ONBIRES: ONtology-based BIological Relation Extraction System

Minlie Huang[1], *Xiaoyan Zhu[1], Shilin Ding[1], Hao Yu[1] and Ming Li[1,2]

[1]*State Key Laboratory of Intelligent Technology and Systems (LITS), Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China, {huangml00,th,dsl05}@mails.tsinghua.edu.cn*

[2]*Bioinformatics Laboratory, School of Computer Science, University of Waterloo, N2L 3G1, Ontario, Canada, mli@uwaterloo.ca*

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. In this paper, we present an ontology-based framework of biological relation extraction system. This framework is unified and able to extract several kinds of relations such as gene-disease, gene-gene, and protein-protein interactions etc. The main contributions of this paper are that we propose a two-level pattern learning algorithm, and organize patterns hierarchically.

1. Introduction

Biological data, including both experimental data and textual information, are growing tremendously in these decades. However, most of important biological knowledge, such as protein-protein interaction and gene-disease interaction, is still locked in a large number of literatures, remaining not computer-readable. The heaven burden of accessing, extracting and retrieving biological knowledge of interests is left to the human user. To expedite the process of functional bioinformatics, it is absolutely important to develop information extraction systems to automatically process these online biological documents and extract biological knowledge such as protein-protein interaction (PPI), gene-disease correlation, sub-cellular location of protein and so on. A number of database, for example, DIP for PPI [1], KEGG for biological pathways [2], BIND for molecular interactions [3], accumulate such relations.

The portability is another major problem that impedes the wide use of IE tools in online biological documents. Some systems are aimed to extract PPIs [4,5,6], some are designed to mine gene-disease relation, some are able to discover gene-function correlation [7], but none of them can extract these kinds of relations in a unified framework. In other words, it is not easy or unable to adopt these systems from this kind of relation extraction to another one. Most of the approaches are more focused on a specific application to solve a specific kind of problem.

Ontology is a formal conceptualization of a particular domain that is shared by a group of people [8]. Each concept in an ontology has a canonical and consistent definition, and they are organized in a hierarchical tree, thus knowledge can be easily communicated, shared and reused across applications. In recent decades, a

* Corresponding author: zxy-dcs@tsinghua.edu.cn Tel: 86-10-62796831 Fax: 86-10-62771138

number of biological ontologies have been designed and developed for public usage, including Gene Ontology [9], MeSH [10], and LocusLink [11]. These ontologies provide a controlled vocabulary or conceptualization for biological concepts such as gene, protein, disease and function etc, and thus supply a shared understanding of knowledge among biology communities. When an IE system is structured in an ontology-style way, it is more portable and less dependent on applications.

In this paper, we propose an ontology-based biological relation extraction system to automatically extract biological relations from a huge number of online MEDLINE abstracts. Compared with the previous methods, the main contributions of our method are:

1) External ontology integration. Currently, we have integrated four external ontologies, including GO, MeSH, LocusLink, OMIM [12]. Concepts in these ontologies have been converted into a uniform format, and each concept is described by a set of synonymous terms (i.e. *synset*);

2) Ontology-based semantic annotation of online biological documents. Our method will recognize and identify several categories of biological entities, including GENE, PROTEIN, DISEASE, PROCESS, FUNCTION, CELLULAR COMPONENT (CELLC);

3) Two-level pattern learning, i.e., token pattern learning and syntactic pattern learning. We organize patterns in a hierarchy and then a weighted pattern matching scheme is applied.

The rest of the paper is organized as follows: in Section 2, we present an overview of the architecture of ONBIRES. In Section 3, we state that how external ontologies are integrated into our system and how concepts are organized uniformly. In Section 4, a pattern hierarchy is introduced, followed by the detailed pattern learning algorithm in Section 5. Then, experiments and evaluations are shown in Section 6. At last, we make our conclusion in Section 7.

2. Architecture of ONBIRES

A large number of methods have been proposed and various systems are developed to extract biological knowledge from biological literature such as extracting protein-protein interactions, or integrated systems as [13]. However, most systems do not provide a unified framework and most algorithms are heavily dependent on a specific application. Also there is a lack of a mechanism for automatic learning of pattern for such information extraction tasks.

We proposed a novel framework that can extract several kinds of relations with a mechanism of automatic pattern learning. Our algorithm is much less dependent on a specific problem to be solved. It is able to learn patterns and extract relations in a unified way. The system architecture is shown in Figure 1. Compared with previous methods and systems, our approach has several significant advantages. First, we utilize several external ontologies to try to capture as many synonyms as possible for each type of biological entities, and organize them in a uniform format. Second, a hierarchical pattern structure is introduced, on which a weighted pattern matching scheme is used to balance precision and coverage.

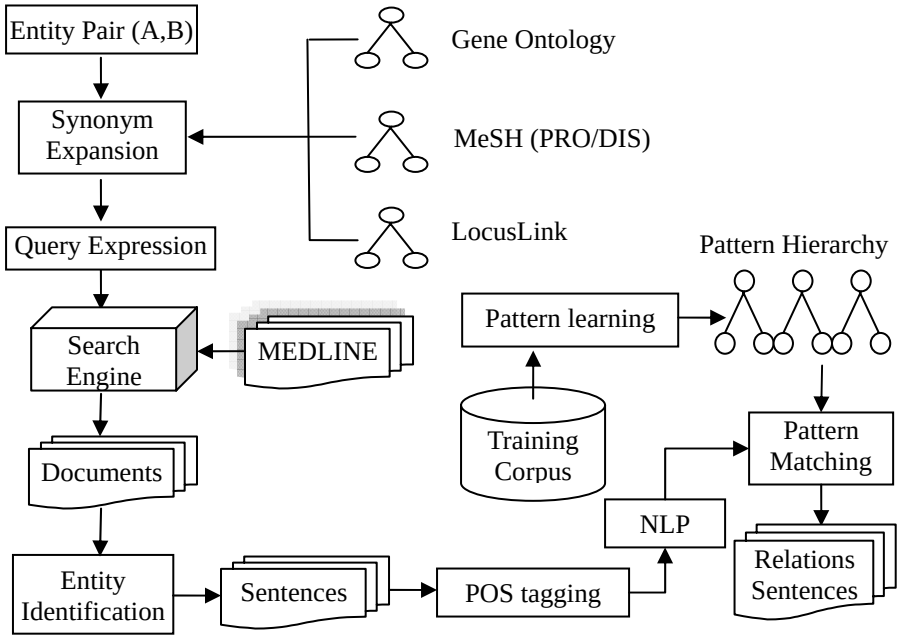

Figure 1. ONBIRES architecture.

There are several steps in our system as follows:

1) For each entity pair (A,B), we will search synonyms in our local ontologies for A and B, respectively. The set of synonyms are later called a synset. The semantic type of each entity is also returned. If no synonym is found, the user has to specify a semantic type.

2) According to the synsets of A and B, a query expression is formed. If the two synset sets are $A=\{a_1,a_2,\ldots a_m\}$ and $B=\{b_1,b_2,\ldots b_n\}$, the query expression is "(a_1 OR a_2 OR $\ldots a_m$) AND (b_1 OR b_2 OR $\ldots b_n$)". This expression is input into a search engine to retrieve MEDLINE documents (currently only abstracts).

3) For each document, we will do semantic tagging using the synsets of entity A and B. Then, documents are segmented into sentences and we only save those sentences that contain both A and B for the further processing.

4) Sentences are part-of-speech (POS) tagged. At the training stage, patterns are learned from a training corpus, whose sentences have been labeled as positive or negative. At the matching stage, sentences are processed by a natural language processing (NLP) module. We have several shallow parsing techniques in the NLP module as describe in [14].

5) A weighted pattern matching algorithm is applied against the pattern hierarchy. Sentences whose matching scores exceed a threshold are declared to have relations.

6) Both the extracted relations and relevant documents are presented to the user in a user interface. We provide PMID, title and abstract of a relevant document.

3. External Ontology Integration

Biological named entity recognition is a great challenge for IE communities [15]. A number of methods, such as machine learning based ones [16], have been devised to improve the performance, but they are still far away from real applications. In our system, we try to collect external ontologies to enhance the results of entity identification. The first one is Gene Ontology, which has been well-known as a controlled vocabulary for gene annotation of documents. This ontology consists of three subjects, that is, biological process, biological function and cellular component. Accordingly, we extract three kinds of entities, that is, PROCESS, FUNCTION, and CELLC, to form our own synset ontology. The number of the three types of entities amounts to 9852, 7576 and 1679, respectively.

The second ontology we used is MeSH (Medical subject Heading). MeSH models a hierarchical terminology of disease, chemical and drug and so on. In this system we only consider two sub-branches of the hierarchical tree, that is, the disease branch (labeled with C##.###, each '#' is a digit), and the protein branch (labeled with D12.776.###). Totally, we obtain 1610 proteins and 226 diseases from MeSH. 1,303,625 genes are extracted from LocusLink and 9315 from OMIM and another 3125 diseases are obtained from OMIM.

Finally, these data are organized uniformly in the format as shown in Table 1. *UID* is the unique identity of an entity, where this identifier is directly reproduced from the original ontology thus we could search the ontology via this symbol. Synset is a set of terms describing the same entity. Six entity types are defined, that is, PROCESS, FUNCTION, CELLC, PROTEIN, GENE, and DISEASE.

Table 1. Uniform concept format and examples. Synonyms are separated with '#'.

UID	Entity Description	Entity Type	Source	Synset
D12.776.503	Lectins	PROTEIN	MeSH	Animal Lectins# Isolectins#
U_GO:0050285	sinapine esterase activity	FUNCTION	GO	sinapine esterase activity#

4. Pattern Hierarchy

We have defined a pattern hierarchy according to the generalization power of each pattern. An example of the hierarchy is shown in Figure 2.

Syntactic pattern consists of a sequence of part-of-speech tags. This kind of pattern reveals the syntactic constraints that a pattern must conform to. Syntactic patterns are learned by aligning sequences of part-of-speech tags from token patterns.

Token pattern comprises keywords that are commonly used to describe relations. And many token patterns may share the same form of syntactic constraints. They have less generalization power than syntactic patterns, and at the same time, they are more precise. Token patterns are generated by aligning sequences of words from instance patterns.

Instance pattern is a sentence which has been labeled as positive. Token pattern can be learned from positive samples.

We note that the generalization power of a pattern decreases from the top to the bottom along the hierarchy, and the accuracy increases. With a weighted pattern matching scheme at different levels, we could obtain a balance between the accuracy and extensibility. This is the major motivation why we organize patterns hierarchically.

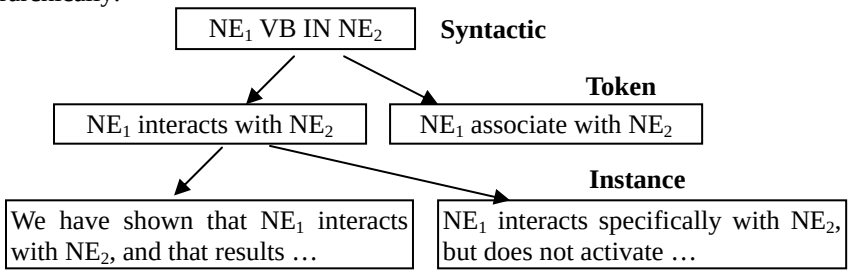

Figure 2. Pattern Hierarchy

5. Automatic pattern learning

The idea of using dynamic programming to automatically learn patterns is used by [5,7,14,17]. The major contribution of our method is that we use a two-level pattern learning algorithm and organize patterns hierarchically, and furthermore we adopt a weighted pattern matching scheme on the structure. We generate patterns at a token level and syntactic level. At each level, sequence algorithm alignment is used to generate patterns. The pattern structure used in our system is Eliza-style [18]. A pattern will be represented in a 5-tuple: <*prefiller*, *NE₁*, *midfiller*, *NE₂*, *postfiller*>, where *NE₁* and *NE₂* are two entities concerned with a specific application. *Prefiller* is a pattern element before the entity *NE₁*, *midfiller* is a pattern element between *NE₁* and *NE₂*, and *postfiller* is a pattern element after *NE₂*. They are all lists of words or tags. For instance, given a sentence as "We/NNP found/VBD that/IN NEDD8/PROTEIN modifies/VBZ CUL1/PROTEIN in/IN Drosophila/NN.", the algorithm may learn a token pattern {"", $PROTEIN_1$, modifies, $PROTEIN_2$, ""}, and a syntactic pattern {"", $PROTEIN_1$, VBZ, $PROTEIN_2$,""}. The sentence itself is an instance pattern (may be positive or negative). A sentence is also represented in a similar 5-tuple.

It is well known that local alignment is a dynamic programming algorithm as formula (1a-b).

$$sim(i,0) = sim(0,j) = 0; i = 1,2,...,M,\ j = 1,2,...,N \quad (1a)$$

$$sim(A_{1,2,...,i}, B_{1,2,...,j}) = \max \begin{cases} 0 \\ sim(A_{1,2,...,i-1}, B_{1,2,...,j-1}) + s(a_i, b_j), if\ a_i = b_j \\ sim(A_{1,2,...,i-1}, B_{1,2,...,j}) + s(a_i, GAP), \\ sim(A_{1,2,...,i}, B_{1,2,...,j-1}) + s(b_j, GAP), \end{cases} \quad (1b)$$

During token pattern learning, we take $s(w,w)$ =1 and $s(w_1,w_2)=-1, w_1 \neq w_2$, which means that if two words share the same base form, the score is 1, otherwise the score will be -1. Therefore only those words that have the same base form can be aligned together.

During syntactic pattern learning, the local alignment algorithm is applied again on sequences of part-of-speech tags of token patterns. The scores *s(a,b)* is adopted from [5].

In our pattern learning algorithm, we use a pattern frequency to record how many times each pattern is aligned during the pairwise alignment. Those whose frequencies are less than a user-specified threshold are removed from the pattern set.

When a pattern hierarchy is obtained, a weighted pattern matching scheme is used. The matching score for a sentence is defined in formula (2):

$$Score(S_j) = \arg\max_{P_{tok}} \{w_{tok} * Sim(P_{tok}, S_j) + w_{syn} * Sim(P_{syn}(P_{tok}), S_j)\} \quad (2)$$

where P_{tok} is a token pattern, $P_{syn}(P_{tok})$ is the syntactic pattern of P_{tok}. w_{tok} is the weight for token pattern, and w_{syn} for syntactic pattern. When this score exceeds a user-specified threshold, we can say definitely that this sentence describes a relation. For those sentences that have more than two entities, all possible combinations of two entities are considered. Since syntactic patterns are much less precise than token patterns, w_{tok} is set to be larger than w_{syn}. We also apply other constraints on syntactic patterns. For example, if two words match in syntactic patterns, but do not match in token patterns, the semantic similarity is computed by using WordNet. The matching score from syntactic pattern is added to the overall score only when the similarity is larger than a threshold (0.7 currently).

The reason why we have to weight between different levels derives from this fact: if we only consider one level of patterns, either the matching precision is quite low, or the coverage is narrow. For example, if we have two patterns, as shown in Figure 2, a token pattern {“”, NE_1,”interacts with”, NE_2,””} and a syntactic pattern {“”, NE_1,VB IN, NE_2,””}, for a sentence “…NE_1 associates with NE_2…”, there is no match at the token level, while it can be matched at a syntactic level. Similar cases are also observed for the problem of low-precision.

6. Experiments

Evaluating the precision and recall of ONBIRES is very difficult because a huge collection of online MEDLINE abstracts is involved. For a small number of documents, it is possible to annotate them manually and compute the precision and recall. In the current version of our system, we evaluate our approach on two applications, i.e. gene-disease interactions, and protein-protein interactions.

The first experiment is to extract protein-protein interactions. We collect the training corpus from http://www.biostat.wisc.edu/~craven/ie/ [19]. Each sentence is annotated as either a negative sample or positive sample. Positive samples are labeled with relation tuples which were gathered from the MIPS Comprehensive Yeast Genome Database. We used 1102 positive samples to generate patterns. 1024

sentences from GENIA corpus are used for evaluation [20]. GENIA corpus is available at: http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/. These sentences are manually annotated by experts, where there are 238 positive samples.

The second experiment is to extract gene-disease correlations. This corpora is also downloaded from http://www.biostat.wisc.edu/~craven/ie/. The relation tuples were gathered from the Online Mendelian Inheritance in Man (OMIM) database. There are 636 positive samples in this corpus, which are all used for learning patterns. Since the corpus is comparatively small, 100 of the training samples and another 177 negative samples are randomly selected for evaluation.

In each experiment, we compare the performance of token patterns and that of token patterns plus syntactic patterns. These results are shown in Figure 3 and Figure 4. During pattern learning, we provide a vocabulary to restrict which words can be contained by a pattern. Patterns whose frequencies are less than one are removed. The statistics of extracted interactions in these experiments are listed in Table 2.

From these results, we could see that token patterns plus syntactic patterns outperform only token patterns. The two curves converge to the same curve when the threshold becomes larger because sentences with large matching scores have matched token patterns perfectly, and syntactic patterns have tiny contributions to these sentences. With a smaller threshold, the performance is improved remarkably when syntactic patterns are used.

We also investigated into those sentences that can not be extracted correctly. There are three kinds of errors:

1) Incorrect patterns. Although we have limited the vocabulary of patterns, and have removed patterns with low frequencies, a small proportion of incorrect patterns are still left. Unfortunately, they have a fairly high frequency. Therefore, more sophisticated techniques need be developed to assess each pattern.

2) Errors caused by complicated grammatical structures. This method treats a sentence as a linear sequence, thus it is not competent to process complicated grammatical structures. Although we have done long sentence splitting, appositive and coordinative structure recognition as shown in [14], there are more structures that can not be handled. For example, there is a sentence:

The oxygen radical scavenger N-acetyl-L-cysteine, but not an inhibitor of nitric oxide synthase, inhibited LIF -induced HIV replication. where underlined parts are identified as proteins. It matches a pattern "PROTEIN inhibit PROTEIN", which is erroneous.

3) Errors caused by named entity identification. In our system, we used a dictionary-based method to recognize named entities. However, in many cases, this method produces errors. Particularly, it can not discriminate proteins from genes, since most genes and proteins have the same lexical symbols. An example is listed here:

Taken together, our data indicate that MS-2 mediates induction of the CD11b gene as cells of the monocytic lineage mature. The underlined terms are identified as proteins, but the second should be recognized as a gene.

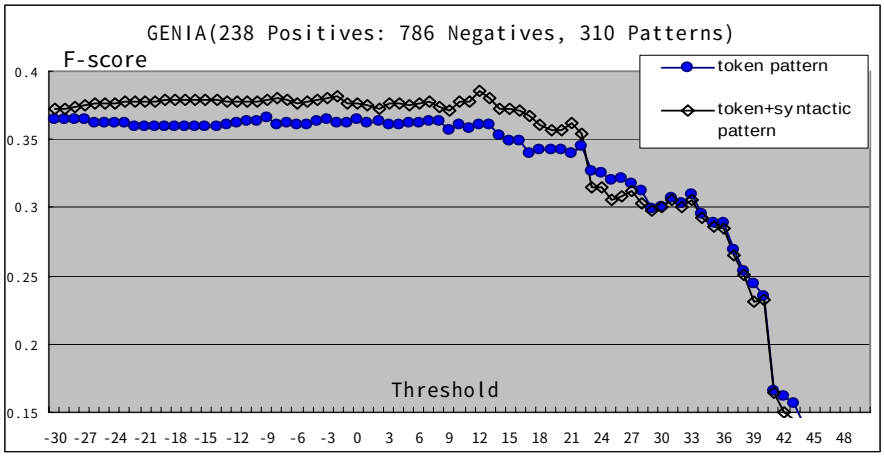

Figure 3. F-Score curve over matching score threshold for GENIA corpus.

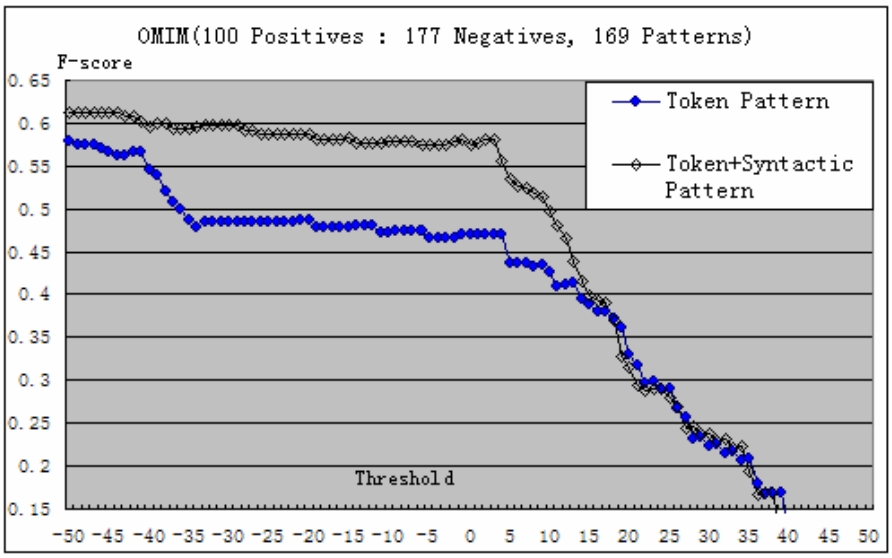

Figure 4. F-Score curve over matching score threshold for OMIM corpus.

Table 2. Statistics of extracted interactions with the best threshold. TP indicates the number of correct samples. ET denotes the number of extracted samples.

Corpus	Experiment	Precision	Recall	F-score	TP	ET
GENIA	Token pattern	0.424	0.235	0.302	56	132
	Token+syntactic	0.411	0.243	0.306	58	141
OMIM	Token pattern	0.532	0.42	0.469	42	79
	Token+syntactic	0.58	0.58	0.58	58	100

7. Conclusion

In this paper, we have proposed an ontology-based information extraction system to search biological relations from online documents. This system, which is ontology-based, has a unified framework and is less dependent on specific applications. Several external ontologies are integrated to improve the structure and organization of concept. A two-level pattern learning algorithm is applied to generate patterns which are then organized in a hierarchy. A weighted matching scheme is devised to balance the accuracy and coverage of the system.

The experimental results show that our system is promising to extract knowledge from a huge number of MEDLINE abstracts. Future work will be focused on how to evaluate patterns more efficiently, process complicated grammatical structures and handle named entity recognition errors.

Acknowledgments

The work was supported by Chinese Natural Science Foundation under grant No. 60272019 and 60321002, the Canadian NSERC grant OGP0046506, CRC Chair fund, and the Killam Fellowship. We also would like to thank Xiaozhe Li and Zhiyuan Liu for coding programs and converting data from several publicly available external resources.

References

1. I. Xenarios, E. Fernandez, L. Salwinski, X.J. Duan, M.J. Thompson, E.M. Marcotte and D. Eisenberg. (2001) DIP: The Database of Interacting Proteins: 2001 update, Nucleic Acids Reg., 29, pp. 239–241.
2. M. Kanehisa and S. Goto. (1997) A systematic analysis of gene functions by the metabolic pathway database. In "Theoretical and Computational Methods in Genome Research" (Suhai, S., ed.), pp. 41–55, Plenum Press
3. G.D. Bader, I. Donaldson, C. Wolting, B.F. Quellette, T. Pawson and C.W. Hogue. (2001) BIND –The Biomolecular Interaction Network Database, Nucleic Acids Research, 29(1), pp. 242–245.
4. T. Ono, H. Hishigaki, A. Tanigami and T. Takagi. (2001) Automated extraction of information on protein–protein interactions from the biological literature, Bioinformatics, 17(2), pp. 155–161.
5. M.L. Huang, X.Y. Zhu, Y. Hao, D.G. Payan, K. Qu and M. Li. Discovering patterns to extract protein-protein interactions from full-texts. Bioinformatics, Dec, 2004; 20(18):3604-3612.
6. E.M. Marcott, I. Xenarios and D. Eisenberg. (2001) Mining literature for protein–protein interactions, *Bioinformatics*, 17(4), pp. 359–363.
7. J.H. Chiang and H.H. Yu. (2003) MeKE: discovering the functions of gene products from biomedical literature via sentence alignment, *Bioinformatics*, 19(11), pp. 1417–1422.

8. T.R. Gruber. “A Translation Approach to Portable Ontology Specifications,” Knowledge Acquisition, vol. 5, pp. 199–220, 1993.
9. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nat. Genet., 25, 25–29. http://www.gene ontology.org/.
10. MeSH: Medical Subject Heading. http://www.nlm.nih.gov/mesh/meshhome.html.
11. K.D. Pruitt and D.R. Maglott. (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res., 29, 137–140. http://www.ncbi. nlm.nih.gov/LocusLink/.
12. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000. http://www.ncbi. nlm.nih.gov/omim/.
13. X.H. Hu, T.Y. Lin, I.Y. Song, X. Lin, I. Yoo, M. Lechner, M. Song. Ontology-Based Scalable and Portable Information Extraction System to Extract Biological Knowledge from Huge Collection of Biomedical Web Documents. Web Intelligence 2004: 77-83.
14. M.L. Huang, X.Y. Zhu, and M. Li. A hybrid method for relation extraction from biomedical literature. 2005, accepted by International Journal of Medical Informatics.
15. L. Hirschman, J.C. Park, J. Tsujii, L. Wong and C.H. Wu. Accomplishments and challenges in literature data mining for biology. Bioinformatics, 18, 1553-1561, December 2002.
16. G.D. Zhou, J. Zhang, J. Su, D. Shen, and C.L. Tan. “Recognizing names in biomedical texts: a machine learning approach”. Bioinformatics Vol. 20 no. 7. 2004, pages 1178-1190.
17. E. Agichtein and L. Gravano. Snowball: extracting relations from large plain-text collections. ACM DL 2000: 85-94.
18. J. Weizenbaum. (1966) ELIZA – A Computer program for the study of natural language communications between men and machine, Communications of the Association for Computing Machinery, 9, pp. 36–45.
19. S. Ray and M. Craven. Representing Sentence Structure in Hidden Markov Models for Information Extraction. IJCAI 2001: 1273-1279. Seattle, USA.
20. J.D. Kim, T. Ohta, Y. Teteisi and J. Tsujii. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics. 19(suppl. 1). pp. i180-i182. Oxford University Press.