A kernel based SVM for Semantic Relations Extraction from Biomedical Literature

Abstract

To identify and extract semantic relationships among named entities, relation extraction is a significant approach in knowledge representation. In order to capture the semantic as well as syntactic structures in text and to enable deep understanding of biomedical literatures, relation expression become essential. The automatic extraction of disease gene relations is presented by utilizing shallow linguistic features of global and local word sequence context with string kernel based supporting vector machine (SVM) for efficient disease-gene relation extraction. The performance of the proposed work shows that the bag-of-features kernel-based SVM classification is a promising resolution for specific disease-gene association mining.

Keywords: Biomedical Relation Extraction, Natural Language Processing, Machine Learning, Biomedical Literature.

1. INTRODUCTION

The semantic relations of the entities in biomedical literature are expressed in the form of attributes which can be static facts, or dynamic events that exist between entities. Information extraction (IE) concerns itself with extraction of those entities and their semantic relations from the unstructured text. In this work, the interest is focused in extracting fact relationships between biomedical entities, e.g. those facts that may imply a biological entity (gene or gene product) being a biomarker candidate of certain disease entity.

In almost all the information extraction systems, the bottom-up strategy is followed to extract structured information frames from unstructured text. The focus is to fill the predefined data frame with information extracted from the text similar to the relational database. In a simple example of extracting name, title, contact number from highly heterogeneous web pages, the IE task is to parse the web content and extract person name as named entity, and its attributes including title, phone number etc. This task consists of tokenization of the input text, analyze the morphological and lexical structure, analyze the syntactic structure, and combine above annotated components in a domain knowledge framework representing the entities and their relationships. Since the natural language has characteristic of long-distance dependency (Jianfeng Gao 2005), the co-reference or anaphora issue needs to be resolved in order to extract relations between entities. It is significant for particular domains like social study, where co-reference of names is frequent among sentences.

However this problem is much less significant in biomedical domain, therefore not handled in this work. The co-reference or anaphora resolution is resolved at each step by removing the ambiguous text words and parsed in a syntactic manner. This work focuses on extracting the semantic relationships in the context of biomarker definition, which can be viewed as a structured information framework containing disease and its associated genes or gene products.

Relation extraction is one of the most important subjects in IE which refers to the method of identification and extraction of semantic relationships between named entities in the text. Semantic and grammatical relations, negation, and co-reference are included in the relation. Protein-protein interaction and disease-associated gene mining are two examples of biomedical applications. The relation extraction task can be defined as to identify the relations specified above between two entities in the text, normally at the sentence level, and assign the relation type to one of predefined relation types. Supervised learning methods are used for relation extraction in case of large corpora of annotated data is available, or semi-supervised and bootstrapping method is used in case of limited annotated corpora.

With the help of statistical learning classifiers, context information surrounding the related entities are extracted as features for learning the relation by utilizing annotated positive and negative examples. Feature spaces that are useful for relation classification is reviewed in (Jiang and Zhai 2007). Among them, entity attributes e.g. entity types, bag of words, n-grams, grammar productions, dependency paths etc can be used as discriminative features for feature based classification. The huge feature space in large corpora training set makes it infeasible for search. In (Jiang and Zhai 2007) Bottom-up approach is used for exploiting those feature spaces by starting with a set of minimum features and adding more complex features to experiment the classification performances. Their results show that the basic unit features, which consists of bigrams and syntactic parse tree, is sufficient to achieve state-of-the-art performance while over fitting the classifier by adding complex features may decrease the overall performance. It suggests for each feature space, different feature representations may be redundant, even though it can increase robustness to noise but in the meantime may introduce more errors. Represented features should be selected to achieve better performance for feature based classification.

In (Zelenko et al. 2003) a kernel based relation classification method was introduced which is adapted from kernel method described in (Shawe-Taylor and Cristianini 2004). In contrast to feature based methods that directly rely on extracted features, kernel based methods utilize kernel function to compute the similarity score between pair of objects.

Let $\{x^i, E1^i, E2^i, r^i\}$ represent an input training instance where x^i denotes the sentence, $E1^i$ and $E2^i$ denote entities, r^i denotes the relationship and $r^i \in Y$ (relation types), $1 \leq i \leq N$ (N is the size of the training set). Let X_i denotes the $\{x^i, E1^i, E2^i\}$ of a training instance, and $X = \{E1, E2\}$ denotes a new instance for which the relation is to be predicted. The relation $\widehat{r}$ for the new instance can be computed by:

$$\hat{r} = argmax_{r \in Y} \sum_{i=1}^{N} \alpha_{ir} K(X_i, X) \quad (1)$$

where $K(X_i, X)$ is the kernel function for similarity computing, and α_{ir} can be estimated during training process (Sarawagi 2007). Kernel function $K(X_i, X)$ is defined over structures like parse tree or dependency graph, without the need to convert those structures to flat sequence of features required by feature based methods. In this paper we will present our work of extracting disease-gene relationship from text corpora based on kernel method and SVM classifier.

2. RELATED WORKS

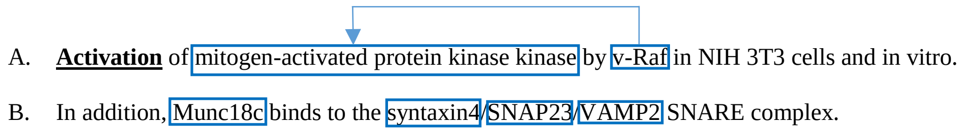
Relation extraction has been extensively studied and reported in newspapers, web content, emails etc. By querying PubMed in biomedical domain, with all known protein names it was found 269,000 out of 1.88 million PubMed abstracts were classified as being containing protein-protein interaction relations (Donaldson et al. 2003). In another study ~150,000 gene and protein relations were extracted from one million PubMed abstracts (Fundel et al. 2007). By the application of high throughput technology the numberer is rising in recent years. To facilitate automatic extraction of biomedical relations from the fast growing literature reports, BioCreative II (Hirschman et al. 2005b) and BioNLP (Pyysalo et al. 2012)

(Björne et al. 2010) have included relation extraction tasks for protein-protein interaction, co-reference, and entity relations extraction. Both events rely on annotated GENIA corpora and focused on PPI, protein-component and subunit complex relation extraction. The annotated corpora is significantly essential in relation extraction task for statistical machine learning based modeling, rule induction using rule-based methods, and for performance evaluation. Table 1 summarized current public annotated corpora for relation extraction in biomedical realm. For disease-gene relation extraction, to our knowledge, so far there is no publically available annotated corpora dedicated to this specific niche.

Table 1 Public biomedical corpora for relation extraction tasks. PPI denotes protein-protein interaction. AImed and HPRD50 are the only two corpora focusing on human PPI only.

Corpora	**Corpora size**	**Type**	**References**
AImed	225 abstracts	PPI (human)	(Bunescu et al. 2005)
BioInfer	1100 sentences	PPI	(Pyysalo et al. 2007)
HPRD50	145 sentences	PPI (human)	(Fundel et al. 2007)
IEPA	303 abstracts	PPI	(Ding et al. 2002)
LLL	77 sentences	PPI	(Nédellec 2005)
GENIA	9372 sentences	PPI	(Kim et al. 2003)
GREC	240 abstracts	Gene regulation	(Thompson et al. 2009)
IntAct	693 sentences	PPI	(Raja et al. 2013)

Generally, relation extraction can be binary or multi-way of directed or undirected entity pairs. For directed pair in subject-object relation, the object of the relation is the target and the subject entity is the agent. The binary relation involves only two entities related to each other. While the multi-way relations involves three or more entities linked by the relationship. The protein-component and subunit complex relation extraction is a multi-way relation extraction where typically more than three proteins or protein subunits form a functional complex. Above two relations are illustrated in figure 1.



A. **Activation** of mitogen-activated protein kinase kinase by v-Raf in NIH 3T3 cells and in vitro.

B. In addition, Munc18c binds to the syntaxin4/SNAP23/VAMP2 SNARE complex.

Figure 1 Illustration of common biomedical relations. A. Directed binary relation (activation) between two gene and protein pair. B. Undirected multi-way relation (binding) between subunits of a protein complex. (PMID 1326789, 16899085).

2.1 Machine learning and statistics based relation extraction

First we define the relation as $r(e_1, e_2, \dots, e_n)$ where e_i entities with relation r in the text. The sentence s from which e_i are identified can be represented as $s = (w_1, w_2, e_1, \dots, w_m, e_n)$, where $w_j(1 \leq j \leq m)$ is the word in the text. Given a corpora of positive and negative relation examples, in which e_i are annotated for a relation, the discriminative classifier can be trained using set of text features representing its local or global context shown below. Hence relation extraction can be represented as a classification problem that can be solved by supervised machine learning. The features in relation extraction which are commonly used are summarized below:

- Bag-of-words, bigrams surrounding the entity (before, between, and after), lemma
- Entity types
- The distance between entities and the word sequence
- Syntactic parse tree paths, tree distance between entities

Parse tree is a tree graph representing syntactic structure of natural language based on formal grammar (Feldman and Sanger 2007).The two types of parse tree used in text mining are Constituent parse tree and dependency parse with former one analyzed by constituency grammars (e.g. phrase structure grammars) and later one analyzed by dependency grammars without considering noun phrase (NP) or verb phrase (VP) categories. The constituency parse tree of a given sentence is more complex than its corresponding dependency parse tree and therefore is more computationally expensive. The constituency parse tree and dependency parse tree from an example PubMed sentence are given below. We generated both parse trees using the annotation pipeline in Stanford Core NLP toolkit.

"Activation of mitogen-activated protein kinase kinase by v–Raf in NIH 3T3 cells and in vitro" (PMID 1326789)

Constituency parse tree:

```
(ROOT
  (NP
    (NP
      (NP (NN Activation))
      (PP (IN of)
        (NP
          (NP
            (NP (JJ mitogen-activated) (NN protein) (NN kinase) (NN kinase))
            (PP (IN by)
              (NP (LS v))))
          (: --)
          (NP
            (NP (NN Raf))
            (PP (IN in)
              (NP (NN NIH) (NN 3T3) (NNS cells)))))))
    (CC and)
    (ADVP (FW in) (FW vitro))))
```

Dependency parse tree:

```
[Activation/NN
   prep_of:[kinase/NN
        amod:mitogen-activated/JJ
        nn:protein/NN
        nn:kinase/NN
        prep_by:v/LS
        dep:[Raf/NN prep_in:[cells/NNS nn:NIH/NN nn:3T3/NN]]]
    cc:and/CC
    advmod:[vitro/FW nn:in/FW]]
```

Using parse tree graph representation of the relation instance feature spaces for relation extraction was systematically exploited in (Jiang and Zhai 2007). Better performance is observed for constituency parse tree than dependency parse tree and sequence feature. But the difference is small, suggesting each of three feature spaces is capable of capturing most structural information between entities. Further comparison between unigram, bigram, and trigram features shows the bigram performs significantly better than unigram, but trigram feature didn't improve it further.

For sentence $s = (w_1, w_2, e_1, \dots, w_m, e_n)$where w_j is the word and e_i is the entity with defined relation, the feature set ideally should include as much discriminative power as possible for e_i while minimizing the computational cost. Feature set containing full syntactic parsing is also called heavy-weighted feature set.

Binary or multiclass relation extraction, different classifiers including SVM, Max Entropy, Naive Bayes etc., can be used for the classification task based on the specific relation extraction problem. Support Vector Machine (SVM) is the most commonly used machine learning classifier for relation extraction.

Figure 2 shows the linear SVM model trained with samples from two classes by the hyperplane H. The machine learning task is to find the hyperplane that can separate two classes of vectors with maximum margin between two of them.
For a training set with sample size of L
$(x_i, y_i), x_i \in R^d, y_i \in \{+1, -1\}, i = 1, 2, \dots, L$

The SVM is to find the hyperplane H: $W^T x + r = 0$

$$Maximize\ W(a) = \sum_{i=1}^{L} \alpha i = \frac{1}{2} \sum_{i,j=1}^{L} \alpha i\ \alpha j\ yi\ yi < xi, xj >$$

Subject to $\sum_{i=1}^{L} \alpha i\ yi = 0$
Where $\alpha i \geq 0, i = 1, 2, \dots, L$ (2)

αi is the non-negative Lagrangian multipliers. Vector $\alpha i > 0$ when x_i is a support vector, and $\alpha i = 0$ when it is not.

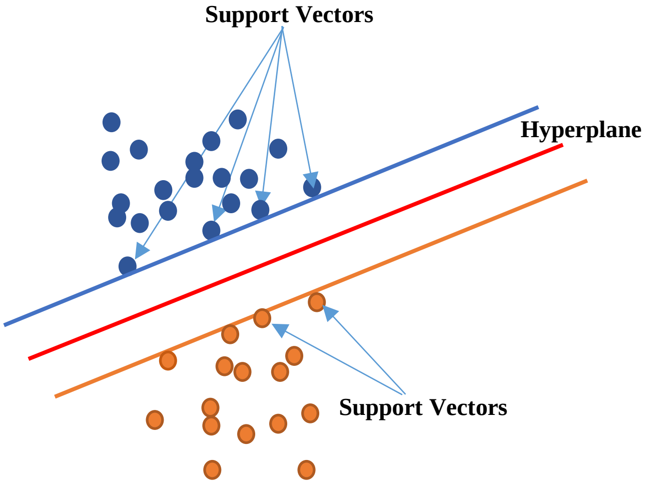


Figure 2 Illustration of Linear Support Vector and Hyperplane Separation.

In case the data points between two classes are not linearly separable, a kernel function is needed to map dataset into higher dimensional space so that classes become separable. Kernel function $K(X_i, X)$ can also be thought of a similarity function for pair of structures X and X_i in the feature space (Kim et al. 2008). Let's represent each training data instance i as $\left(x^i, E_1^i, E_2^i, r^i\right)$ and $X_i = (x^i, E_1^i, E_2^i)$, where r denotes the relationship, E denotes the entity, and x denotes the sentence. For a new instance (x, E_1, E_2) we can classify it by predicting its relation $\hat{r}$ with equation 3 (Sarawagi 2007).

$\hat{r} = argmax_{r \in y} \sum_{i=0}^{N} \propto_{ir} K(X_i, X)$ (3)
The $\propto_{ir}$ is the estimated weight for each training instance i, and y is the class. N denotes number of training sets.

Kernel for computing $K(X_i, X)$ used in relation extraction is based on string kernels proposed in (Lodhi et al. 2002), mathematically represented in equation 4 as:

$$K(X_i, X) = \sum_{u \in U} \emptyset_u(X_i)^T \emptyset_u(X) \tag{4}$$

where U is the set of all possible sub-structure in structure $X_i$and X. The $\emptyset_u(X_i)$ and $\emptyset_u(X)$ are decay factor ∈(0,1). The term "structure" can be generalized to any object including string, sequence of words, parse tree etc. For relation extraction the structures are represented as word sequences before/between/after related entities using Bag of features kernel approach, or parse trees containing the entity using Tree kernel approach (Bach and Badaskar 2007). Kernels developed using above approaches include tree kernel (TK) (Zelenko et al. 2003), dependency tree kernel (DTK) (Culotta and Sorensen 2004), shortest path dependency kernel (SPDK) (Bunescu and Mooney 2005), subsequence kernel (SK) (Bunescu and Mooney 2006), and composite kernel (CK) (Zhang 2006) (Zhang et al. 2011).

If the learning method utilizes only the labeled data for training, it is supervised machine learning. If it utilizes small set of labeled and large set of unlabeled data for training, it is semi-supervised. Semi-supervised methods rely on iterative learning by taking output of learner from last iteration and are becoming an important alternative to supervised approach, due to limited availability of high quality labeled data.

2.2 Pattern-based relation extraction

Handcrafted patterns or automatically generated patterns are used by this approach to extract relations. Patterns can be simple regular expression matching rules, or more complicated surface patterns consisting of POS tags and phrasal structures. Syntactic and semantic structure analysis is included in the most sophisticated pattern representation by full parsing, for instance to produce subject-verb-object (SVO) structure or predicate-argument structure (PAS) (Surdeanu et al. 2003). High Precision is one of the major advantage of manual pattern. Poor generalization from one domain to another is the major disadvantage for handcrafted patterns which also leads to relatively low recall because the manual pattern will not be able to cover all possible relation structures. Automatically generated patterns can be used to reduce the seriousness of this issue. Bootstrapping methods, for example, extract patterns from small set of relation examples (seeds) and iteratively expand the seeds by applying them on new data (Agichtein and Gravano 2000).

2.3 Disease and gene relationship extraction

Biomarker candidates like disease associated genes are used as indicators of diagnosis, disease progression, and treatment efficacy for the past several years. For example, in neurodegenerative diseases including Alzheimer's disease, Huntington's disease, Parkinson's disease, the genetic factor plays a critical role and consequently the disease-causing genes were studied extensively whereas the gene-disease relation extraction from literature haven't received similar level of attention as protein-protein interaction, protein and its sub-cellular localization. Hence there are more changes left to improve performance of disease-gene relation extraction. In this paper, we applied machine learning kernel methods based on works in (Bunescu and Mooney 2005) and (Giuliano et al. 2006) to extract Huntington disease - gene relation from PubMed literatures.

Global mining of general disease-gene association and selective mining of specific disease-gene associations are applied on disease-gene relation mining for information extraction needs. EDGAR is a system for global extraction of genes, drugs, and cell type’s interactions from PubMed literature and can be used to query the disease-gene associations (Rindflesch et al. 2000). BITOLA is a literature-based information extraction system designed to extract relations between different concepts, such as disease and gene association, by association rule algorithm (Hristovski et al. 2005). The association rule will be in the form of $x \rightarrow y(confidence, support)$ where support is their co-occurrence frequency and confidence is the percentage of records containing y concept (e.g. pathological functions or symptoms) with all records containing x concept (e.g. disease). Initially, for disease (x) gene (z) relation association, the algorithm first finds all concepts y such that $x \rightarrow y$ and then finds all concepts z (e.g. genes) such that $y \rightarrow z$. The algorithm then filter out all concepts z whose chromosomal location do not co-localize with chromosomal location of disease concept x by using HUGO gene nomenclature and LocusLink genetic loci information. The remaining set of z concepts (genes or gene products) are ranked as candidate disease associated genes. A rule-based with keyword matching algorithm for disease-gene extraction was also presented (Jung et al. 2013). In another work (Chun et al. 2006a), the binary pair of gene-disease was extracted from PubMed sentences using dictionary based matching approach followed by machine learning NER filtering. In the filtering step large set of false positive introduced are removed by dictionary matching which improved precision of relation extraction by 26.7%, suggesting the critical role of entity recognition step for overall performance of disease-gene relation extraction. Gene relation extraction for particular disease, in (Chun et al. 2006b) annotated corpora for prostate and gastric cancers from PubMed were constructed to train the maximum entropy based NER and relation extractor. The authors reported a 92.1% precision of topic-classified relation recognition.

3. Experiments and Results

3.1. Experiment design and datasets

Huntington disease related gene extraction is used in the experiment and casted it to a binary classification problem. Taken following NER tagged sentence describing association between HD and NR2B, NR2A as an example:

We conclude that these two genes, coding for <GENE>NR2B</GENE> and <GENE>NR2A</GENE> subtypes mainly expressed in the striatum, may influence the variability in AO of <DISEASE>HD</DISEASE>. (PMID 15742215)

In this example two gene entities and one disease entity were identified to have disease-gene association relations *r(NR2B, HD)*, *r(NR2A, HD)*. Here we consider the relation being all molecular interactions including expression, genetic variation, regulatory modification, or general description of associations in the text.

Due to unavailability of annotated corpora for machine learning based disease-gene relation extraction, we started by constructing it using the PubMed citations in Genetic Association Database (GAD) (Kevin Becker, Kathleaen Barnes, Tiffani Bright 2004). GAD is a database containing manually curated genetic association information for human disease with links to corresponding PubMed citations. List of all PubMed ids related to Huntington disease from GAD is compiled and all abstracts are retrieved from PubMed using Entrez e-Util API. Abstracts were automatically split into sentences and tagged with CRF based NER tagger. Sentences with at least one gene mention and one disease mention were selected.

Since the disease-gene relation extraction is considered as a binary relation extraction, in case the sentence contains more than one gene or disease mention, our system automatically makes copies of the sentence (instance of the sentence) so that each gene-disease pair is tagged as a training example. After manual verification and curation of all tagged sentences, a training datasets consisting of 117 positive examples and 64 negative examples was constructed. The annotated corpora was then processed by contextual kernel functions, and used subsequently to train and test on SVM classifier by 10 fold cross-validation. The performance of the SVM classifier was compared against a protein-protein interaction golden standard corpora AImed, which collects only human protein interactions. Table 2 shows the statistics of the two corpora used in our experiments and figure 3 summarized the system architecture of our kernel based relation extraction system.

Table 2 Statistics of two corpora used in the experiments. The constructed Huntington disease corpora from PubMed contains 181 annotated sentences and the AImed corpora contains 5625 annotated sentences.

	AIMED dataset	**Huntington disease dataset**
Positive examples	1008	117
Negative examples	4617	64
Total	5625	181

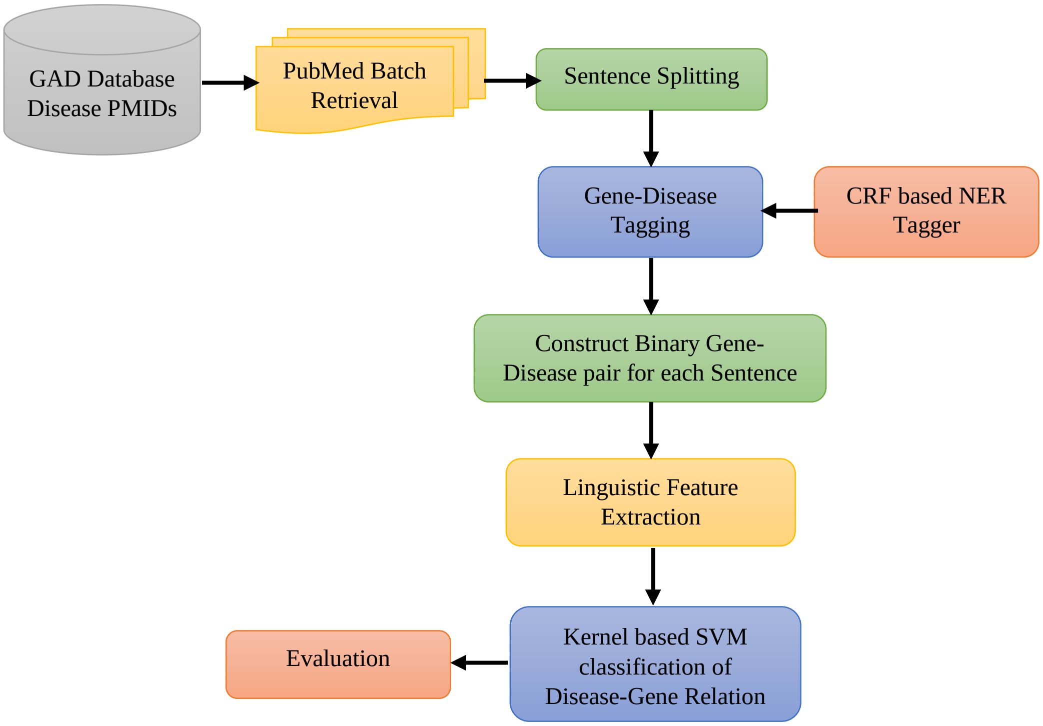


Figure 3 Work flow of kernel based disease-gene relation extraction system.

3.2 Kernel based SVM classifier for relation extraction

The input data are mapped using Kernel methods into a high dimensional feature space so that linearly non-separable classes become separable by a linear algorithm. For our disease-gene classification problem, we used kernel functions implemented in JRSE package (Giuliano et al. 2006) shown below (5 to 10):

$$K(x_1, x_2) = \frac{(\emptyset(x_1), \emptyset(x_2))}{\| \emptyset(x_1) \| \| \emptyset(x_2) \|} \tag{5}$$

The kernel is normalized by 2-norm of embedding vectors $\emptyset(x_1)$and $\emptyset(x_2)$. It is based on string kernel using bag-of-features approach. Given two entities and an interaction relation shown in following examples, (Bunescu and Mooney 2006) found three patterns for words around related entities:

i) Fore-Between (FB): relation is asserted using words before and between the entities. For example "interaction of Entity_1 with Entity_2".

ii) Between (B): relation is asserted using words between entities. For example "Entitye_1 is associated with Entity_2".

iii) Between-After (BA): relation is asserted using words between and after entities. For example "Entity_1 and Entity_2 interaction".

Formally, for the relation R, all three patterns (P) can be represented as a row vector:

$$\emptyset_p(R) = (tf(t_1, P), tf(t_2, P), \dots, tf(t_n, P)) \qquad (6)$$

Where $t_i (1 < i < n)$ is the token in the pattern and $tf(t_i, P)$ is its frequency of occurrence in pattern P. For all three patterns a kernel termed Global Context kernel K_{GC} is defined as:

$$K_{GC}(R_1, R_2) = K_{FB}(R_1, R_2) + K_B(R_1, R_2) + K_{BA}(R_1, R_2) \qquad (7)$$

Where K_{FB}, K_B, and K_{BA} denotes kernels for Fore-Between, Between, and Between-After bag-of-words patterns based on equation 4 respectively.

It is observed in (Bunescu and Mooney 2006) above patterns use no more than 4 words to assert the relation. Therefore in our experiment for disease-gene relation classification, we used tri-grams contiguous tokens kernel.

In addition to above global context kernel KGC, a Local Context kernel (LC) is define to take following four features of related entities into account namely, Token, Lemma, POS tag, Orthographic (capitalization, punctuation, numerals).

The local context $LC = (t_{-w}, \dots, t_{-1}, t_0, t_{+1}, \dots, t_{+w})$ can be formally represented as a row vector:

$$\varphi_L(R) = (f_1(L), f_2(L), \dots, f_n(L)) \qquad (8)$$

For each feature at position L, the feature function f_i returns 1 if it is active, or 0 otherwise. Here we used default window size of 1. The local context kernel K_{LC} for entity $E1$ and $E2$ is therefore defined as

$$K_{LC}(R1, R2) = K_{E1}(R1, R2) + K_{E2}(R1, R2) \qquad (9)$$

Finally, the combo kernel K_{SL} combining both global and local context kernel is defined as:

$$K_{SL}(R1, R2) = K_{GC}(R1, R2) + K_{LC}(R1, R2) \qquad (10)$$

Table 3 summarized the kernels and its configuration used in our experiments.

Table 3 Kernels and configuration used in the experiments.

Kernel	**Feature**	**Configuration**
Global Context (GC)	Fore-between	Tri-gram
	Between	Tri-gram
	Between-after	Tri-gram
Local Context (LC)	Token, lemma, POS, orthographic	Windows size = 1
Shallow Linguistic (SL)	GC + LC	Tri-gram, window = 1

All kernels in the toolkit are embedded into SVM package LIBSVM (Chang and Lin 2011) for model training and testing.

3.3 Evaluation of linguistic context based kernel method on AImed corpora

Before applying above kernel based classification methods on Huntington disease corpora, the performance on the human protein-protein interaction corpora AImed is evaluated. Table 4 shows the performance matrices (precision, recall, and F-measure) using 10-fold cross-validation. The results indicate global context kernel performs significantly better than local context kernel, and the combined kernel slightly increased the F-measure by 0.76%.

Table 4 Performance evaluation of three kernel based methods on human protein-protein interaction corpora AImed.

LC			**GC**			**LC + GC**		
Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
0.4424	0.7332	0.5493	0.6245	0.777	0.6912	0.6212	0.8014	0.6988

3.4. Disease-gene relation extraction from Huntington disease corpora

The linguistic kernel based SVM classification is applied on our Huntington disease corpora. Similar to the results in 3-3, table 5 shows global context kernel outperformed local context kernel in our binary disease-gene relation classification task, with significant increase of recall by 15.31% and F-measure by 9.34%. Compared with global context kernel, the combined kernel decreased the F-measure by 4.7% and recall by 7.93%. From the analysis, it shows the most discriminative linguistic characteristics are largely contained in tri-grams global context before, between, and after two related entities in our annotated corpora.

Table 5 Kernel based disease-gene classification using annotated Huntington disease corpora.

LC			**GC**			**LC + GC**		
Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
0.9621	0.7205	0.8211	0.9623	0.8736	0.9145	0.9614	0.7943	0.8675

4. Conclusion and Future work

Full shallow linguistic parsing estimates the basic text structure using bag-of-features approach in opposition to full syntactic tree parsing. The initial results obtained using shallow linguistic kernel methods on an annotated Huntington disease corpora suggest the global tri-grams context surrounding related entities are critical and essential for disease-gene relation extraction, which is in agreement with PPI relation extraction evaluation using AImed corpora. However it is noted that, due to limited Huntington disease PubMed citations from GAD, there are chances that our annotated dataset is relatively small, which will likely miss some complicated sentences in real-world. The linguistic kernel based machine learning approach is exploited in extracting relations between disease and gene. The results suggest bag-of-features kernel-based SVM classification is a propitious decision for specific disease-gene association mining. With future expansion of the training corpora, it can be applied on real-world problem for known disease associated gene extraction and novel gene prediction. As a future work, the corpora size can be increased by adding new PubMed abstracts referenced by other gene-disease relation databases.

REFERENCES

1. Agichtein E, Gravano L (2000) Snowball : Extracting Relations from Large Plain-Text Collections.
2. Bach N, Badaskar S (2007) A survey on relation extraction. Lit. Rev. Lang. Stat. II
3. Björne J, Ginter F, Pyysalo S, et al. (2010) Complex event extraction at PubMed scale. Bioinformatics 26:i382–90. doi: 10.1093/bioinformatics/btq180
4. Bunescu RC, Mooney RJ (2005) A shortest path dependency kernel for relation extraction. Proc Conf Hum Lang Technol Empir Methods Nat Lang Process pages:724–731. doi: 10.3115/1220575.1220666
5. Bunescu RC, Mooney RJ (2006) Subsequence kernels for relation extraction. Adv Neural Inf Process Syst 18:171.
6. Chang C-C, Lin C-J (2011) LIBSVM. ACM Trans Intell Syst Technol 2:1–27. doi: 10.1145/1961189.1961199
7. Chun H-W, Tsuruoka Y, Kim J-D, et al. (2006a) Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Pacific Symp Biocomput 15:4–15.
8. Chun H-W, Tsuruoka Y, Kim J-D, et al. (2006b) Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. BMC Bioinformatics 7 Suppl 3:S4. doi: 10.1186/1471-2105-7-S3-S4
9. Culotta A, Sorensen J (2004) Dependency tree kernels for relation extraction. Proc 42nd Annu Meet Assoc Comput Linguist ACL 04 4:423–es. doi: 10.3115/1218955.1219009
10. Donaldson I, Martin J, de Bruijn B, et al. (2003) PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics 4:11. doi: 10.1186/1471-2105-4-11
11. Feldman R, Sanger J (2007) The text mining handbook: advanced approaches in analyzing unstructured data. Casualty Actuar Soc E-Forum, Spring 2010 423.
12. Fundel K, Küffner R, Zimmer R (2007) RelEx--relation extraction using dependency parse trees. Bioinformatics 23:365–71. doi: 10.1093/bioinformatics/btl616
13. Giuliano C, Lavelli A, Romano L (2006) Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. EACL 18:401–408.
14. Hirschman L, Yeh A, Blaschke C, Valencia A (2005b) Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 6 Suppl 1:S1. doi: 10.1186/1471-2105-6-S1-S1
15. Jianfeng Gao HS (2005) Long Distance Dependency in Language Modeling: An Empirical Study. Lect Notes Comput Sci 3248:396–405. doi: 10.1007/b105612
16. Jiang J, Zhai C (2007) A Systematic Exploration of the Feature Space for Relation Extraction. HLT-NAACL
17. Jung J-Y, Deluca TF, Nelson TH, Wall DP (2013) A literature search tool for intelligent extraction of disease-associated genes. J Am Med Inform Assoc amiajnl–2012–001563–. doi: 10.1136/amiajnl-2012-001563
18. Kevin Becker, Kathleaen Barnes, Tiffani Bright AW (2004) The Genetic Associated Database. Nat Genet 36:431–432.
19. Kim S, Yoon J, Yang J (2008) Kernel approaches for genic interaction extraction. Bioinformatics 24:118–26. doi: 10.1093/bioinformatics/btm544
20. Pyysalo S, Ohta T, Rak R, et al. (2012) Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. BMC Bioinformatics 13 Suppl 1:S2. doi: 10.1186/1471-2105-13-S11-S2
21. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 517–28.
22. Sarawagi S (2007) Information extraction. Found Trends Database 1:261–377. doi: 10.1561/1500000003
23. Shawe-Taylor J, Cristianini N (2004) Kernel Methods for Pattern Analysis. 462.
24. Surdeanu M, Harabagiu S, Williams J, Aarseth P (2003) Using Predicate-argument Structures for Information Extraction. Proc. 41st Annu. Meet. Assoc. Comput. Linguist. pp xx–xx
25. Zelenko D, Aone C, Richardella A (2003) Kernel methods for relation extraction. J Mach Learn Res 3:1083–1106.
26. Zhang M (2006) A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features. Proc 21st Int Conf Comput Linguist 44th Annu Meet ACL 825–832. doi: 10.3115/1220175.1220279
27. Zhang X, Gao Z, Rong Z, Zhu Y (2011) Two novel composite kernels for relation extraction. 2011 Int Conf Multimed Technol 5207–5210. doi: 10.1109/ICMT.2011.6002253