A Hybrid Approach to Extract and Classify Relation from Biomedical Text

Abdul Wahab Muzaffar¹, Farooque Azam¹, Usman Qamar¹, Shumyla Rasheed Mir² and Muhammad Latif¹

¹Dept. of Computer Engineering, College of E&ME, National University of Sciences and Technology (NUST), H-12, Islamabad, Pakistan
²Dept. of Computer Science and Software Engineering, University of Hail, Hail, Saudi Arabia

Abstract - Unstructured biomedical text is a key source of knowledge. Information extraction in biomedical is a complex task due to the high volume of data. Manual efforts produce the best results; however, it is a near impossible task for such a large amount of data. Thus, there is a need of tools and techniques in biomedical text to extract the information automatically. Biomedical text contains relationships between entities of importance for practitioner and researcher. Relation extraction is an important area in biomedical which has gained much importance. The main work is done on rule based and machine learning techniques in biomedical relation extraction. Recently the focus has changed to hybrid algorithms which have shown better results. This research proposed a hybrid rule based approach to classify relations between biomedical entities. This approach uses Noun-Verb based rules for the identification of noun and verb phrases. It then uses support vector machine, a machine learning technique to classify these relations. Our approach has been validated on standard biomedical text corpus obtained from MEDLINE 2001, an accuracy of 94.91%.

Keywords: Relation Extraction; NLP; SVM; Classification

1 Introduction

With the huge information hidden in the biomedical field, in the form of publications, that is growing exponentially, it is not possible for researchers to keep himself updated with all developments in specific field [1, 2]. The emphasis of biomedical research is shifting from individual to whole systems, with the demand of extracting relationships between entities, e.g. protein-protein interaction, diseases-genes) from biomedical text to produce knowledge [3, 4]. Manual efforts to transform unstructured text into structured are a laborious process [5]. Automatic techniques for relation extraction are a solution to the problem [6].

Numerous relation extraction techniques for biomedical text have been proposed [8-10]. These techniques can be categorized into four major areas i.e. co-occurrence based, pattern-based, rule-based, and ML-based approaches. Co-occurrence is the simplest technique that identifies entities co-occurs in a sentence, abstract or document [11]. Pattern-based systems use a set of patterns to extract relations; these patterns can be manually defined patterns or automatically generated patterns. Manually define patterns involve domain experts for defining patterns and is the time consuming process and have low recall [12]. To increase recall of manually generated patterns automatically generated patterns can be used. An automatic pattern generation can be used bootstrapping [13] or generate directly from corpora [14]. In rule-based systems set of rules can be built to extract relations [15, 16]. Rules-based systems can also be manually defined and automatically generated from training data. When the annotated corpora on biomedical is available machine learning based approaches become more effective and ubiquitous [17, 18].

Most approaches use supervised learning, in which relations extraction tasks are modeled as classification problems. Broadly, any relation extraction system consists of three generalized modules i.e. text preprocessing, parsing and relation extraction. In this paper, we proposed a hybrid approach to extract relations between disease and treatment from biomedical text.

2 Related Work

Research proposed by [19] focused on a hybrid approach to discover semantic relations that occur between diseases and treatments. Cure, Prevent and Side Effects are the semantic relations that have considered to be extracted between entities (Disease, Treatment). The authors claimed better results compared to previous studies done on this topic. Results show different figures for each of the three relations mentioned: Accuracy for Cure relation is 95%, Prevent relation has 75% Accuracy, 46% accuracy for Side Effect relation has been claimed.

This paper [20], primarily focused to extract the semantic relations from biomedical text. Relations are extracted between
The authors propose an approach, which is a hybrid in nature, having two different techniques, to extract the semantic relations. In first technique relations extracted by pattern based on human expertise and in second one machine learning technique based on support vector machine classification. This new hybrid approach mainly relies on manual patterns when available relations examples are less, while feature values are used more when the number of available relations examples are sufficient. The authors claimed an overall F-measure of 94.07% for a cure, prevent and side effect relation.

### 3 Proposed Framework

The proposed work is a hybrid approach which uses rule-based and machine learning technique for the extraction of semantic relations between disease and treatment entities from biomedical text. High level view of the proposed approach is as under:

#### Biomedical Text Corpus

#### Preprocessing Phase

#### Feature Extraction

#### Vector Representation and Class Label

#### Relation Classification

Fig. 1: Proposed Relation Extraction Approach

The framework is divided into five major steps as follows:

- Biomedical Text Corpus
- Preprocessing phase
- Feature Extraction
- Vector Representation and Class Label
- Relation Classification

Detailed flow of the proposed framework is given in Fig. 2 below:

#### 3.1 Biomedical Text Corpus

We used the standard text corpus that is obtained from [21]. This corpus/data set contain eight possible types of relationships, between TREATMENT and DISEASE. This dataset was collected from Medline 2001 abstracts. Relations are annotated from sentences taken from titles and abstracts.

Table I, presents the original data set, as published in previous research showing relationships and number of sentences.

<table>
<thead>
<tr>
<th>Sr #</th>
<th>Relationship</th>
<th>No of Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cure/Treat for Dis</td>
<td>830</td>
</tr>
<tr>
<td>2</td>
<td>Prevent</td>
<td>63</td>
</tr>
<tr>
<td>3</td>
<td>Side Effect</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>Disonly</td>
<td>629</td>
</tr>
<tr>
<td>5</td>
<td>Treatonly</td>
<td>169</td>
</tr>
<tr>
<td>6</td>
<td>Vague</td>
<td>37</td>
</tr>
<tr>
<td>7</td>
<td>To See</td>
<td>75</td>
</tr>
<tr>
<td>8</td>
<td>No Cure/Treat Not for Dis</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>None</td>
<td>1818</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>3655</strong></td>
</tr>
</tbody>
</table>

Table I. Original Dataset [21]
3.2 Preprocessing phase

In this phase following preprocessing steps has been carried out on the text corpora. We used GATE 7.1 [22, 23], for the preprocessing of corpus and used the ANNIE [23] plug-in of GATE. Following modules of ANNIE are used in this research to tag the corpus for further used of relation extraction phase:

This phase was specially aimed to extract noun and verb phrases from corpora in order to provide semantic features to classifier.

3.2.1 ANNIE English Tokenizer in GATE

The Tokenizer convert the whole text into splits the text into small tokens, i.e. different type of words, punctuations and numbers [24].

3.2.2 Annie Sentence Splitter

The sentence splitter splits the text into sentences, which is required for tagger. The sentence splitter uses a dictionary list of abbreviations to differentiate between full stops and other token types [25].

3.2.3 ANNIE POS Tagger

This module produces a part-of-speech tag and annotates each word or symbol in the text. Part of speech tags can be a verb, noun, adverb or adjectives. Tagger can be customized by changing rule set given to it.

3.2.4 GATE Morphological analyzer

The Morphological Analyzer takes as input a tokenized GATE document. It identifies the lemma and an affix of each token by considering token’s part of speech tag, one at a time. These values are then added as features on the Token annotation. Morpher is based on certain regular expression rules [26]. This module uses to identify the common root of words in the text.

3.2.5 Noun-Verb-Noun Rules

We write rules and implement in JAPE [27]. These rules extract the Noun and verb phrases from text corpora. First three rules are extracting the noun phrases from text based on different criteria given in existing literature, while the last rule is extracting verb phrases. The rules are as under:

**Rule: NP1**

```
\n(nonePhrase -->
  nonePhrase.NP = {kind="NP", rule=NP1}
)
```

This rule is chunking a noun phrase whenever the chunker finds an optional determiner (DT) or personal pronoun (PRP) followed by zero or more adverbs (RB,RBR,RBS) followed by zero or more adjectives (JJ,JJR,JJS) followed by one or more singular or plural noun (NN,NNS).

**Rule: NP2**

```

(nonePhrase -->
  nonePhrase.NP = {kind="NP", rule=NP2}
)
```

This rule is simple one describing that one or more proper nouns are to be annotated as the noun phrase.

**Rule: NP3**

```

(nonePhrase -->
  nonePhrase.NP = {kind="NP", rule=NP3}
)
```

This rule describes that a personal noun (PRP) can be optionally preceded by a determiner (DT).

**Rule: NounandVerb**

```

(nonePhrase -->
  nonePhrase.NP = {kind="NP", rule=NP1}
)
```

This rule chunk all the verb phrases on the basis of one or more types of verbs ( VB ,VBD,VBG, VBN ,VBZ ).

With the use of above mentioned rules we chunked all the verb phrases and noun phrases from the text. Table II demonstrates the 10 examples of both most frequent phrases occurred in the text.

<table>
<thead>
<tr>
<th>S/No</th>
<th>VP</th>
<th>NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>be treat</td>
<td>the treatment</td>
</tr>
<tr>
<td>2</td>
<td>be perform</td>
<td>the use</td>
</tr>
<tr>
<td>3</td>
<td>to evaluate</td>
<td>the effectiveness</td>
</tr>
<tr>
<td>4</td>
<td>be administer</td>
<td>a combination</td>
</tr>
<tr>
<td>5</td>
<td>to prevent</td>
<td>surgical management</td>
</tr>
<tr>
<td>6</td>
<td>to moderate</td>
<td>this retrospective study</td>
</tr>
<tr>
<td>7</td>
<td>to determine</td>
<td>the risk</td>
</tr>
<tr>
<td>8</td>
<td>should be consider</td>
<td>successful treatment</td>
</tr>
<tr>
<td>9</td>
<td>to examine</td>
<td>the purpose</td>
</tr>
</tbody>
</table>
be to compare the efficacy and safety

### 3.3 Feature Extraction

A feature is anything that can be determined as being either present or absent in the item [34]. Feature extraction tries to find new data rows that can be used in combination to reconstruct rows of the original dataset and rather than belonging to one cluster, each row is created from a combination of the features [28].

#### 3.3.1 Unigrams features on basis of POS

We considered unigram features on the basis of part of speech (POS). Unigram consist of a single word labeled with a single attribute. Bag-of-word features and non-conjunctive entity attribute features are unigrams [29]. One vocabulary initially builds at this stage to refine unigrams features.

For unigram feature extraction, we build the vocabulary based on words of mainly four POS groups, i.e., Adjectives (a), Adverbs (r), Verbs (v) and Nouns (n). Instead of each token string for a word its term root is used obtained with use of Morphological analyzer. By considering these term roots vocabulary size is much reduced. For simplicity and later comparison, we grouped the tagger POS designation into following equivalent POS.

<table>
<thead>
<tr>
<th>POS</th>
<th>Tagger’s POSs</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>JJ, JJR, JJS, JJSS</td>
</tr>
<tr>
<td>r</td>
<td>RB, RBR, RBS</td>
</tr>
<tr>
<td>v</td>
<td>VB, VBD, VBG, VBN, VBP, VBZ, MD</td>
</tr>
<tr>
<td>n</td>
<td>NN, NNP, NNS, NNPS</td>
</tr>
</tbody>
</table>

#### 3.3.2 Filtered on the basis of word size

Normally, words with length less or equal to 2 are stop words or erroneous and creating noise in classification task not such as “to”, “be” and “s” etc. So we consider word size > 2 char in our unigram vocabulary.

#### 3.3.3 Phrase Features

As shown from literature noun phrases and verb phrases are much important for relation classification we consider these phrasal features along with the unigrams.

Verbs and noun phrase chucker separate these phrases in the text based on the above described rules. We build a separate vocabulary for these phrasal features.

### 3.4 Vector Representation and Class Label

This phase mainly focuses on conversion of text into vectors so that it latter be used for classification. Vector is represented with the use of features and one important decision at this stage is to select each feature weight. Feature weight also affects the classification performance as shown from literature.

#### 3.4.1 Term Frequency:

How many times a feature actually occurred in the text. Easy to formulate, but suffer from limited classification performance.

#### 3.4.2 Term Presence

Either a feature occurred in a piece of text or not, it is a binary representation. Term Presence has following advantages over the other weighting techniques.

- Easy to formulate a document or text vector
- Less processing and computation in machine learning tasks
- Better classification results as evidence from literature

#### 3.4.3 TFIDF (Term frequency – inverse document frequency):

New weighting scheme gains importance for machine learning tasks in previous years. It is a relative weight of the feature with its document frequency. It requires more calculation to form a vector.

#### 3.4.4 Class Labels:

Following are class labels used with vector representation against each setup in multi class classification.

<table>
<thead>
<tr>
<th>Feature Distribution</th>
<th>Unigram</th>
<th>VP</th>
<th>NP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setup # 1</td>
<td>1267</td>
<td>31</td>
<td>101</td>
<td>1399</td>
</tr>
<tr>
<td>Setup # 2</td>
<td>673</td>
<td>8</td>
<td>37</td>
<td>718</td>
</tr>
<tr>
<td>Setup # 3</td>
<td>624</td>
<td>7</td>
<td>36</td>
<td>667</td>
</tr>
</tbody>
</table>

#### 3.3.4 Frequency based filtering

To finally chosen features both unigrams and phrasal vocabularies are filtered by the corpus frequency with a threshold so the important terms are only selected for features.

Feature frequency > 3

Table IV shows the total no of features for unigrams, verb and noun phrases.
3.5 Relation Classification

This phase mainly focuses on the classification of all those relations which exist between disease and treatment entities in the text corpora. We used SVM to classify relations in the dataset. Our main focus was to extract three main relations i.e. cure, prevent and side effect.

3.5.1 Classifier:

Relation extraction is a text classification problem, we used support vector machine (SVM) as SVMs have already been used to yield higher accuracy on related tasks like text categorization [30]. Support vector machines (SVM) have also been widely used in the PPI extraction task, and have shown competitive results over other learning methods ([31, 32]).

We use SVM classifier for our experimentation for which LIBSVM [33] is used that is integrated software for support vector classification, it supports multi-class classification. We used the LIBSVM in both setting i.e. Linear Kernel and Radial Based Function (RBF) Kernel. For linear kernel best results are obtained at c=0.5 while other parameters are on default settings, for RBF kernel best results are when g=0.05 and c=8.

4 Results and Discussion

4.1 10-fold Class Validation

The model reported here is a combination of unigram, verb and noun-phrases, biomedical, with an SVM classifier. We use cross validation when there is a limited amount of data for training and testing. We swap the role of data i.e. the data used once used for training will be used for testing and the data used for testing will use for training.

10 fold means we split the data into 10 equal partitions and the process for 10 times while ensuring that each partition is used for testing at once. Such that 10 % data for testing and 90% for training. Average the performance results of 10 x iteration to get the final results.

Table VI represents a comparison of the accuracy results obtained in previous work by [20] and our proposed approach. As we can see from the table, our technique has a major improvement over previous results in setup 2 and setup 3 and a slight less improvement in setup 1. As it can be seen, our approach is very constant in all three setups. Our results improve the previous ones in setup 1, setup 2 and setup 3 with the difference varying from: accuracy in setup 1 is almost same as that of previous approach 18 percentage point improvement (setup 2) to 46 percentage point improvement (setup 3).

4.2 Discussion

Table 7 shows the comparison of previous techniques and our approach on the same data set produced by [21]. In the current settings our approach gives the best accuracy. Oana Frunza and Diana Inkpen [20], claims even better results for these relations using UMLS. Our results differ in two aspects: firstly our assumption was to extract multiple-relations-per-sentence as done by Asma Ben Abacha and Pierre Zweigenbaum [22], while the objective of the authors in [20] was the relation detection in a one-relation-per-sentence assumption. Secondly, our contribution is a hybrid approach which is combination of rule based approach and machine learning approach. Rule based techniques have higher precision but low recall while machine learning has low precision when few training examples are available. In order to increase the recall of the rule based system, machine learning approaches can be combined with rule based approaches. To increase the precision of machine learning approaches, while training set is small, we may use rule base features.

5 Conclusion and Future Work

This research focuses on hybrid framework for extracting the relations between medical entities in text documents/ corpus. We mainly investigated the extraction of semantic relations between treatments and diseases. The proposed approach
relies on (i) a rule-based technique and (ii) a supervised learning method with an SVM classifier using a set of lexical and semantic features. We experiment this approach and compared it with the previous approach [20]. The results taken by our approach shows that it considerably outperforms the previous techniques and provides a good alternative to enhance accuracy of relation extraction in the biomedical domain, if few training examples are available.

In future we are planning to test our approach with other types of relations and different corpora; we will also work on multi-stage classifier to enhance the performance of relation extraction.

6 References


[24] https://gate.ac.uk/sale/tao/splitch6.html#x9-1300006.2

[25] https://gate.ac.uk/sale/tao/splitch6.html#x9-1400006.4

[26] https://gate.ac.uk/sale/tao/splitch23.html#x28-55200023.12


