A Framework for Extracting Biological Relations from Different Resources

Enas M.F. El Houby
Systems & Information Department, Engineering Division,
National Research Centre,
Dokki, Giza, Egypt

ABSTRACT
The World Wide Web provides a vast source of information of almost all types. Biological data specifically have increased dramatically in the past years because of the exponential growth of knowledge in biological domain. It is very difficult to search for the required data in unstructured documents. Text documents often hide valuable structured data. This data can be exploited if available as a relational table that could be used to answer queries or to perform data mining tasks. Manually extracting biological relations from published literature and transforming them into machine-understandable knowledge is a difficult task because biological domain comprises huge, dynamic, and complicated knowledge. Automatic extraction of semantic relation between biological terms from unstructured documents is challenging in information extraction and important task for deep information processing and management.

In this research, a framework has been developed to extract different relations between various biological entities from documents. Semi supervised approach has been used to develop the framework. It requires the user to just provide a handful of valid pairs as initial seeds of the target relation, with no other training. Different patterns can be generated from initial seeds, and then from these patterns additional relation pairs can be extracted. The results has showed that different relations can be extracted such as gene-disease, protein-protein.

Keywords
Bootstrapping; Information extraction; Semantic relation; Semi-supervised.

1. INTRODUCTION
There exists a vast amount of unstructured electronic text on the Web including newswire, blogs, email communications, governmental documents, chat logs, and so on. Biological domain represents one of data rich domain whose text hides data that would be best utilized in structured form. This data ranging from DNA databases to lists of diseases. However, this information is often scattered among many web servers and through various documents using many different formats. If structured tables could be extracted from the information hidden in unstructured text, they would form an unprecedented source of information, and then more complex queries would be able to run and be analyzed over these tables and reported precise results [1, 2].

Information extraction is the task of automatically extracting structured information from unstructured or semi-structured machine readable documents. With the continuous growth of biological knowledge, the information extraction tools become more and more important for researchers of the biological domain. It is important to develop information extraction system to automatically process online biological documents and extract biological relations between biological entities such as protein-protein interaction (PPI), gene-disease correlation and so on. Relation Extraction (RE) process is an important not a trivial task which deals with the problem of finding associations between terms within a text phrase. Extracting semantic relations between entities in biological text is a crucial step towards natural language understanding applications. Also queries could be answered more precisely if a table listing all the biological terms pairs and associated relations that are mentioned in the documents collection is available [3].

Automated discovery and extraction of biological relations from online documents, particularly MEDLINE texts, has become essential and urgent because such literature data are accumulated in a tremendous growth. Extracting relationships between biological terms is important to help the researches to know about the kinship that combines these terms. Relation extraction methods are useful in discovering several kinds of relations such as gene-disease, gene-gene, protein-protein interactions, and gene-binding conditions. Patterns like "Protein X binds with Protein Y" are often found in biomedical texts where the protein names are entities which are held together by the "bind" relation. Such protein-protein interactions are useful for applications like drug discovery etc. Other relations of interest are a protein's location inside an organism [3, 4]. Three different approaches have been applied for relations extraction which is supervised approach, unsupervised approach and semi supervised and bootstrapping approach.

The supervised approach is to train the system over a large manually tagged corpus, where the system can apply machine learning techniques to generate extraction patterns [5]. This approach has typically been applied to small corpora such as a collection of news wire stories, and has difficulty scaling to the Web. The difficulty with this approach is the need for a large tagged corpus, which involves a significant amount of manual labor to create. The applied techniques learn a language model or a set of rules from this set of hand-tagged training documents, and then apply the model or rules to new texts. Models learned in this manner are effective on documents similar to the set of training documents, but they extract quite poorly when applied to documents with a different genre or style. As a result, this approach has difficulty scaling to the Web due to the diversity of text styles and genres on the web and the prohibitive cost of creating an equally diverse set of hand tagged documents [1, 6].

On the other hand, unsupervised approach aims to overcome the difficulty of supervised approaches, i.e. the need for hand-tagged data. Unsupervised methods of relation extraction [7]

apply clustering algorithms in order to group similar pairs of entities to the same cluster. Each cluster represents a relation between these pairs.

The semi supervised and bootstrapping approach has been an attractive alternative in automatic text processing. Bootstrap learning uses unlabeled examples for training; it only requires a small set of tagged seed instances or a few hand-crafted extraction patterns per relation to launch the training process, the systems' output is used to generate the training input for the next iteration. It works best in an environment like the World-Wide Web, or big corpus where the table tuples to be extracted will tend to appear in uniform contexts repeatedly in the collection documents. It exploits the regularity of language and the data redundancy in the collection to extract the target relation with minimal training from a user. It requires that the user just provide a handful of valid tuples of the target relation, with no other training. (This is in contrast to the way traditional information extraction systems operate) [2, 8].

In this research, the problem of extracting relations from biological sources will be addressed. Let's formulate the problem more formally:

Let **D** be a large database of unstructured information such as the biological corpus.

Let **R** = $r_1, \ldots, r_n$ be the target relation.

Every tuple pair **T** of **R** occurs one or more times in **D**. Every such occurrence consists of the pair of **T** which is represented as strings occurring in close proximity to each other in **D**. Every such occurrence of **T** is connected by substring text keywords which represent pattern **P**.

In this research, the focus will be on the method of recognizing relations between terms in unstructured text in biological sources with minimal human intervention. A framework for extracting different structured biological relations from unstructured biological documents will be developed. The developed framework is based on semi supervised approach which exploits the duality between sets of patterns and relations to grow the target relation starting from a small sample. To extract a structured relation (or table), the proposed framework requires that the user just provide a handful of valid tuples of the target relations, with no other training. Different biological relations that appear in a given corpus can be extracted according to the initial seeds relations types such as if the initial seeds are protein-protein pairs then the extracted relation pairs will be protein-protein and so on.

The remainder of the paper is organized as follows: in section 2, an overview for the previous works related to our subject is presented. In section 3, the materials and methods of the proposed framework is described. In section 4, the results are produced, before drawing conclusions and future work in section 5.

2. RELATED WORK

A lot of previous works were studied and a large number of methods have been developed to extract relations from biological documents. They can be classified into supervised relation extraction, semi-supervised relation extraction and unsupervised relation extraction methods.

Much of the previous work on relation extraction has focused on the use of supervised learning techniques such as Craven M et al. proposed a system that focuses on detecting associations between proteins and sub cellular locations by using machine learning methods to introduce times for extracting facts from text [9]. Abulaish M & Dey L proposed an ontology-based biological information extraction and query answering (BIEQA) system which extracts biological relations from MEDLINE abstracts for tagged documents. The extracted relation is assigned a fuzzy value according to its frequency in the corpus. The extracted relations are stored in a database which is integrated with a query-processing module [10]. Frunza, Oana, et al. proposed a machine learning based methodology for building an application that extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments. The proposed methodology's outcomes could be integrated in an application to be used in the medical care domain [11]. Wen-Juan Hou &Hsiao-Yuan Chen used automatic rule-learning approach to gene–disease relationship extraction. All possible rules have been learnt that discriminate relevant from irrelevant sentences. The scores of the learned rules have been computed in order to select high ranked rules [12]. Kang Ning, et al., developed a knowledge-based relation extraction system using training data, and applied the system for the extraction of adverse drug events from biomedical text. The system consists of a concept recognition module that identifies drugs and adverse effects in sentences, and a knowledge-base module that establishes whether a relation exists between the recognized concepts [13].

Unsupervised learning has been applied in relation extraction to overcome the problem of hand-labeled training data, and attempted to find inherent patterns in the data that can then be used to determine the correct output value for new data instances [14]. Hasegawa et al. developed such an unsupervised approach. Their primary assumption is that pairs occurring in similar contexts share the same type of relation, hence they can be clustered together. They considered that two entities form a pair, if they co-occur in the same sentence and are separated by at most n intervening tokens [7]. Quan Changqin, et al. presented an unsupervised method based on pattern clustering and sentence parsing to deal with biomedical relation extraction. Pattern clustering algorithm is based on polynomial kernel method, which identifies interaction words from unlabeled data; these interaction words are then used in relation extraction between entity pairs. This approach has been applied on two different tasks which are protein–protein interactions extraction, and gene–suicide association extraction [15].

Relation extraction systems have addressed scalability with semi supervised approach and bootstrap learning techniques. This method of research is the used one in our work. Minlie Huang et al. proposed ontology-based biological relationship extraction system to automatically extract biological relationships from a huge number of online MEDLINE abstracts. Authors made ontology-based semantic annotation of online biological documents [3]. Carlson et al. used semi-supervised learning method for information extraction to extract new instances of concept categories and relations using an initial pre constructed categories and relations of ontology [16]. Chao Chen, et al. developed an automatic approach REV (Relation Extraction with Verification) to extract relations from World Wide Web, which just requires a few user specified seed instances as input set with the form of <e1, e2, keyword>. These instances are used to generate extraction rules that in turn result in new instances [17]. A bootstrapping, semi-supervised learning approach has developed to iteratively extract and rank drug–gene pairs

according to their relevance to drug pharmacogenomics. The availability of a comprehensive pharmacogenomics-specific drug–gene relationship knowledge base is important for personalized medicine to deliver the right drug to the right patient in the right dose [18]. Xu, Rong, et al. had used disease-manifestations pairs from existing biomedical ontologies as prior knowledge to automatically discover disease-manifestations specific syntactic patterns. Additional pairs have been extracted from MEDLINE using the learned patterns. Correlations between disease manifestations and disease-associated genes and drugs had been analyzed to demonstrate the potential of this newly created knowledge base in disease gene discovery and drug repurposing [19]. Xu Rong, et al., used a semi-supervised iterative pattern learning approach to build a precise, large-scale disease-disease risk relationship (D1 →D2) knowledge base (dRiskKB) from a vast corpus of free-text published biomedical literature [20].

3. Material and method

3.1 Material

In this research, different resources which are MEDLINE, OMIM and UNIPORT are used to collect corpus documents. MEDLINE is a massive biomedical corpus for biomedical researchers; it is one of the most comprehensive textual sources of biomedical information, it covers topics in biology, biotechnology, medicine, biochemistry, and other related fields. MEDLINE abstracts or free full text articles can be accessed through the PubMed database (http://www.ncbi.nlm.nih.gov/pubmed/). Online Mendelian Inheritance in Man (OMIM) is one of the most well-known databases that contains gene–disease annotations. The full-text overviews in OMIM contain information on all known Mendelian disorders and genes. It is a comprehensive knowledge base of human genes and genetic diseases. For biomedical researchers, OMIM serves as an important resource to support Mendelian inheritance information. It can be accessed through (http://omim.org/). UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. The human diseases in which proteins are involved are described in UniProtKB entries with a controlled vocabulary. It can be accessed through (http://www.uniprot.org/).

A dictionary has been created to recognize biological named entities in the documents. Biological named entity recognizer has been collected from OMIM and UNIPORT to capture as many biological entities as possible and organize them in a uniform format. It contains disease names list, genes names, protein names, cellular component and others.

3.2 The proposed Framework for the Biological Relation Extraction

In this research, a biological relation extraction framework has been developed to automatically extract several kinds of relations from biological documents. It can extract relations using semi supervised approach which can learn patterns and extract biological related pairs from the learned patterns.

The proposed semi supervised approach can extract a table listing as much biological terms pairs that are mentioned in the documents collection as possible. It consists of 1) corpus pre-processing phase to prepare the used corpus documents for the relations extraction phase, 2) relations extraction phase that comprises two steps; the first step is to generate patterns from different documents which relate different biological terms, and the second step is to extract related pairs from the corpus using the generated patterns.

Figure 1 illustrates the overall architecture of the proposed framework for the biological relations extraction. First, each document in the corpus is pre-processed by tagging it using Named Entity Recognizer (NER). Different biological resources and databases have been used to create NER which provide diseases' names, genes' names and other biological entities. The output of this phase is a tagged corpus. Next, the relation extraction phase which start by any related biological terms' pairs as initial seeds to generate patterns by searching the tagged corpus which in turn are used to generate new seeds pairs and so iteratively discover new patterns and extract new pairs with newly discovered patterns until no significant patterns and terms pairs can be extracted. The extracted biological relations are stored in the associated knowledge base. Where the extracted relations depend on the initial seeds pairs, if the user provides gene-disease pairs, it is expected that the generated patterns are related to gene-disease and so the extracted pairs from generated patterns.

3.2.1 Corpus Preprocessing Phase

In corpus pre-processing phase, the biological named entity recognizer has been used to identify all appearances of biological entities in the documents of the corpus. A dictionary-based method has been used to build named entity recognizer. It collects biological entities from different resources and databases as mentioned before. It has been used to recognize biological terms in the corpus' documents including recognizing genes names, diseases names, proteins names and other biological terms. Table1 shows a sample of generated NER dictionary which is used to tag the corpus.

Table1: Sample of Named Entity Recognizer (NER)

Biological Term	Term's Type
Cervical cancer	Disease
Chordoma	Disease
Cystic fibrosis	Disease
IGHG3	Gene name
PARK7	Gene name
PKD2	Gene name
Protein AF1q	Protein name
Serum albumin	Protein name
repairosome	cellular_component
protein deneddylation	biological_process
……..	……..

3.2.2 Relations Extraction Phase

Relations extraction phase is the main phase in the framework; its purpose is extracting knowledge bases or tables listing different biological relations between different biological entities from corpus' documents. It extracts various semantic relations exemplified by a given small set of seed instances. It requires only a handful of training examples of the target relations from the users. These examples are used to generate patterns that in turn result in new pairs being extracted from the corpus documents which can be used to generate new patterns and so on, the process can be repeated and collect as much relations from documents as possible. It terminates when a predefined stopping condition is met some stopping criteria, such as no new seed relations are extracted or after specific number of iterations.

3.2.2.1 Patterns Generation

The pattern generation module is initially given a handful of example relation pairs. For each pair, the segments of text in the corpus' documents that connect terms' pair which occur close to each other are analyzed. A key step in generating patterns is finding where biological terms, which were previously tagged using named-entity recognizer, occurred in the text. The occurrences of biological terms will be focused on and compared with the searched terms of seeds' pairs, and then the contexts that connect the occurrences of the searched pairs are analyzed. That is mainly because in English-language documents, the middle context between two terms is considered as the most indicative of the relationship between these terms especially between biological terms. The similar connected substrings for different occurrences of pairs are grouped together to represent different generated patterns. The frequencies are counted for different groups to record how many times each pattern occurred. Those whose frequencies are less than a user-specified threshold are removed from the pattern set and those higher than threshold represent the set of new generated patterns P, these patterns can be ranked according to their frequencies.

Figure 2 shows a sample of the retrieved sentences that contain the searched biological terms pair <sickle cell disease, HBB>. To retrieve these sentences, the tagged biological terms which are enclosed between tagged marks <> and </> are focused, where <D> </D> to enclose disease name and <G> </G> to enclose gene name. By analyzing the substrings that connect these terms, 2 different keywords are found. Similar keywords which connect these terms in different sentences are grouped and their frequencies are counted. As shown from example, 2 different patterns can be generated from grouped keywords which are:

< D> **associated with** <G > **3**
< D> **caused by** <G > **2**

3.2.2.2 Pairs extraction

After generating patterns, the most frequent patterns which are above threshold will be used to extract new pairs from the corpus. So the corpus' documents are scanned to search for segments of text that match the patterns. Any pair of terms connected by the keywords of patterns will not be matched unless that they are tagged as biological terms using named-entity recognizer. As a result of this process, new pairs are extracted and can be used as new seeds in the next iteration. The relation pair <Term1, Term2> is extracted if there is a searched keyword that matches the substring of the middle of the text segment that connects this pair.

As an example, the generated pattern < D> **caused by** <G > in the pattern generation step can be used to extract new pairs from the corpus by searching for the keyword "**caused by"** which is enclosed by tagged marks <D> and <G> and by collecting disease name which is enclosed between <D> </D> and gene name which is enclosed between <G> </G> that are found with keyword in the same sentence, new pairs can be collected. And so on many pairs can be collected using different generated patterns. Figure 3 shows examples of retrieved sentences that contain the searched pattern. The pairs of terms connected by the keyword "**caused by**" and tagged as disease and gene using named-entity recognizer are extracted as new pairs and can be added to the list of disease-gene pairs as shown in table2.

Algorithm1 sums up the full process to extract relations and collect different knowledge bases using semi supervised approach. In step 1 the biological terms have been tagged using NER. Steps (2 - 20), collect different relations pairs for different knowledge bases according to the initial seeds which are provided in step 3. Steps from 4-19 to extract relations for the same knowledge base until satisfy termination condition. Steps from 5-9, to generate patterns from the provided seeds, then the generated patterns are ranked in steps (10-12). Steps from 13- 17 to extract related pairs from the generated patterns and to add new pairs in the KB. the extracted pairs are used as new seeds in the next iteration as in step18.

Table2: The extracted disease-gene pairs

Disease	Gene
microdeletion syndrome	TBX1
Gaucher disease	GBA
Stargardt disease	ABCA4 ABCR

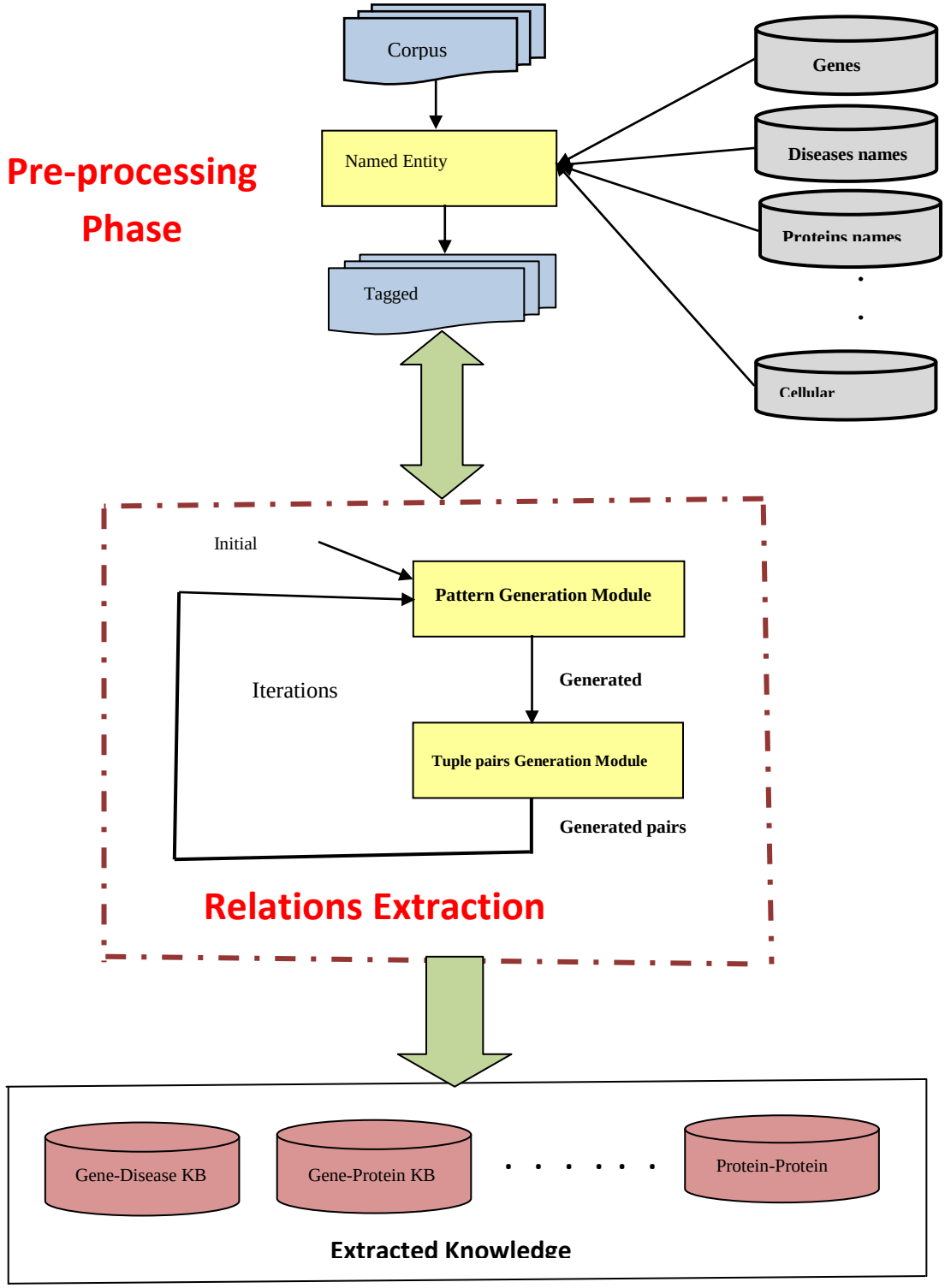

Figure 1: The architecture of the proposed framework for the biological relation extraction.

1. <D> sickle cell disease</D> **associated with** <G> (HBB) </G> gene.
2. <D> "sickle cell disease" </D> encompasses a group of symptomatic disorders **associated with** pathogenic variants in <G> HBB </G>.
3. <D> Sickle cell disease </D> (SCD) in Saudi patients from the Eastern Province is **associated with** the Arab-Indian (AI) <G>HBB </G>.
4. <D> Sickle cell disease </D> (SCD) is the most common human genetic disease which is **caused by** human β-globin <G> HBB </G> gene.
5. <D> Sickle cell disease </D> (SCD) and beta thalassaemia**, caused by** lesions that affect the <G>HBB </G>.

Figure 2: examples of sentences contain the searched biological terms

1. <D> microdeletion syndrome </D>, which is mainly **caused by** <G>TBX1</G> gene mutations.

2. <D>Gaucher disease<D> (GD) is the most common of the lysosomal storage disorders and is **caused by** defects in the <G>GBA</G> gene encoding glucocerebrosidase (GlcCerase).

3. <D> Stargardt disease </D> (STGD1) is a macular dystrophy **caused by** mutations in the <G> ABCA4 ABCR </G> gene.

Figure 3: examples of sentences contain the searched pattern

```
1. Use N.E.R. to tag biological terms in the corpus
2. Do to collect different relations types for different KBs
3.    Provide initial seeds of relation pairs according to target KB
4.    Repeat
5.      Repeat
6.        Analyze the segments of text that connects seeds pairs
7.        Group similar substrings that connect seeds pairs
8.        Count the frequency of each group
9.      Until no new group can be generated
10.     For each created group
11.       If group frequency > threshold
            Then generate pattern
12.     Rank generated patterns by frequencies
13.     Repeat
14.       Use generated patterns to extract relation pairs
15.       Compare the extracted relation pairs with those in the target KB
16.       If No similar relation's pairs in KB
            Then add new pairs to the target KB
17.     Until no new pairs can be extracted
18.     Use extracted relation pairs as seeds
19.   Until termination condition
20. End Do
```

Algorithm 1: The full process to extract relations and collect various knowledge bases

4. EXPERIMENTAL RESULTS

This section shows the empirical results of the proposed framework. To evaluate the proposed framework a small corpus consists of 1000 abstracts is considered as experimental data, involving a lot of genes, proteins, diseases and other biological terms. It is collected from MEDLINE, OMIM and UNIPORT as mentioned before.

Since semi supervised methods of relation extraction are always applied on large amounts of data. Therefore, getting an exact measure of precision and recall is difficult. So a small sample from the output is treated as a representative of the output and manually checked. Then, an approximate estimation of the precision is calculated. This procedure for evaluation has been used by [2] and [21].

So in this research a sample of documents is collected and used to manually compute approximated values for precision, recall and F-Score. The actual values can't be computed and therefore the values of error and coverage cannot be identified. The precision, recall and F-Score are computed as follow:

$$Precision = \frac{TP}{TP + FP} \qquad (1)$$

$$Recall = \frac{TP}{TP + FN} \qquad (2)$$

$$F\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (3)$$

A set of 20 documents which contain the (sickle cell disease, HBB gene) pair has been collected as sample corpus to manually verify the results and compute the precision, recall and F-score.

Starting with seed pairs: < **sickle cell disease**, **HBB** >

The generated patterns with their frequencies are:

< Sickle cell disease> **mutation** <HBB > **5**
< Sickle cell disease> **caused by** <HBB > **5**
< Sickle cell disease> **associated with** <HBB > **4**
< HBB> **inherited in** <sickle cell disease > **1**

By manually inspecting the results of the generated patterns, they are found as follow TP=11, FP=4 and FN=6

$$Precision = \frac{11}{11 + 4} = 73.3\ \%$$

$$Recall = \frac{11}{11 + 6} = 64.7\ \%$$

$$F\text{-}Measure = \frac{2 \times 73.3 \times 64.7}{73.3 + 64.7} = 68.73\ \%$$

If threshold is set as 3 then the pattern < **inherited in**> will be neglected where its frequency is less than threshold and the other 3 patterns will be considered for pairs extraction step where theirs frequencies are greater than threshold. These patterns are as follow:

<D> **mutation** <G > **5**
<D> **caused by** <G> **5**
<D> **associated with** <G> **4**

When the generated pattern <D> **caused by** <G> was used to extract new (disease, gene) pairs, a set of (disease, gene) pairs was extracted. Table3 illustrates a sample of extracted (disease, gene) pairs using <D>**caused by**<G> pattern. Table4 illustrates a sample of extracted (disease, gene) pairs using <D>**associated with**<G> pattern. Similarly <D>**mutation**<G> can be used to collect other related biological pairs. For the preliminary experiments, the stopping condition is set as one iteration starting from seeds pairs to generate pattern and then using the generated patterns to extract related pairs.

Starting by different seed pairs, the system can provide different patterns then extract different relations' pairs which are stored in the associated knowledge base. Since some patterns are common so that they can be matched to extract different pairs representing various relations target (knowledge bases), so the extracted pairs have been checked through tagging to be listed in the correct knowledge base. Table 5 shows a sample of the generated patterns and relations pairs which target different knowledge bases.

A crucial step in the table extraction process is the generation of patterns that will be used to find new pairs in the documents. Ideally, the patterns should to be both selective, so that they do not generate incorrect pairs, and to have high coverage, so that they identify many new relation pairs. As shown in table 5, although <*> in <*> is a correct pattern and through different runs always has high frequency (the same for pattern such as <*> and <*> but it is not selective enough to generate correct relation, so it should be omitted from the list of patterns. Until now the task of omitting generic patterns has been done manually, so some criteria and measurements should be set to select the confident patterns. Also the same for the extracted relations' pair they should be selected before generating new patterns. Domain expert is needed to inspect and verify the extracted results.

Table3: A sample of extracted (disease, gene) pairs using <caused by> pattern

Disease	Gene
Krabbe	GALC
Glycogen storage disease	GBE
Cholesteryl ester storage disease	LIPA
autoimmune disease	AIRE
Rett syndrome	MECP2
Charcot-Marie-Tooth disease	NDRG1
Niemann-Pick	NPC1
macular dystrophy	ABCA4
polycystic kidney	PKD1

Table4: A sample of extracted (disease, gene) pairs using <Associated with> pattern

Disease	Gene
dystonia	TOR1A
Moyamoya disease	TGFB1
galactosemia	GALE
Alexander disease	GFAP
Alzheimer disease	APP
Alzheimer disease	CAV1
periodic paralysis	KCNJ2
Colorectal cancer	CCND1

Table 5: A sample of the generated patterns and relations' pairs

Target Knowledge base	Pattern (Relation Type)	Extracted Pairs
Protein-Protein	Interact with	(Battenin, BIP)
Gene-Protein	Encode	(RAC2, GTP-binding)
Protein-Disease	Cause	(Presenilin-2, Alzheimer)
Gene-Protein	Encode	(LBR, bifunctional protein)
Gene-Disease	Affected with	(NKX2.5, congenital heart disease)
Gene-Disease	in	(NKX2-5, ventricular septal defects)
Protein-Protein	interact with	(heterogeneous nuclear ribonucleoprotein, TAR DNA-binding protein 43)

5. CONCLUSTION AND FUTURE WORK

In this research a framework has been developed which can extract different relations between various biological pairs. Semi supervised approach has been used. It requires no training other than providing a handful of initial seeds pairs to generate patterns which are used in turn to extract related biological pairs. The aim of the framework is to capture as many relation pairs mentioned in the corpus as possible and build different knowledge bases for different relations' pairs.

In the future the framework needs more evaluation and more analysis for the results by a specialist, then it can be applied in a big corpus or through World Wide Web. It is important to find method to select high confident patterns and relations' pairs and rank them according to their confidence values. Also it is important to find automated method to calculate the precision and recall. Some language related problems should be solved such as using different expressions, synonyms or abbreviations for the same scientific terms by different authors or even the same author, which may differ from that found in the dictionary.

6. REFERENCES

[1] Agichtein, E., & Gravano, L., Snowball: Extracting relations from large plain-text collections, Proceedings of the Fifth ACM International Conference on Digital Libraries, 2000.

[2] Brin, S., "Extracting patterns and relations from the world wide web", WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT '1998.

[3] Minlie Huang, Xiaoyan Zhu, Shilin Ding, Hao Yu and Ming Li, "ONBIRES: Ontology-based Biological Relation Extraction System", In Proceedings of the Fourth Asia Pacific Bioinformatics Conference, 2006.

[4] Liu, Y., Shi, Z., & Sarkar, A. (2007). Exploiting rich syntactic information for relationship extraction from biomedical articles. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers (pp. 97–100). Rochester, New York: Association for Computational Linguistics.

[5] D. Fisher, S. Soderland, J. McCarthy, F. Feng, and W. Lehnert. Description of the UMass systems as used for MUC-6. In Proceedings of the 6th Message Understanding Conference. Columbia, MD, 1995.

[6] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "WebScale Information Extraction in KnowItAll", ACM 1-58113-844-X/04/0005, May, 2004, New York, USA.

[7] Hasegawa, T., Sekine, S. and Grishman, R. "Discovering relations among named entities from large corpora," in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ser. ACL '04. Stroudsburg, PA, USA, (2004).

[8] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods", In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pages 189–196. Cambridge, MA, 1995.

[9] Craven M, Kumlien J, Constructing Biological Knowledge Bases by Extracting Information from Text Sources, Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology 1999.

[10] Abulaish M, Dey L,"Biological relation extraction and query answering from medline abstracts using ontology-based text mining", Data Knowl Eng, 2007;61(2):228–62.

[11] Frunza, Oana, Diana Inkpen, and Thomas Tran. "A machine learning approach for identifying disease-treatment relations in short texts." Knowledge and Data Engineering, IEEE Transactions on 23.6 (2011): 801-814.

[12] Wen-Juan Hou, Hsiao-Yuan Chen, "Rule extraction in gene–disease relationship discovery", Gene 518 (2013) 132–138.

[13] Ning Kang, Bharat Singh, Chinh Bui, Zubair Afzal, Erik M van Mulligen and Jan A Kors, "Knowledge-based extraction of adverse drug events from biomedical text." BMC bioinformatics 15.1 (2014): 64.

[14] Jump up, Carvalko, J.R., Preston K, On Determining Optimum Simple Golay Marking Transforms for Binary Image Processing, IEEE Transactions on Computers 21: 1430–33.doi:10.1109/T-C.1972.223519,1972.

[15] Quan, Changqin, Meng Wang, and Fuji Ren. "An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature." PloS one 9.7 (2014): e102039.

[16] Carlson, Andrew, et al., 2010. Toward an architecture for never ending language learning. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, pp. 1306–1313.

[17] Chao Chen, Liang He, Xin Lin, " REV: Extracting Entity Relations from World Wide Web", In proceeding of ACM, 978-1-4503-1172-4, ICUIMC'12, February 20–22, 2012, Kuala Lumpur, Malaysia.

[18] Xu, Rong, and QuanQiu Wang. "A semi-supervised approach to extract pharmacogenomics-specific drug–gene pairs from biomedical literature for personalized medicine." Journal of biomedical informatics 46.4 (2013): 585-593.

[19] Xu, Rong, Li Li, and QuanQiu Wang. "Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature." Bioinformatics 29.17 (2013): 2186-2194.

[20] Xu, Rong, Li Li, and QuanQiu Wang. "dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text." BMC bioinformatics 15.1 (2014): 105.

[21] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. IJCAI '07: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India.