Extracting and Visualizing Semantic Relationships from Chinese Biomedical Text

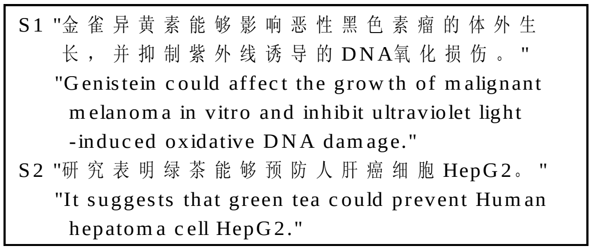
Qingliang Miao, Shu Zhang, Bo Zhang, Yao Meng, Hao Yu
Fujitsu R&D Center Co., Ltd.
{qingliang.miao, zhangshu, zhangbo, mengyao, yu}@cn.fujitsu.com

Abstract

In this paper, we study how to automatically extract and visualize food (or nutrition) and disease relationships from Chinese publications of Nutritional Genomics. Different from previous approaches that mostly apply handcrafted rules or co-occurrence patterns, we propose an approach using probabilistic models and domain knowledge. In particular, we first utilize encyclopedia to construct a domain knowledge base, and then develop a sentence simplification model to simplify complicated sentences we meet. Afterwards, we treat relation extraction issue as a sequence labeling task and adopt Conditional Random Fields (CRFs) models to extract food and disease relationships. Finally, these relationships are visualized. Experimental results on real-world datasets show that the proposed approach is effective.

1 Introduction

Advancements in biomedical science has led to large volume of published research articles, especially in Nutritional Genomics, an emerging interdisciplinary that studies the relationship between human genome, food and diseases (Hakenberg *et al.*, 2010; Sharma *et al.*, 2010; Tsuruoka *et al.*, 2011). For example, many researches in Nutritional Genomics study the relationships between "green tea", "soy", "fish oil" and "tumor diseases". Mining and drawing a full picture of these relationships can be adopted in many practical fields, such as public health services, drug discovery, etc. However, due to the considerable number of unstructured data, it is unrealistic to go through and obtain the panoramas of relationships manually. Consequently, automatically relation extraction and visualization techniques become ever more important and necessary. Some prior work has studied how to extract food and disease relationships from English biomedical text (Yang *et al.*, 2011). On Chinese biomedical text, however, there is relatively little investigation conducted on food and disease relation mining. In this paper, we focus on extracting and visualizing food and disease relationship from Chinese biomedical text.



S1 "金雀异黄素能够影响恶性黑色素瘤的体外生长，并抑制紫外线诱导的DNA氧化损伤。"
"Genistein could affect the growth of malignant melanoma in vitro and inhibit ultraviolet light -induced oxidative DNA damage."
S2 "研究表明绿茶能够预防人肝癌细胞HepG2。"
"It suggests that green tea could prevent Human hepatoma cell HepG2."

Figure 1: Example of relation-bearing sentences in Chinese and their English translation.

Figure 1 shows two examples of Chinese biomedical sentences and their English translation. The objective of semantic relationship mining is to extract all the binary semantic relationships between food and diseases, such as <金雀异黄素, 影响, 黑色素瘤> (<genistein, affect malignant melanoma>), <绿茶, 预防, 人肝癌细胞 HepG2> (<green tea, prevent, human hepatoma cell HepG2>).

In order to facilitate the explanation, we first introduce two basic terminologies of relation-bearing sentences.

Definition 1: Multiple Relation-bearing Sentence

Multiple relation-bearing sentence (*MRS*) contains more than two entities and mutual relationships.

26 th Pacific Asia Conference on Language, Information and Computation pages 99-107

Take Sentence 1 for example, there is one food entity—genistein, and two disease entities—malignant melanoma and DNA damage, and two relationships. Generally speaking, MRS could be represented by the following patterns, where *M-M*, *O-M* and *M-O* respectively represent many-to-many, one-to-many and many-to-one relationships. Table 1 below shows the multiple relation patterns, where *e* represents entity, *r* represents relation words/phrase.

Pattern	Multiple relation patterns
M-M	$\{e_1, e_2, ..., e_m, r, e_1^{'}, e_2^{'}, ..., e_n^{'}\}$ $\{e_1, e_2, ..., e_m, (r_1), e_1^{'}, (r_2), e_2^{'}, ..., (r_n), e_n^{'}\}$
O-M	$\{e, r, e_1^{'}, e_2^{'}, ..., e_n^{'}\}$ $\{e, (r_1), e_1^{'}, (r_2), e_2^{'}, ..., (r_n), e_n^{'}\}$
M-O	$\{e_1, e_2, ..., e_m, r, e^{'}\}$

Table 1: Multiple relation patterns.

Definition 2: Single Relation-bearing Sentence

Single relation-bearing sentence (SRS) contains two entities and one relationship. Take Sentence 2 for example, we can see there are two entities (one food entity and one disease entity) and one relationship.

Mining semantic relationships from Chinese biomedical text is very challenging, because the sentence structure is complicated and most of the sentences contain multiple relationships. According to our statistic analysis of 3000 sentences from Chinese biomedical text, about 66% of the sentences are multiple relation-bearing sentences. Worse still, fewer biomedical resources such as USDA food database[1] and UMLS Meta thesaurus[2] are available in Chinese. Due to the complicated structure of multiple relation-bearing sentences, traditional methods could not perform effectively to extract food and disease relationships. Consequently, we have to simplify them, and then adopt extraction models to obtain food and disease relationships.

The remainder of the paper is organized as follows. In the following section we review the existing literature on semantic relation extraction. Then, we introduce the proposed approach in section 3. We conduct comparative experiments and present the results in section 4. At last, we conclude the paper with a summary of our work and give our future working directions.

2 Related Work

In the field of semantic relation mining, there are three dominant methods, namely, rule-based, pattern-based and learning-based methods (Finkelstein-Landau, M. and E. Mori, 1999; Bach and Badaskar, 2007; Weikum and Theobald, 2010; Zweigenbaum *et al.*, 2007). Next we will introduce these methods respectively.

Rule-based methods utilize predefined rules to extract relationships based on part of speech information (Weikum and Theobald, 2010). For example, if we want *isInstanceOf* relation, we can design extraction rules like <NP$_0$ such as {NP$_1$, NP$_2$,…NP$_n$}>. Some more sophisticated methods exploit syntactic information. For example, Fundel et al., first used a lexicalized parser to generate the dependency trees of each sentence, and then adopted four extraction rules to find protein and gene interactions (Fundel *et al.*, 2007). Rinaldi *et al.*, (2007) also utilized dependency parsing and lexicon to extract protein and gene relationships. However, rule-based methods mainly rely on handcraft rules, and suffer from low recall due to the sparseness of extraction rules. In addition, rule-based methods that incorporate syntactic information can be computationally costly in larger corpus.

Due to the sparseness issue in handcraft rules, pattern-based methods aim to construct comprehensive rules automatically (Hearst, 1992). Specifically, they are based on the duality of relationships, and usually adopt bootstrapping paradigm. For example, Brin (1998) proposed a pattern-based relation extraction system named DIPRE, which starts with a small set of seed facts for one or more relations of interest. Then it automatically looks for linguistic patterns in underlying sources as indicators of facts. Finally it utilizes these patterns to identify new fact candidates as further hypotheses to populate relationships. Agichtein and Gravano (2000) proposed a system called Snowball, which adopts similar strategy with DIPRE. However, Snowball does not use exact match, but a similarity function to group similar patterns instead. Snowball's

[1] http://ndb.nal.usda.gov/ndb/foods/list

[2] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

flexible matching system allows for slight variations in token or punctuation. In pattern-based methods, the initial patterns may shift during iterative processes, consequently it is inevitable to bring in noise. Girju and Moldovan (2002) extract lexico-syntactic patterns that refer to the causal relation.

Machine learning-based methods such as SVM and CRFs (Bundschus *et al.*, 2008; Lafferty *et al.*, 2001) can also be used in relationship extraction. Some work views relation extraction as classification issue, and adopt kernel features to train extraction models (Bunescu and Mooney, 2005; Zelenko *et al.*, 2003). Others treat relation extraction as a sequence labeling issue, and adopt HMM or CRFs to extract relationships. Bundschus et al., (2008) adopted CRFs model to extract treatment and disease relationships. However, effective learning features of these supervised approaches are derived from syntax parsers. Unfortunately, due to the complicated structure of biomedical sentences, few parsers perform well in Chinese biomedical sentences. When the sentence structure is complicated or the sentence contains multiple relationships, traditional methods cannot perform well (Jonnalagadda *et al.*, 2009).

3 The Proposed Approach

In this section, we will first introduce the architecture of the mining system, and then illustrate how to build domain knowledge base. After that, sentence simplification model will be introduced. In the end, we will explain how to utilize CRFs model to extract food and disease relationship on the basis of sentence simplification.

3.1 System Architecture

Figure 2 shows the architecture of the mining system. The inputs are unstructured biomedical texts, and the outputs are food and disease relationships. The system consists of four modules: (1) biomedical data server (BDS); (2) knowledge mining engine (KME); (3) relationship mining engine (RME); and (4) relationship visualization engine (RVE).

BDS collects biomedical texts by crawling scientific literature website such as *wanfang.com.* Then, web pages are cleaned to remove HTML tags, after that, abstracts in biomedical articles are extracted and splitted into sentences according to punctuations. Finally, word segmentation and part of speech tagging are conducted.

KME utilizes encyclopedia and biomedical corpus to construct knowledge base. Firstly, KME extracts food and disease entities from encyclopedia. Treating food and disease entities as anchor, KME adopts association rules to discover relation words from biomedical corpus. Finally, KME combines entities with relation words to construct domain knowledge base.

RME is the key part of the system, which includes three steps. Firstly, RME utilizes CRFs models and domain knowledge to extract food and disease entities. Secondly, it uses food and disease entities as anchors to simplify multiple relation-bearing sentences. Finally, CRFs models equipped with domain knowledge and other learning features are trained to extract relation words from simplified biomedical sentences.

RVE visualizes food and disease relationships. Figure 3 illustrates the visualization results of green tea and tumor disease relationships.

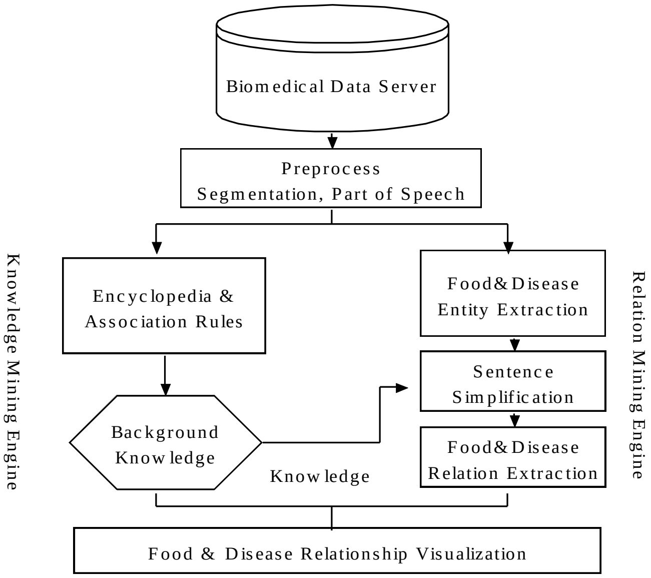
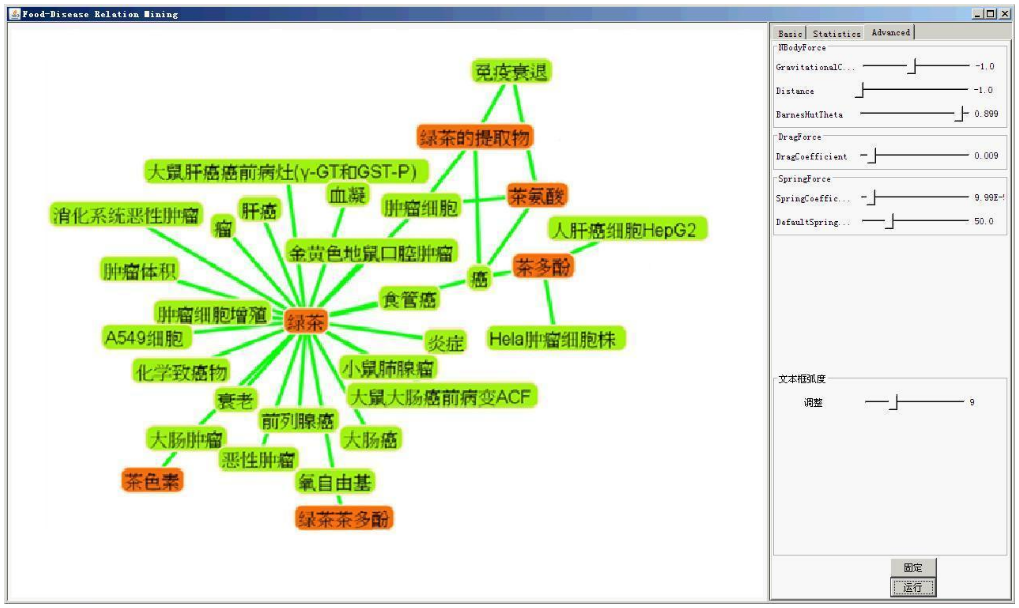


Figure 2: Flowchart of the proposed approach.



NBodyForce
GravitationalC... -1.0
Distance -1.0
BarnesHutTheta 0.899

DragForce
DragCoefficient 0.009

SpringForce
SpringCoeffic... 9.99E-!
DefaultSpring... 50.0

文本框弧度
调整 9

Figure 3: Food and disease relationship visualization results, red nodes represent green tea and its extractions, while green ones represent tumor disease entities.

3.2 Knowledge Base Construction

To construct a knowledge base, we need to extract food and disease entities and relation words. In particular, we first extract food and disease entities from three original data sources: *Wikipedia* Chinese version, *Baidu Baike*, and *Hudong Baike*. In these encyclopedias, concepts belonging to the same class are organized together. Therefore, we select 11 related categories such as "健康饮食 (healthy food)", "营养学(nutrition)" and "疾病 (disease)". After that, we collect food and disease entities from these 11 categories and assign each entity a Uniform Resource Identifier (URI). The URI is defined according to the following schema "*kb/category/entityName*". In the schema, field "*category*" is used to alleviate homonyms issues. For example, in our knowledge base, the URI of "apple" is defined as "*kb/fruit/apple*" instead of "*kb/company/apple*".

Through analyzing the content of each page, we extract 5 types of contents to construct domain knowledge, "*Title*", "*Alias*", "*Category*", "*Redirect*", "*Related Term*". Besides the above 5 types of contents, we also extract "*Function*" and "*Primary Constituent*" for food entities. We use Dublin Core (DC) metadata and Simple Knowledge Organization System (SKOS) to manage these contents. We will explain them in details as follows:

Title:

The titles in Hudong Baike are used as labels for the corresponding food and disease entities directly. Field "*entityName*" in URI is the same as title, which is represented by *dc:title*.

Alias:

In Wikipedia, editors may use alias to represent the same entity. For example, [[樱| 樱桃]] ([[cherry| prunus]]) will produce a link to \樱桃 while the displayed anchor is \樱. We call the displayed anchors as the aliases and represent them using *skos:exactMatch*.

Category:

Categories describe the subjects of a given entity, and we use *dcterms:subject* to present categories for the corresponding entities. *skos:broader* and *skos:narrower* are used to represent hyponymy relationships.

Redirection:

Encyclopedias usually use redirections to solve the synonymous problem. Redirection relations are described by *skos:closeMatch* to connect two entities.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
      <rdf:Description rdf:about="http://kb/food/soy_isoflavones">
          <dc:title>soy_isoflavones</dc:title>
          <skos:exactMatch>http://kb/food/isoflavones</skos:exactMatch>
          <dcterms:subject>food</dcterms:subject>
          <skos:relatedMatch>http://kb/food/soybean_saponin</skos:relatedMatch>
          <kb:function>http://kb/disease/osteoporosis</kb:function>
          <kb:constituent>http://kb/food/daidzin</kb:constituent>
          <kb:relationWord>http://kb/relationWord/prevent</kb:relationWord>
      </rdf:Description>
</rdf:RDF>
```

Figure 4: A snippet of domain knowledge base.

Related Term:

In Hudong Baike and Baidu Baike, there are related entities of a given entity. For example, related entities of “大豆异黄酮 (soy isoflavones)” are “大豆皂苷(soybean saponin)”, “葛根异黄酮 (pueraria isoflavones)”. *skos:relatedMatch* is used to represent Related Terms.

Function:

Function represents therapeutic efficacy of corresponding food. For example, “大豆异黄酮 (soy isoflavones)” has effect on “骨质疏松 (osteoporosis)” and “乳腺癌 (breast cancer)”. *kb:function* is used to represent Function.

Primary Constituent:

Primary constituent of a given food are represented by *kb:constituent*, for example the primary constituent of “大豆异黄酮(soy isoflavones)” includes “大豆甙(daidzin)”, “大豆甙元(daidzein)” and “染料木甙(genistin)”.

After concepts extraction, we utilize food and disease entities as anchor to extract relation words from biomedical corpus. In relation-bearing sentences, relation words are usually verbs, verb or prepositional phrases, such as “prevent”, “reduce mortality” and “with the increased risk of”, etc. Specifically, we use extraction patterns like “*<F verb D>*”; “*<F verb phrase D>*” and “*<F prepositional phrase D>*” to extract relation words. “*F*” and “*D*” represent food and disease entity, respectively. After relation words extraction, we filter out relation words those less than 5 times. We also assign a URI *kb/relationWord/word* to each relation word and use *kb:relationWord* to represent relations.

Finally, we use Resource Description Framework (RDF) to describe the knowledge base. Due to the limited space, Figure 4 shows a snippet of domain knowledge base.

3.3 Sentence Simplification

As discussed above, the characteristic complexity of the sentences in biomedical text challenges the relationship mining task. Recently, researchers have paid attention to simplifying sentences (Bach *et al.,* 2011; Jonnalagadda *et al.,* 2009). However, these approaches usually use syntax information as learning features or to generate rules. This is a chicken and egg problem. Inspired by (Bach *et al.*, 2011), we develop a new sentence simplification model without using syntax parser. Moreover, ours uses domain knowledge to incorporate more constrains to reduce the search space and computational complexity. Benefits of this sentence simplification model are twofold: 1) Sentence structure is simplified, second, 2) Since we can obtain more simple sentences that contain only one-one relationship, it alleviates the data sparseness problem.

For a given multiple relation sentence, let *SF* and *SD* be food and disease entity set and *SV* be verb set. By combination, we have $n=|SF|*|SV|*|SD|$ simple sentences in candidate set *C*. *HSS* uses Function (1) and (2) to find out $m=|SF|*|SD|$ qualified simple sentences as the simplified results. Where s_i is simple sentence candidate and c is the complicated sentence. w^T is the weight vector, which needs to be estimated from training data. $f(s_i)$ is the feature function vector.

$$\arg\max \sum_{i=1}^{m} p(s_i \mid c)$$

$$p(s_i \mid c) = \frac{\exp(w^T f(s_i))}{\sum_{j=1}^{n} \exp(w^T f(s_j))}, s_i \in C$$

Besides the word count and distance features in (Bach *et al.*, 2011), we adopt several other learning features such as semantic features to model where the verb is semantic related to the relation words in domain knowledge base; entity class features to ensure that subject and object of simple sentences are food and disease entities; context features to model the part-of-speech information in relation words' contexts.

The workflow of the sentence simplification model is as follows: First, we extract all the food and disease entities by CRFs model and domain knowledge, and then we combine the food and disease entities with verbs to form simple sentence candidates. If we get *n* entities and *m* verbs, we can obtain *n*m*(n-1)* simple sentence candidates. Finally, we use the constraints to find true simple sentences.

Figure 5 illustrates an example of the sentence simplification procedure. In Figure 5, the initial sentence contains two disease entities "HepG2" and "gastric cancer", one food entity "green tea" and two verbs "suggest" and "prevent". Therefore, we have *3*2*2=12* simple sentence candidates as shown in Figure 5. Through semantic feature and entity class feature constraints, sentences using verb "suggest" as predicate verbs and sentences using disease entities as subject are filtered out from the candidate set. Finally, two sentences in shaded rectangles are obtained as single relation-bearing sentences.

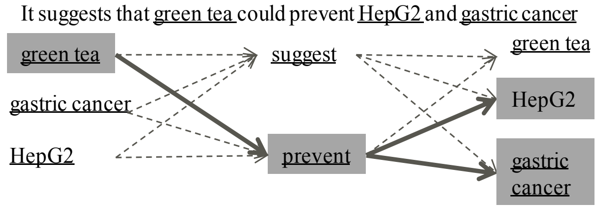


Figure 5: Workflow of the sentence simplification model.

3.4 Semantic Relation Mining

3.4.1 Extraction Model

We adopt CRFs models to extract relation words, because CRFs models are considered to be effective to solve the sequence labeling problem (Lafferty *et al.*, 2011). In addition, we can adopt flexible and abundant features such as lexical features, linguistic features and contextual clues to the process of CRFs model learning. Given a simple sentence of tokens, $x=x_1x_2...x_n$, we need to generate a sequence of labels $y=y_1y_2...y_n$. We define the set of possible label values as BIO to represent relation word.

We use a linear-chain CRF based on an undirected graph *G=(V, E)*, where *V* is the set of random variables. $Y=\{Y_i|1\leq i\leq n\}$ *and* $E=\{(Y_{i-1},Y_i)|1\leq i\leq n\}$ is the set of edges forming a linear chain. For a given sentence *x*, the conditional probability of a sequence of labels *y* is defined as follows:

$$p(y \mid x) = \frac{1}{Z(X)} \exp\left\{ \sum_{e \in E, k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V, k} \mu_k g_k(v, y|_v, x) \right\}$$

$$Z(x) = \sum_{y} \exp\left\{ \sum_{e \in E, k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V, k} \mu_k g_k(v, y|_v, x) \right\}$$

where f_k and g_k are binary feature indicator functions and λ_k and μ_k are weights assigned for each feature functions. *Z(x)* is a normalization factor of all state sequences.

3.4.2 Features Sets

One character that makes CRFs so attractive is that they transform the sequence labeling problem into finding an appropriate training feature set. In this paper, we define the following training features for each token/word x_i in an input sentence *x*.

Word Features:

We use two types of word features: unigram and bigram as learning features. In particular, we first remove stop words and then extract every single word as unigram feature and every two adjacent words as bigram feature. Bigram features can capture useful relation information, such as "reduce risk" and "decreased mortality", etc.

Part of Speech Features:

As relation words are mainly verbs, prepositional and verb phrases, part of speech might also play an important role in contributing to relation extraction.

In particular, we adopt Stanford tagger[3] to produce part of speech features.

Lexical Features:

In addition to word features and part of speech features, the model could also benefit from domain knowledge. In this research, we incorporate domain knowledge in the form of lexical features. For each token x_i, we include a binary feature that indicates whether or not the token is in our domain knowledge base.

4 Experiments

In this section, we first describe the dataset used in the experiments and then we report our experiment results to demonstrate the effectiveness of our approach.

4.1 Datasets and Evaluation Criteria

Since there is no open and available dataset for food and disease relationship mining task available in Chinese, we collect experimental dataset from *wanfang.com* and annotate it by three interns. We collected 3108 relation-bearing sentences, and used them as Dataset 1 to evaluate the performance of food and disease entity extraction. We randomly selected 706 sentences as Dataset 2 to evaluate the performance of food and disease relation extraction. The statistics of the annotated results are shown in Table 2.

In order to verify the degree of agreement among three annotators, we adopted Fleiss' Kappa (Sim and Wright, 2005) to evaluate the consistency of annotated results. The Fleiss' Kappa of Dataset 1 and Dataset 2 are 0.87 and 0.82, which shows strong consistency. To construct the final gold standard, we adopted the following procedure. For sentences that have received the same labels from all three annotators, we assigned this agreed-upon label. For a small number of sentences that have received different assessments, we had all three annotators go through these sentences and discuss their assessments with each other in a face-to-face meeting. We then used their consensual assessment as the final label.

Based on the above manually constructed gold standard, *precision*, *recall* and *F-Measure* are used in our experiments to evaluate the proposed approach, in which *precision* is defined as the ratio between the number of correctly extracted entities/relationships and the total number of entities/relationships extracted by the system, while *recall* is calculated as the number of correctly extracted entities/relationships divided by the total number of entities/relationships in the original sentences and *F-measure* is the weighted harmonic mean of the *precision* and *recall*.

$$F-measure = \frac{2\,precision * recall}{precision + recall}$$

	#Sentence	#Entities	#Relationships
Dataset 1	3108	2035	/
Dataset 2	706	629	1485

Table 2: Statistics of the datasets.

4.2 Food and Disease Entity Extraction Results

We use Dataset 1 to evaluate food and disease entity extraction performance. Specifically, we randomly select 50% as training data and the rest as testing data and repeat the experiment 10 times. We adopt CRFs as extraction models. Table 3 shows the average *precision*, *recall* and *F-measure*. From Table 3, we can see that CRFs model achieves promising results. Since sentence simplification model exploits entity type information as anchors to simplify multiple relation-bearing sentences, effective entity extraction model is very important for relation extraction.

	Precision	Recall	F-measure
Food Entity	98.7	84.6	91.1
Disease Entity	99.2	84	91

Table 3: Food and disease entity extraction results.

4.3 Food and Disease Relation Extraction Results

We implement a pattern-based method using strategy (Brin, 1998) and Yang's method as baselines. Table 4 shows the average *precision*, *recall* and *F-measure*. From Table 4, we can see

[3] http://nlp.stanford.edu/software/tagger.shtml

FDRM outperforms both PB and Yang's method, and FDRM increases *precision*, *recall* and *F-measure* by 2.4%, 2.3% and 2.4% respectively.

Method	Ave Precision	Ave Recall	Ave F-measure
PB	0.681	0.689	0.677
Yang	0.738	0.747	0.732
FDRM	0.762	0.77	0.756

Table 4: Food and disease relation extraction results.

Figure 6 shows the *F-measure* in l0 experiments. From Figure 6, we can see that FDRM outperforms the baselines across all experiments.

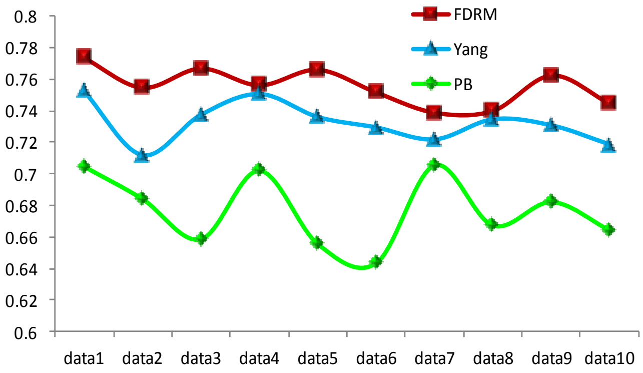


Figure 6: Relation mining results on F-measure.

We also conduct pairwise *t*-test to evaluate the improvement is significant or not. The *p*-values of FDRM and Yang, PB are 1.6E-04 and 3.75E-06 respectively and indicate the improvement is significant.

5 Conclusion

In this study, we propose a hybrid approach to extract and visualize food and disease relationships from Chinese biomedical text. As part of our work, we construct a domain knowledge base and develop a sentence simplification model. Experimental results on real-world datasets show the approach is promising. In addition, we find some interesting relationships, such as "<fresh milk, increase risk, lung cancer>". We believe that this study is just the first step in food and disease relationship mining and much more work needs to be done to further explore the issue. In our ongoing work, we will utilize more sophisticated nature language processing techniques such as co-reference resolution in the mining process. And we also plan to analyze polarity and strength of food and disease relationships.

References

Agichtein, E., and Gravano, L. 2000. Snowball: Extracting relations from large plain-text collections. *Proceedings of the 5th ACM International Conference on Digital Libraries*, pp.85-94.

Bach, N., and Badaskar, S. 2007. A review of relation extraction, *Literature review for Language and Statistics II*.

Bach, N., Gao, Q., Vogel, S., and Waibel, A. 2011. TriS: A statistical sentence simplifier with log-linear models and margin-based discriminative training, *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pp. 474-482.

Brin, S. 1998. Extracting patterns and relations from the World Wide Web. *Proceedings of the World Wide Web and Databases*, 1590(2), pp. 172-183.

Bundschus, M., Dejori, M., Stetter, M., Tresp, V., and Kriegel, H.P. 2008. Extraction of semantic biomedical relations from text using conditional random fields, *BMC Bioinformatics*, Vol. 9.

Bunescu, R. C., and Mooney, R. J. 2005. Subsequence kernels for relation extraction. *In Advances in Neural Information Processing Systems*, pp. 171-178.

Finkelstein-Landau, M. and E. Mori. 1999. Extracting semantic relationships between terms: Supervised vs. unsupervised methods. *Proceedings of the Int. Workshop on Ontological Engineering on the Global Information Infrastructure,* pp. 71-80.

Fundel K., Kuffner R., and Zimmer R. 2007. RelEx-relation extraction using dependency parse trees, *Bioinformatics*, 23:365-71.

Girju, R. and Moldovan, D. 2002. Text mining for causal relations. *Proceedings of the FLAIRS Conference*, pp. 360-364.

Hakenberg, J., Leaman, R., Vo, N.H., Jonnalagadda, S., Sullivan, R., Miller, C., Tari, L., Baral, C., and Gonzalez, G. 2010. Efficient extraction of protein-protein interactions from full-text articles, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3).

Hearst, M. A. Automatic acquisition of hyponyms from large text corpora. 1992. *Proceedings of the 14th COLING*, pp. 539-545.

Jonnalagadda, S., and Gonzalez, G. 2009. Sentence simplification aids protein-protein interaction extraction, *Proceedings of the 3rd International Symposium on Languages in Biology and Medicine*.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*.

Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., Andronis, C., Konstandi, O., and Persidis, A. 2007. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach, *Artificial Intelligence in Medicine*, 39(2):127-36.

Sharma, A., Swaminathan, R., and Yang, H. 2010. A verb-centric approach for relationship extraction in biomedical text. *Proceedings of the fourth IEEE International Conference on Semantic Computing, Pittsburg*.

Sim, J., and Wright, C. C. 2005. The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements, *In Physical Therapy*, 85(3), pp. 257-268.

Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J., and Ananiadou, S. 2011. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics*, 27(13), pp. i111-i119.

Weikum, G., and Theobald, M. 2010. From information to knowledge: harvesting entities and relationships from web sources. *Proceedings of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems of Data*, pp, 65-76.

Yang, H., Swaminathan, R., Sharma, A., Ketkar, V., and Silva, J.D. 2011. Mining biomedical text towards building a quantitative food-disease-gene networks, *book chapter in Learning structures and schemas from documents.*

Zelenko, D., Aone, C., and Richardella. 2003. A. Kernel methods for relation extraction. *Journal of Machine Learning Research*.

Zweigenbaum P., Demner-Fushman D., Yu H., and Cohen K.B. 2007. Frontiers of biomedical text mining: current progress, *Briefings in Bioinformatics*. 8(5). pp. 358-375.