Research of ddi based on multi-label conditional random field

Yangzhi Yu [1], Hongtao Deng [1,a] and Xun Zhu [1]

[1] *Mathematics and Computer Science Department of JiangHan University, China*

Abstract. The detection of drug name and drug-drug interaction(DDI) is considered as a sequence labeling task in this paper. We present the multi-label CRF method to complete it. Compared to the traditional method, our method can not only identify drug names, but also can identify drug-drug interaction. According to the characteristics of medical texts, this paper extracts the good features of the description of DDI. The proposed method has good performance in DDIExtraction 2013 evaluation corpus.

1 Introduction

The detection of DDI is an important research task in ensuring patient safety since these DDI can be so dangerous and increase health care costs. The phenomenon of drug combination is becoming more and more common. It can enhance the efficacy and improve the therapeutic effect. However, it also may result in increased frequency of adverse drug reactions and increased severity of adverse reactions. The DDI has become a vital factor affecting the drugs rational use. More and more researchers pay their attention to the DDI extraction research.

To address the extraction of DDI from biomedical texts, two main steps were proposed: the detection and classification of pharmacological substances and the detection and classification of drug–drug interactions.

There is a large number of entity relationship extraction researches, such as protein extraction, RNA molecules relationship extraction, DDI extraction and the drug adverse effects event extraction and so on. Hasegawa T[1] used the method based on co-occurrence to extract entity relationships. They argue that all entities that co-occur in a text fragment should have a relationship. Deng B[2] used the method based on semantic pattern to identify entity relationships in Chinese. They extract the rules by hand, and then match with the actual situation to determine whether the two entities are related. Now the mainstream method is based on pattern matching[3], which requires less manual intervention.

2 Task description

DrugBank[4] is a constantly updated database of saving chemical informatics, bioinformatics resources and so on, which holds a large number of drugs and drug targets.

[a] Hongtao Deng: hongtaodeng@qq.com

2.1 Drug entity recognition

Name entity recognition estimates whether a character sequence represents a name entity, and determines its types, namely to find name entities and tagging name entity. Drug entity recognition refers to the identification of name entities in the text in the description of pharmacological substances. The task includes the determination of the boundary of the entity name and the determination of the entity type. The types of drug entities are classified into four types: Drug, Brand, Group, and No-Human. The following is a brief introduction of the four types of entities:

1) Drug type: This type refers to all have been approved for treatment, prevention and diagnosis of chemicals or drugs.

2) Brand Type: This type refers to original drug firstly developed by a pharmaceutical company.

3) Group Type: This type represents a class of drug having a certain function.

4) No-Human Type: This type represents the drug not approved for using as a clinical medicine, and has an effect on the body's organs.

2.2 Drug-drug interaction detection

The ultimate goal of the task is to extract DDI, which refers to the relationship between the two drug entities within a sentence. The types of DDI are divided into four types: Advice, Effect, Mechanism, and Int.

1) Advise: When two drug entities are suggested to use together in the text, the relationship between the two drugs is defined as Advise type.

2) Effect: Mechanisms of drug action is generally divided into pharmacodynamic mechanism and pharmacokinetic mechanism. Pharmacodynamics mechanism refers to the effect of a drug affected by another drug. When the article mentioned the existence of such effects of the two drugs, it is defined as Effect type.

3) Mechanism: Pharmacokinetics mechanism refers to that the absorption, diffusion, metabolism and excretion of drugs are affected. When there is pharmacokinetics mechanism effect between the two drugs, this relationship is called the Mechanism.

4) Int: When the text only mentioned the existence of a relationship between the two drugs, but did not make a specific description of the relationship, unable to determine the type of relationship, it is defined as Int.

3 CRF method

3.1 Introduction of CRF algorithm

CRF is a kind of algorithm based on probabilistic graph model, which is used to label and segment sequence data. Lafferty, J., McCallum and Pereira[5] proposed CRF model in 2001. CRF is a discriminative undirected probabilistic graphical model. It is often used for labeling or parsing sequential data, such as natural language text, biological sequences[6] and computer vision[7]. CRF acquires good performances in shallow parsing[8], named entity recognition[9] and gene finding, among other tasks, being an alternative to the related hidden Markov models. We use CRF to solve drug NER as sequence annotation problem.

Lafferty[5] define a CRF on observations X and random variables Y as follows: Let G=(V,E) be a graph such that $Y=(Y_V)$, $_V \in V$, so that Y is indexed by the vertex of G. Then (X, Y) is a conditional random field when the random variables Y_V, conditioned on X, obey the Markov property with respect to the graph:

$p(Y_V \mid X, Y_W, _W \neq _V) = p(Y_V \mid X, Y_W, _W \sim _V)$.

Where W~ V means that w and v are neighbours in G.

3.2 CRF drug entity recognition

We need to apply CRF to recognize drug entity first, then apply the multi-label CRF to extract DDI. CRF is often used in named entity recognition. And it has been successfully applied in the field of biomedical RNA, DNA, protein and other entities identifications. There are four types of drug entities that need to be identified in this task, which not only identify the boundary of the entity, but also identify the type of the entity.

3.2.1 Encoding mode

The BEISO label is used to encode the sentences in the paper. Label B represents the beginning word of drug entity, label E stands for the ending word of drug entity, label I represents the middle word of drug entity, S represents a single-word entity, label O stands for the word of non-entity. B, E, I, S, each label has four types respectively (for example B_drug, B_brand, B_group, B_drug_n), corresponding to the four types of drug entity (drug, brand, group, drug_n) .

3.2.2 Feature design

(1) POS feature: getting POS information by Stanford tool, which uses unigram, bigram, trigram feature.

(2) Ignore cases feature：using unigram, bigram, trigram feature.

(3) Word feature: 'A' replacement for uppercase letters, 'a' replacement lowercase letters, '0' to replacement for numbers;

(4) Words Brown cluster prefix：using unigram, bigram, trigram feature.

3.3 DDI based on multi-label CRF

3.3.1 Multi-label CRF algorithm

The traditional method is only applied for CRF sequence labeling，which can be successfully used in the NER task. However its drawback is that it can only identify the continuous entity. In order to avoid this shortcoming, the multi-label CRF is proposed, which adopt 6 classes labels to mark the sentence. They are "B_val", "I_val", "O", "E_val", "S_val" and "number" labels. Since the multi-label method has the ability to mark one symbol with more labels, which can view non-adjacent tokens as an entity. It has achieved good results in the recognition of discrete entities mission[10]. And then we made some changes to the algorithm so as to identify the entity relationship. One is to package multiple consecutive drug entities. Another is to improve the label coding. In this paper, we removed the original I label, which represents the internal word of the entity. Because of identifying various types, label B and E is divided into four types, corresponding to the four types of relationship.

Label coding algorithm. We adopt the multi-label to mark two entities which have relations with label B, E. If an entity has more than one label, it is required to add the label to the overlay algorithm. According to the characteristics of the task, we removed the I type labels. Multi-labels of entity relationship recognition contain five types of labels: "B_val", "E_val" and "O", "S_val", "number". The B type label represents the previous entity of DDI, and the E type label represents the latter entity of DDI. O type label represents a type of non drug substance. S type label represents drug entity of no DDI. Binary values are represented by the number and val.

Label decoding algorithm. The output of CRF method is just a label sequence, which we need to decode the relationship we need. Decoding process is divided into two steps. First the coordinate drug package is separated to several drugs, and then the drugs with interacted relation should be combined as a pairs.

3.3.2 Packing algorithm

The phenomenon is very common that different drug entity have interaction with the same entity in a sentence. For example,"Benzthiazide may interact with alcohol, bloodthinners, decongestantdrugs (allergy, cold, and sinus medicines), diabetic drugs, lithium, norepinephrine, NSAIDs like Aleve or Ibuprofen, and high blood pressure medications." In the sentence, Benzthiazide has interaction with other entities. If we match the two drugs with relation directly, the label value will be too large. The sparsity of large value label will lead to the reduction of identification effect. The multi-label CRF algorithm limits that the number of interaction entity pairs is less than six in one sentence, so we should pack parallel drug as a drug entity by some rules we specified. The parallel entities are generally connected by a high frequency words, such as 'comma', 'and', 'or', 'such as', and 'like' so on. We first pick up 30 words which had higher frequency before and after the drug from the training corpus. If the frequency word is between two drug entities, they will be packaged together and represented by the first drug. If an entity's type is determined in decoding process, all coordinate entities with it should have the same type of relationship.

3.3.3 Feature design

According to the characteristics of biomedical text, we selected a number of targeted features. There are POS feature, word feature ignored case, dependency feature and two entities' distance feature, as shown in Table1.

Table 1. Multi-label CRF training data example.

word	POS	Entity label	Dependcy word	Dependcy word location	Entity location	Output label
concurrent	JJ	N	therapy	2	0	O
therapy	NN	N	recommended	9	0	O
with	IN	N	ORENCIA	4	0	O
orencia	NN	brand	therapy	2	1	B_{advise} 1
and	CC	N	ORENCIA	4	0	O
antagonists	NNS	group	ORENCIA	4	2	E_{advise} 1
is	VBZ	N	recommended	9	0	O
not	RB	N	recommended	9	0	O
recommended	VBN	N	null	0	0	O
.	.	N	recommended	9	0	O

4 Experiment

DrugBank corpus is divided into training data and testing data. The training data contains 572 articles, a total of 5675 sentences. The testing data contains 158 articles, a total of 973 sentences. Statistical results of drug entity and DDI are shown in Table 2.

Table 2. Quantity of drug substance and drug relation in the corpus.

	Training	Testing
documents	572	158
sentences	5675	973
Drug	8197	1518
Group	3206	626
Brand	1423	347
Drug_n	103	21
Mechanism	1260	279
Effect	1548	301
Advice	819	215
Int	178	94

We analyze the corpus.Table3 scribes the quantity of each type drug. As shown, the majority are Drug and Group, and there are 82 entities of Drug_n type. The length of the entity token of Drug and Group type are less than 3. Almost all type entity's name are made up of single token.

Table 3. The number of each type of entity name.

Drug	Group	Brand	Drug_n
1583	1403	488	82

We mainly utility the Tokenizer, Part-Of-Speech Tagger[11], Dependency Parser[12], Constituency Parser[13] in the feature designing. The steps of multi-label CRF to classifying the drug and DDI are file parsing, pre-process, entity coding, entity packing, relational coding, decoding, and analyzing results.

Table4 is the recognition result. It is considered that the entity is correctly identified, only when the entity boundary and the entity type are identified correctly.

Table 4. The result of recognition.

	Accuracy	Recall	F value
Drug	0.857	0.888	0.872
Brand	0.836	0.811	0.823
Group	0.829	0.774	0.801
Drug_n	0.764	0.567	0.651
Total	0.840	0.823	0.831
Advise	0.483	0.421	0.449
Effect	0.468	0.444	0.456
Mechanism	0.394	0.386	0.390
Int	0.351	0.364	0.357
Total	0.461	0.411	0.439

5 Conclusion and future work

It is significant of NER and RE task from huge medical text. In this paper, we mainly research the extraction of the interaction between two drugs utilizing multi-label CRF model in medical text. The multi-label CRF model realized the extraction of non-adjacent named entity. We also extracted a variety of features to improve the model efficiency from the medical text. Our task may provide a reference for future work.

Acknowledgment

We thank all reviewers for the insightful comments. This work is supported by the Graduate Innovation Foundation, key disciplines "Management Science and Engineering" of Jianghan University, Hubei Province, China.

References

1. Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 415.
2. Deng B, Fan X, Yang L. Entity relation extraction method using semantic pattern[J]. Jisuanji Gongcheng/ Computer Engineering, 2007, 33(10): 212-214.
3. Tikk D, Thomas P, Palaga P, et al. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature[J]. PLoS Comput Biol, 2010, 6(7): e1000837.

4. Wishart D S, Knox C, Guo A C, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration[J]. Nucleic acids research, 2006, 34(suppl 1): D668-D672.
5. Lafferty, J., McCallum, A., Pereira, F. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann. pp. 282–289.
6. He, X.; Zemel, R.S.; Carreira-Perpinñán, M.A. (2004). "Multiscale conditional random fields for image labeling". IEEE Computer Society. CiteSeerX:10.1.1.3.7826
7. Sha, F., Pereira, F. (2003). "shallow parsing with conditional random fields"
8. Settles, B. (2004). "Biomedical named entity recognition using conditional random fields and rich feature sets". Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. pp. 104–107.
9. Tsuruoka Y, Tsujii J. Boosting precision and recall of dictionary-based protein name recognition[C]//Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13. Association for Computational Linguistics, 2003: 41-48.
10. Lin Wutao. Research of medical information extraction based on multi label 2015[D]. CRF Wuhan: Wuhan University, 2015
11. Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Si nger. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.
12. Chen D, Manning C D. A Fast and Accurate Dependency Parser using Neural Networks[C]//EMNLP. 2014: 740-750.
13. Zhu M, Zhang Y, Chen W, et al. Fast and Accurate Shift-Reduce Constituent Parsing[C]//ACL (1). 2013: 434-443.