:Bidirectional long stort term memory
LSTM:long stort term memory
DDI:Drug Drug Interaction
NLP:Natural language processing
CRFs:Conditional Random Fields
PPI:Protein protein interactions
ChemProt:Chemical protein
CBOW:Continuous bag of words
NLTK:Netural language Toolkit
TP:True Positive
FP:False Positive
FN:False Negative
TN:True Negative

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Availability of data and materials

The datasets used and analysed during the current study are available in the BioCreative VI (Resources) and HULAT workshop
https://biocreative.bioinformatics.udel.edu/media/store/files/2017/ChemProt_Corpus.zip
<https://github.com/isequra/DDICorpus>.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Natural Science Foundations of China under Grant 61463050.

Author's contributions

Xiaobing Zhou is responsible for theoretical guidance, and Huiping is responsible for specific modeling, experiments and writing manuscript.

Acknowledgements

Special thanks to Professor Zhou for his guidance in the theory and experimental process. Special thanks to the BioCreative VI organizers and HULAT workshop organizers for providing a communication platform for this study.

References

1. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
2. GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*, pages 427–434, 2005.
3. Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *Coling 1992 volume 2: The 15th international conference on computational linguistics*, 1992.
4. Yudong Liu, Zhongmin Shi, and Anoop Sarkar. Exploiting rich syntactic information for relationship extraction from biomedical articles. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 97–100, 2007.
5. Angus Roberts, Robert Gaizauskas, and Mark Hepple. Extracting clinical relationships from patient narratives. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 10–18, 2008.
6. Wei Zheng, Hongfei Lin, Ling Luo, Zhehuan Zhao, Zhengguang Li, Yijia Zhang, Zhihao Yang, and Jian Wang. An attention-based effective neural model for drug-drug interactions extraction. *BMC bioinformatics*, 18(1):445, 2017.
7. One Model To Rule Them All. Semi-supervised target-level sentiment analysis via variational autoencoder. 1810.
8. Sepp Hochreiter and Jürgen Schmidhuber. Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479, 1997.
9. ZHOU Ming and WU Jin HUANG Chang-Ning. A fast learning algorithm for part of speech tagging: An improvment on brill's transformation-based algorithm [j]. *CHINESE JOURNAL OF COMPUTERS*, 4, 1998.
10. Zhehuan Zhao, Zhihao Yang, Hongfei Lin, Jian Wang, and Song Gao. A protein-protein interaction extraction approach based on deep neural network. *International Journal of Data Mining and Bioinformatics*, 15(2):145–164, 2016.
11. Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, 2016, 2016.
12. Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, Shaowu Zhang, Yuanyuan Sun, and Liang Yang. A hybrid model based on neural networks for biomedical relation extraction. *Journal of biomedical informatics*, 81:83–92, 2018.
13. Diana Sousa, Andre Lamurias, and Francisco M Couto. Using neural networks for relation extraction from biomedical literature. *arXiv preprint arXiv:1905.11391*, 2019.
14. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
15. Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017.
16. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
17. Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.
18. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
19. María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920, 2013.
20. Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaiky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wieggers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.
21. Andre Lamurias, Diana Sousa, Luka A Clarke, and Francisco M Couto. Bo-lstm: classifying relations via long short-term memory networks along biomedical ontologies. *BMC bioinformatics*, 20(1):1–12, 2019.
22. SS Deepika, M Saranya, and TV Geetha. Cross-corpus training with cnn to classify imbalanced biomedical relation data. In *International Conference on Applications of Natural Language to Information Systems*, pages 170–181. Springer, 2019.
23. Dinh Phuong Nguyen and Tu Bao Ho. Drug-drug interaction extraction from biomedical texts via relation bert. In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–7. IEEE, 2020.
24. Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.
25. Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgård, Francisco S Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, et al. Chemprot: a disease chemical biology database. *Nucleic acids research*, 39(suppl.1):D367–D372, 2010.

26. Rezarta Islamaj Dogan, Andrew Chatr-aryamontri, Sun Kim, Chih-Hsuan Wei, Yifan Peng, Donald C Comeau, and Zhiyong Lu. Biocreative vi precision medicine track: creating a training corpus for mining protein-protein interactions affected by mutations. In *BioNLP 2017*, pages 171–175, 2017.
27. Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. Attention guided capsule networks for chemical-protein interaction extraction. *Journal of Biomedical Informatics*, 103:103392, 2020.
28. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
29. Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003.
30. Show-Jane Yen and Yue-Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, 2009.
31. Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.
32. Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.
33. Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.
34. Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.
35. James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*, 2020.
36. Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *International conference on learning representations*, 2018.
37. Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.
38. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
39. Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.
40. Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.
41. Deyu Zhou, Lei Miao, and Yulan He. Position-aware deep multi-task learning for drug–drug interaction extraction. *Artificial intelligence in medicine*, 87:1–8, 2018.
42. Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
43. Sangrak Lim, Kyubum Lee, and Jaewoo Kang. Drug drug interaction extraction from the literature using a recursive neural network. *PloS one*, 13(1):e0190926, 2018.
44. Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. A pre-training technique to localize medical bert and enhance biobert. *arXiv preprint arXiv:2005.07202*, 2020.
45. Cong Sun, Zhihao Yang, Ling Luo, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. A deep learning approach with deep contextualized word representations for chemical–protein interaction extraction from biomedical literature. *IEEE Access*, 7:151034–151046, 2019.
46. Sijia Liu, Feichen Shen, Yanshan Wang, Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli, Vipin Chaudhary, and Hongfang Liu. Attention-based neural networks for chemical protein relation extraction. *Training*, 1020(25.247):4157, 2017.
47. Cong Sun, Zhihao Yang, Leilei Su, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. Chemical-protein interaction extraction via gaussian probability distribution and external biomedical knowledge. *arXiv preprint arXiv:1911.09487*, 2019.

Table 1: Data distribution

DDI	effect	mechanism	advise	int
	1212	946	633	146
ChemProt	CPR:4	CPR:3	CPR:9	CPR:5
	3345	1318	1184	289

Table 2: Experimental comparison

Model	DDI	Chemprot
PM-BLSTM [41]	0.7299	-
CNN[42]	0.6975	-
treeLSTM(mi)[43]	0.826	-
Bio-bert[1]	0.799	0.7702
ouBioBERT [44]	0.811	0.75
Blue-Bert(multi-task) [24]	0.799	0.744
DS-LSTM[45]	-	0.6944
Att-GRU[46]	-	0.506
Bert-Att-Capsule[27]	-	0.7470
GaussianProbabilityDistributionOperation(Bert)[47]	-	0.7656
Ourmodel	0.8767	0.802

Table 3: Ablation experiment

Model	DDI	Chemprot
Ourmodel	0.876	0.804
Ourmodel(no Capsule)	0.866	0.792
Ourmodel(no SquashSparseTranformer)	0.834	0.784
Ourmodel(no BiLSTM)	0.846	0.757
Ourmodel(no Kmaxflatten)	0.870	0.742

Table 4: Experimental parameter settings

	lr	0.001
	batch_size	32
sparse self-attention	Epoch	5
	Head	10
	Head-size	20
	rate	2
Capsule	num_capsule	Sent_lenth
	dim_capsule	200
	routing	3
Dropout	rate	0.25
BiLSTM	unit	50
	activation	relu
K_Max_Faltten	k	3
Dense1	unit	128 or 64
	regularizer	L2=0.003
	activation	relu
Dense1	unit	Num_label
	activation	softmax

Figures

Figure 1

The SSCBKM model.

Figure 2

Squash Sparse Transformer

Figure 3

Capsules Mapping method

(a) DDI Data set label distribution

(b) ChemProt Data set label distribution

Figure 4

(a) DDI Data set label distribution (b) ChemProt Data set label distribution