A Logic Approach to Mine Syntactic Fragments of Mesh Keywords in Biomedical Literature

Annalisa Appice[1,2]

[1]Dipartimento di Informatica, Università degli Studi di Bari
via Orabona, 4 - 70126 Bari - Italy
[2]CILA (Centro Interdipartimentale per la ricerca in Logica e Applicazioni)
appice@di.uniba.it

Abstract. The discovery of new and potentially meaningful dependencies between biomedical keywords in literature can take great advantage from the application of a logic approach to text mining. This is motivated by the peculiarity of logic approaches to be able to express and manipulate relations between objects. We investigate the application of a logic approach to the task of frequent pattern discovery in Medline abstracts. Frequent patterns represent frequent syntactic fragments of MeSH keywords. Preliminary experiments with a collection of abstracts obtained by querying Medline on a specific disease are reported.

1 Introduction

The exponential increase in publication rate of new articles in biomedicine makes it difficult for researchers to keep up with research progresses without the help of automatic knowledge discovery techniques. Over 16 million references to biomedical journal articles are currently contained in the Medline collection, the main online resource of biomedical research literature. On the other hand, textual data as Medline articles are generally unstructured and the available resources (e.g., PubMed, the search engine interfacing Medline) do not still provide adequate mechanisms for helping humans in "deeply analyze" very large amount of content. The need to analyze this volume of unstructured data has prompted the use of text mining tools to automatically extract key biological information.

Several successes may be attributed to text mining research in biomedicine and many methods have been presented so far [3, 1, 5, 8]. Most of the text mining methods provide keyword based search functionalities, which are based on the frequencies of surface information such as words and parts of speech (e.g., sentences). These methods ignore more abstract information such as syntactic structures. Although, the surface information of two sentences is similar, the syntactic structures may be completely different. By the way, syntax-based search methods have been proposed in text mining literature. In [13, 4], sentences are parsed according to a user-defined structural pattern and sentences, whose parse trees match the given pattern, are returned. In [10], both a syntax-based search functionality and a keyword based interface are used to retrieve subject-object

relationships between user-defined keywords. The user is not required to edit a structural pattern, but to specify the keywords of interest. Relationships simply express the existence of a verbal interaction between subject and an object without distinguishing among different kind of interactions (e.g., the subject "causes" the object or the subject "inhibits" the object and so on). The main attempts of applying syntax based search methods to biomedical corpus are reported by Page and Craven [12]. Also in his case, the verb is used to isolate potentially meaningful part of speech. In particular, manually-written rules which identify the common verb of expressing protein interactions in natural language are applied to isolate the part-of speech where proteins can be searched.

In this work, we investigate the problem of how to mine an unstructured biomedical text corpus in order to identify any syntactic fragments of biomedical keywords which appear frequently in a training text corpus retrieved with a PubMed query to Medline. We propose a knowledge discovery framework, called BioSOV-FP (BIOlogical Subject-Object(s)Verb based Frequent Pattern discovery system), which automatically annotates the biomedical keywords in the training corpus, integrates these keywords in structured syntactic environments, uses a first-order language to represent these environments and resorts to a logic approach for frequent pattern discovery to identify frequent combination of syntactical dependencies among biomedical keywords. Similarly to the system proposed in [10], BioSOV-FP uses both keyword based interface and syntax-based search functionality. Anyway, BioSOV-FP exhibits fundamental differences with the competitor system. First, BioSOV-FP is not a general purpose system, but it is appositely designed for biomedical domain. Biomedical keywords are automatically annotated in the training corpus as they appear in the MeSH (Medical Subject Headings) taxonomy. Second, the syntactic dependencies between biomedical keywords are 3-structures in the form of subject-object(s)-verb. Differently from most of works in [12], the verb is not used to only extract part of speech of interest, but it is mined in order to discover *how* a subject keyword is frequently related to its object keyword(s). The first-order logic representation of these 3-structures allows to take advantage of the syntactical structure of text corpus according to the assumption that a more knowledge intensive technique is likely to perform better when applied on the tasks of text mining due to expressive power gained through relations [6]. The choice of the expressive power of first-order logic language is further motivated by the fact that it is so close to the human one that it is easy to guess the original sentence underlying this formal description. The first-order logic language allows to exploit a background knowledge such as the MeSH taxonomy, a domain specific knowledge as well as a language bias. Frequent patterns play the role of relevant syntactic knowledge shared by the articles under study. These patterns can be employed to formulate new queries to Medline on the topic of interest.

The paper is organized as follows. Section 2 gives an overview of BioSOV-FP framework. Section 3 describes the extraction of syntactic structures from a biomedical text corpus. Section 4 presents the discovery of frequent syntactic fragments of MeSH keywords. Finally, a preliminary application is reported.

2 An overview of BIOSOV-FP Framework

The framework of BIOSOV-FP is reported in Figure 2. A biomedical text corpus is poured into the BioSOV-FP framework and the knowledge discovery process is triggered. Initially, the unstructured text corpus is allowed to express itself in a structured format. This transformation is in charge of a four-stepped job which includes acronym processing, stopword removing, syntactic structure extraction and MeSH keywords annotation. Transformation job uses acronym dictionary, stopword list and MeSH term taxonomy as input. The structure is added to the training corpus in the form of the subject-objects(s)-verb tuples. MeSH keywords are annotated in the nominal components (subject and object(s)) of these environments. The MeSH taxonomy allows the representation and management of MeSH keywords at different levels of granularity (from the most general, at the top level of taxonomic, to the most specific, at the bottom level of taxonomy). The subject-objects(s)-verb tuples which express verbal dependencies between MeSH keywords are represented in a first-order language and stored in the extensional part of a Datalog database. Domain specific knowledge is stored as a set of rules in the intensional part of the Datalog database to support qualitative reasoning. A language bias is used to specify constraint specifications for interesting patterns. Datalog database and language bias are passed down to the relational frequent pattern discovery module. The minimum support $\sigma[l]$ is given for each level of MeSH taxonomy before frequent pattern discovery starts.

Details on how the text corpus is transformed from the original un-structured format to the structured format and how the relational frequent pattern discovery is performed are reported in the next Sections.

3 Extracting Syntactic Structures of MeSH Keywords

The text processor identifies verbal components and nominal components in the text, assigns each nominal component with either the role of subject or the role of object of the corresponding verbal component, and annotates the MeSH keywords in the nominal components. In this way, syntactic 3-structures, that is, subject-object(s)-verb structures, are extracted. For each subject-object(s)-verb structure, the subject is the modifier of the dependency expressed by the verb, while the object(s) are modifee(s) of the same verbal dependency. The syntactic 3-structures whose nominal components include MeSH keywords are output as a structured representation of the text and passed down to the frequent pattern discovery module together with the MeSH taxonomy. Details of steps of transformation job are described below.

Acronym processing. A dictionary based approach is used to identify the occurrences of biomedical acronyms in the training text corpus. Acronyms are replaced by the expanded definitions as in the dictionary. This is essential to avoid part-of-speech tagging errors due to multi-worded nouns.

Stopword removing. Stopwords (e.g., article such as "the" or modal verbs such as "can") reported in the stopword list are removed from the text.

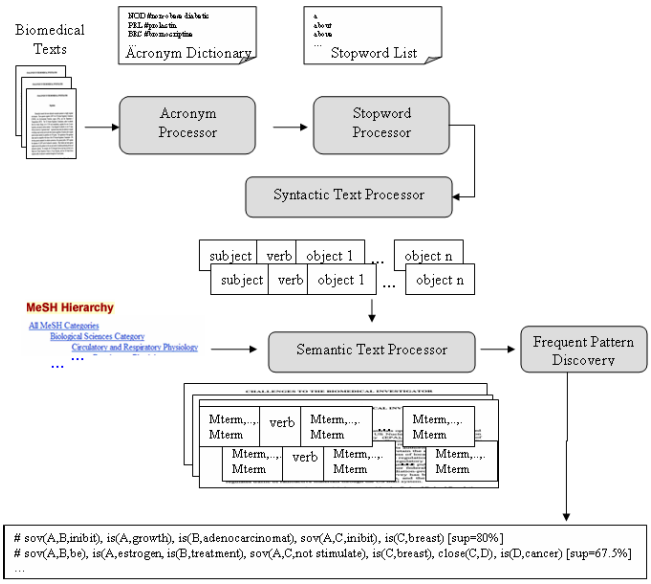


Fig. 1. BIOSOV-FP framework.

Subject-Object(s)-Verb extraction. The text is analyzed by means of the generic text processor for English corpora, called MontyLingua [9]. Montylingua provides the functionalities to tokenize, part-of-speech tag, chunk, stem texts and transform corpora into a set of subject-object(s)-verb (SOV) 3-structures.

Definition 1. *A subject-object(s)-verb (SOV) structure is defined as the extended BNF grammar in the followings.*

$\langle$ SOV $\rangle$ ::= {$\langle$SUBJECT$\rangle$ { $\langle$OBJECT$\rangle$ } $\langle$ VERB$\rangle$}
$\langle$ SUBJECT $\rangle$ ::= { $\langle$ TERM $\rangle$ } $\langle$ OBJECT $\rangle$ ::= { $\langle$ TERM $\rangle$ }
$\langle$ VERB $\rangle$ ::= $\langle$ INFINITIVE_FORM_OF_VERB $\rangle$

The text is firstly tokenized in individual sentences according to a regular expression recognizer of sentence delimiters (e.g., full-stop, ellipse, exclamation mark and question mark, at the end of a word) and then each sentence is atomized in words using white spaces as word delimiter. Common abbreviated words, such as "isn't" are expanded into the words "is" and "not". Each word is tagged by a Brill Tagger [2] which firstly tags each word according to a lexicon which contains the likely tag for each word and then corrects the original tags by using lexical and contextual rules. Some SOVs structures are reported in Example 1.

Example 1. Let us consider a fragment of a Medline article entitled "Hormone replacement therapy: the perspectives for the 21st century", that is:

"*Progestogens* can *modify* the *cellular response* of *normal* as well as *cancer breasts.* The *possible protective effect* of *continuous progestogen addition* is very *interesting* and *needs* further *investigation.*"

The text without stopwords is in italics. SOV structures extracted by MontyLingua from the text in italics are listed below:

$$\text{SOV}=[\underbrace{\langle progestogen\rangle}_{subject}, [\underbrace{\langle cell, response, normal\rangle}_{object}, \underbrace{\langle cancer, breast\rangle}_{object}], \underbrace{modify}_{verb}]$$

$$\text{SOV}=[\underbrace{\langle continuous, progestogen, addition\rangle}_{subject}, [\underbrace{\langle interest\rangle}_{object}], \underbrace{be}_{verb}]$$

$$\text{SOV}=[\underbrace{\langle continuous, progestogen, addition\rangle}_{subject}, [\underbrace{\langle investigation\rangle}_{object}], \underbrace{need}_{verb}]$$

MeSH keywords annotation. The BioTeKS Text Analysis Engine provided within the IBM UIM Architecture [7] is used to annotate the MeSH keywords which occur in the nominal components of the SOV structures. We use the "canonical" form of each MeSH term, which is available in the MeSH taxonomy. Terms without the MeSH tag are removed from the corresponding nominal environments of the SOV structure. A definition of a MeSH annotated subject-object(s)-verb structure (MeSH SOV) is reported in Definition 2.

Definition 2. *A MeSH annotated subject-object(s)-verb (MeSH SOV) structure is defined as the extended BNF grammar in the followings.*

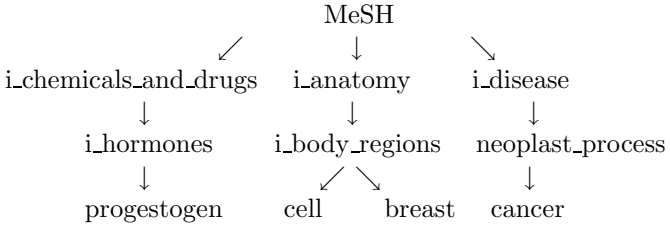
⟨ MeSH SOV ⟩ ::= {⟨SUBJECT⟩ { ⟨OBJECT⟩ } ⟨ VERB⟩}
⟨ SUBJECT ⟩ ::= { ⟨ MeSHTerm } ⟩ ⟨ OBJECT ⟩ ::= { ⟨ MeSHTerm } ⟩
⟨ VERB ⟩ ::= ⟨ INFINITIVE_FORM_OF_VERB ⟩

An example of MeSH SOV structure is reported in Example 1.

Example 2. Let us consider the SOV structure reported in Example 1:

[⟨*progestogen*⟩, [⟨*cell*, response, normal⟩ ⟨*cancer*, *breast*⟩], *modify*]

MeSH keywords (in italics) are annotated as leaves of MeSH taxonomy, that is,



Terms which are not annotated as MeSH keywords are removed from the nominal components, thus the following MeSH SOV structure is constructud:

[⟨*progestogen*⟩, [⟨*cell*⟩ ⟨*cancer*, *breast*⟩], *modify*].

4 Frequent Pattern Discovery

The frequent pattern discovery is in charge of SPADA [11] which addresses the task of frequent pattern discovery from a Datalog database by dealing with taxonomy knowledge and language bias. In the followings, we present how to store MeSH SOV structures in a Datalog database and to express patterns which

describe syntactic fragments of MeSH keywords. Search strategy adopted by SPADA to discover frequent patterns is also reported.

Text corpus representation. Let t_i be a text in the training corpus T. t_i represents a unit of analysis and it is represented by means of the MeSH SOV structures which are extracted from t_i. Each MeSH SOV structure as well as the hierarchical relations of the MeSH taxonomy which identify the Mesh keywords in the structure are stored as ground atoms in the extensional part D_I of a Datalog database D. Predicates used are the followings:

- **text**(t): t is a text in the corpus;
- **verb**(t, v): t is a text and v is a verb;
- **nominal**(t, n): t is a text and n is a nominal component;
- **mesh**(n, m): n is a nominal component and m is a MeSH term;
- **is_a**($m1, m2$): $m1$ and $m2$ are nodes of the MeSH taxonomy, there is a hierarchical relation between $m1$ and $m2$ in the taxonomy;
- **close**($t, w1, w2$): t is a text, $w1$ and $w2$ are MeSH keywords which appear to be close in the nominal component of t;
- **sbj_vb_obj**($n1, v, n2$): $n1$ (subject) and $n2$ (object) are the nominal components, v is the verb.

An example of the use of these predicates is reported in Example 3.

Example 3. The MeSH SOV structure [$\langle progestogen \rangle$, [$\langle cell \rangle$ $\langle cancer, breast \rangle$], $modify$] in Example 2 is extracted from a text t_1 and stored in D_E as follows:

text(t1). nominal(t1,n1). nominal(t1,n2). nominal(t1,n3).
mesh(n1,progestogen). mesh(n2,cell). mesh(n3,cancer). mesh(n3,breast).
verb(t1,modify). close(t1,cancer, breast).
is_a(progestogen, i_hormones). is_a(cell, i_body_regions).
is_a(breast, i_body_regions). is_a(cancer, i_neoplast_process).
is_a(i_hormones, i_chemical_and_drugs). is_a(i_body_regions, i_anatomy).
is_a(i_neoplast_process, i_disease). is_a(i_chemical_and_drugs, mesh).
is_a(i_anatomy, mesh). is_a(i_disease, mesh).
sbj_vb_obj(n1,modify,n2). sbj_vb_obj(n1,modify,n3).

where "is_a(_,_)" atoms express the hierarchical relations of MeSH taxonomy (BK) (see Example 2) which are used to annotate the MeSH keywords.

The domain knowledge is formulated as a normal logic program which defines the intensional part D_I. This intensional part allows deductions to be made (i.e. concluding additional atoms) from data stored in D_E. Example 4 shows the normal logic program stored in D_I and deductions made from atoms in D_E.

Example 4. The domain knowledge includes the intensional definition of the predicate "mesh_verb_mesh(_,_)", which expresses a verbal dependency between MeSH keywords (and not the nominal components):

mesh_vb_mesh(T, M1, V, M2) :− text(T, N1), text(T, N2),
sbj_vb_obj(N1, V, N2), mesh(N1, M1), mesh(N2, M2).

By considering the extensional atoms reported in Example 3, this domain knowledge entails the atoms "mesh_vb_mesh(t1, progestogen, modify, cell)",

"mesh_vb_mesh(t1, progestogen, modify, breast)", "mesh_vb_mesh(t1, progestogen, modify, cancer)" and "mesh_vb_mesh(t1, cell, modify, cancer)".

Pattern representation. Relational patterns to express syntactic fragments of MeSH terms at a level l of the MeSH taxonomy are atomsets in the form "$text(T), \mu_l(T)$ [s]" where text(T) is the atom that identifies each single text in the training text corpus, while $\mu_l(T)$ is a conjunction of Datalog atoms which provide a description of a fragment of the text T at the level l of the MeSH taxonomy. Each atom in $\mu_l(T)$ describes either an extensionally/intensionally defined predicate in D. is_a atoms map each MeSH keyword with the granularity level l of the MeSH taxonomy. The support s estimates the probability p($\{text(T), \mu_l(T)\}$) on D. This means that s of training texts matches the syntactic fragment $\{text(T), \mu_l(T)\}$, that is, a substitution $\theta = T \leftarrow t$ exists such that $\{text(T), \mu_l(T)\}\theta \subseteq D$. The support of a pattern depends on the granularity level l. To be more precise, a pattern $P[s]$ is frequent at level l if $s \geq \sigma[l]$ and all ancestors of P with respect to the MeSH taxonomy are frequent at their corresponding levels. The definition of ancestor relation adopted in this work is based on the MeSH taxonomy as reported in Definition 3.

Definition 3. *A pattern P at granularity level l of the MeSH taxonomy is an ancestor of the pattern P' at granularity level l' with $l' < l$, if P' can be obtained from P by replacing each variable X representing a Mesh keyword at the granularity level l with a variable X' which is more specific than X in the taxonomy and is mapped into the granularity level l'.*

Relational patterns which are related according to the ancestor relation in MeSH taxonomy are reported in Example 5.

Example 5. A possible top-level relational pattern is in the form:

P1: text(T), mesh_vb_mesh(T, M1, reveal, M2),
is_a(M1, mesh), is_a(M2, mesh)

where both M1 and M2 are mapped with "mesh" that is the root of the taxonomy. By descending one level of taxonomy (from more general term to more specific terms), we can find the relational pattern

P2: text(T), mesh_vb_mesh(T, M1,reveal,M2),
is_a(M1, i_anatomy), is_a(M2, i_disease)

where M1 is mapped with "i_anatomy", while M2 is mapped with "i_disease" and "mesh" is hierarchically related to both of them in the taxonomy. By descending to the bottom level of the MeSH taxonomy, M1 and M2 are mapped with the leaf nodes which contain the MeSH keywords as they are annotated in the text,

P3: text(T), mesh_vb_mesh(T, M1, reveal, M2),
is_a(M2, breast), is_a(M2, cancer).

P1 is ancestor of P2 and P3, P2 is ancestor of P3.

Discovery. The set of ground atoms in D_E is partitioned into a number of non-intersecting units of analysis, that is, subsets $D[t]$ each of which includes atoms which describe the nominal components, verbal components, MeSH keywords

which belong to a text t of the training corpus. SPADA mines frequent patterns across the different units of analysis by performing both an intra-level search and an inter-level search. Intra-level search is performed in the space of patterns where the is_a atoms refer to MeSH keywords defined at the same level of MeSH taxonomy. Pattern space is ordered according to the θ-subsumption generality order between patterns. In the inter-level search, SPADA takes advantage of statistics computed at a level l when it searches in the space of more specific MeSH keywords at level $l + 1$. By descending through the MeSH taxonomy it is possible to view the same term levels of abstraction (or granularity) and discover patterns at different levels of granularity. During the frequent pattern generation, patterns which do not satisfy pattern constraints defined in language bias are filtered out. A detailed description of SPADA can be found in [11].

5 The Application

BioSOV-FP has been preliminary evaluated over a corpus of abstracts of biomedical articles extracted by Medline. The dictionary used for the acronym processing is defined by a domain expert. The PubMed query "Alzheimer Drug Treatment Response" is formulated by biomedical researchers. The abstracts of twenty-five articles randomly selected in the retrieved article set are used as a training biomedical text corpus for BioSOV-FP. By processing the text corpus, BioSOV-FP identifies 1957 nominal components (on average 78.28 per abstract) and 834 verbal components (on average 33.36 perr abstract) for a total of 834 SOV structures. By annotating the MeSH keywords, 164 MeSH SOV structures are extracted and stored in a Datalog database for a total of 25 "text(_)" atoms, 241 "verb(_,_)" atoms, 122 "close(_,_,_)" atoms, 311 "nominal(_,_)" atoms, 171 "sbj_vb_obj(_,_,_)" atoms, 374 "mesh(_,_)" atoms, 910842 "is_a(_,_)" atoms stored in extensional database and 241 "mesh_vb_mesh(_,_,_,_)" entailed by normal logic program in the intensional database. Multi-level frequent relational patterns are discovered with $\sigma[l] = 0.08$ for each level l of a six-level MeSH taxonomy. The maximum length of a pattern is set to 9. The language bias is used to ask for patterns containing only the atoms "mesh_vb_mesh(_, _)" and "close_to(_, _)".

A syntactic fragment that is discovered at the top level of the MeSH taxonomy is reported below:

P1 [s=0.4]: text(T),
mesh_vb_mesh(T, M1, cause, M2), mesh_vb_mesh(T, M1, cause, M3),
is_a(M1, mesh), is_a(M2, mesh), is_a(M4, mesh), is_a(M3, mesh).

This pattern cannot be considered informative, but descending to the second level of taxonomy we can find patterns whose ancestor is P1, but which provide a deeper insight in the nature of M1, M2 and M3. For example, the pattern:

P2 [s=0.12]: text(T),
mesh_vb_mesh(T, M1, cause, M2), mesh_vb_mesh(T, M1, cause, M3),
is_a(M1, i_biological_sciences), is_a(M2, i_chemicals_and_drugs),
is_a(M3, i_chemicals_and_drugs).

Table 1. Number of patterns and elapsed time (secs) of learning job on an Intel Centrino (1.66 GHz - 2Gb RAM) PC running WinXP.

Text representation	Patterns	Time
MeSH keywords (B1)	78541	15710.985
MeSH subject-object(s) structures (B2)	1720	229.188
MeSH subject-object(s)-verb structures (BioSOV FP)	800	22.109

By descending to the bottom level (level of leaves), we find the pattern P3 such that P2 is ancestor of P3 and M1, M2, and M3 are mapped to the MeSH keywords as they appear in the text corpora.

P3 [s=0.08]: text(T),
mesh_vb_mesh(T, M1, cause, M2), mesh_vb_mesh(T, M1, cause, M3).
is_a(M1, mutat), is_a(M2, protein), is_a(M3, amyloid).

Knowledge revealed by this bottom level pattern is the existence of a causal dependence (confirmed by domain expert) between mutations in Alzheimer disease and amyloid protein. The training corpora which match this syntactic fragment (in italics) are reported below:

"....Presenilin *mutations* have been hypothesised to *cause* Alzheimer disease either by altering *amyloid* precursor *protein* metabolism or ..."

"... PS *mutations cause* the same functional consequence as mutations on *amyloid* precursor *protein* ..."'

For comparison, we consider the multi-level frequent patterns discovered by resorting to either a keyword based representation (B1) or a subject-object based representation (B2) of the text corpus. In the former case, the Boolean representation is adopted in order to represent only the occurrence of MeSH keywords in the text. In this case, patterns discovered as in [1] express similarity of MeSH term based content in the abstracts without considering differences of the grammatical environment. In the latter case, first-order language is used to represent only subject-object dependencies between MeSH keywords as in [10], the verb is not considered. Patterns are discovered by SPADA. They express similarity of MeSH term based content in the abstracts where MeSH terms are provided as part of subject-object dependencies. The number of the discovered patterns and the elapsed time (in secs) of the learning job are reported in Table 1. As expected, results show a great difference in the number (and elapsed time) of MeSH subject-object(s)-verb patterns discovered by BioSOV FP compared with both the number of keyword based frequent patterns and the number of MeSH subject-object(s) based frequent patterns. In general, for each pattern discovered in BioSOV-FP a pattern involving the same keywords is discovered both in B1 and B2. In this study , the following patterns

P4$\leftarrow B1$ [s=0.12]: mutat, protein, amyloid

P5$\leftarrow B2$ [s=0.08]: text(T), mesh_mesh(T, M1, M2), mesh_mesh(T, M1, M3),
is_a(M1, mutat), is_a(M2, protein), is_a(M3, amyloid).

are discovered, but both of them are less informative than P3.

6 Conclusions

We investigate how to combine syntax search functionality and keyword search functionality to detect subject-object(s)-verb structures of MeSH keywords in Medline abstracts. subject-object(s)-verb structures are represented in first-order logic language. Fragments of frequent syntactic structures are mined as relational frequent patterns. Some preliminary results are discussed. As future work we plan to analyze a larger corpus of biomedical texts.

Acknowledgment

This work is partial fulfillment of the research objective of both ATENEO-2009 project "Estrazione, Rappresentazione e Analisi di Dati Complessi" and DM19410 - Laboratorio di Bioinformatica per la Biodiversità Molecolare.

References

1. M. Berardi, M. Lapi, P. Leo, and C. Loglisci. Mining generalized association rules on biomedical literature. In M. Ali and F. Esposito, editors, *IEA/AIE 2005*, volume 3533 of *LNCS*, pages 500–509. Springer, 2005.
2. E. D. Brill. *A corpus-based approach to language learning*. PhD thesis, USA, 1993.
3. J.-H. Chiang, H.-C. Yu, and H.-J. Hsu. Gis: a biomedical text-mining system for gene information discovery. *Bioinformatics*, 20(1):120–121, 2004.
4. S. Corley, M. Corley, F. Keller, M. W. Crocker, and S. Trewin. The gsearch corpus: Finding syntactic structure in unparsed corpora. In *Computers and the Humanities*, pages 35–2, 2001.
5. D. P. A. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–3213, 2004.
6. S. Ferilli, N. Fanizzi, and G. Semeraro. Learning logic models for automated text categorization. In F. Esposito, editor, *AI*IA 2001*, volume 2175 of *LNCS*, pages 81–86. Springer, 2001.
7. D. Ferrucci and A. Lally. Building an example application with the unstructured information management architecture. *IBM System Journal*, 43(3):455–475, 2004.
8. D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey. Improving literature based discovery support by genetic knowledge integration. *Stud Health Technol Inform*, 95:68–73, 2003.
9. L. Hugo. Montylingua: An end-to-end natural language processor with common sense. In *http://web.media.mit.edu/ hugo/montylingua*, 2004.
10. Y. Kato, S. Egawa, S. Matsubara, and Y. Inagaki. English sentence retrieval system based on dependency structure and its evaluation. In P. Pichappan and A. Abraham, editors, *ICDIM 2008*, pages 279–285. IEEE, 2008.
11. F. A. Lisi and D. Malerba. Inducing multi-level association rules from multiple relations. *Machine Learning Journal*, 55(2):175–210, 2004.
12. D. Page and M. Craven. Biological applications of multi-relational data mining. *SIGKDD Explorations*, 5(1):69–79, 2003.
13. P. Resnik and A. Elkiss. The linguist's search engine: an overview. In *ACL 2005*, pages 33–36. Association for Computational Linguistics, 2005.