Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 21-28 recibido 04-05-12 revisado 04-06-12 aceptado 07-06-12

A Framework for Obtaining Structurally Complex Condensed Representations of Document Sets in the Biomedical Domain

Un marco para la obtención de representaciones condensadas estructuralmente complejas de conjuntos de documentos en el dominio biomédico

Yunior Ramírez-Cruz
Center for Pattern Recognition
and Data Mining
Universidad de Oriente
Santiago de Cuba, Cuba
yunior@cerpamid.co.cu

Rafael Berlanga-Llavori
Department of Languages
and Computer Systems
Universitat Jaume I
Castellón de la Plana, Spain
berlanga@lsi.uji.es

Reynaldo Gil-García
Center for Pattern Recognition
and Data Mining
Universidad de Oriente
Santiago de Cuba, Cuba
gil@cerpamid.co.cu

Resumen: En este artículo presentamos un marco para la obtención de representaciones condensadas estructuralmente complejas de conjuntos de documentos, el cual servirá de base para la construcción de resúmenes, la obtención de respuestas para preguntas complejas, etc. Este marco incluye un método para extraer una lista ordenada de hechos, triplos de la forma entidad - relación - entidad, el cual usa patrones de extracción basados en análisis de dependencias y modelos de lenguajes; y métodos para construir un grafo bipartito que codifique la información contenida en el conjunto de hechos y determinar un orden de recorrido apropiado sobre dicha estructura. Evaluamos los componentes de nuestro marco sobre una subcolección extraída de MEDLINE. Los resultados obtenidos son prometedores.

Palabras clave: minería de textos, recuperación y extracción de información, aplicaciones biomédicas.

Abstract: In this paper, we present a framework for obtaining structurally complex condensed representations of documents sets, which will be used as a base for summarization, answering complex questions, etc. This framework includes a method for extracting a ranked list of facts, triples of the form entity - relation - entity, which relies on dependency parsing-based extraction patterns and language modeling; and methods for constructing a bipartite graph encoding the information contained in the set of facts and determining an appropriate traversing order on that structure. We evaluate the components of our framework on a subcollection extracted from MEDLINE, obtaining promising results.

Keywords: text mining, information retrieval and extraction, biomedical applications.

1 Introduction

Given the exponential growth of the amount of biomedical literature available, clinicians and researchers are forced to use automatic tools to find evidences to support to their tasks and experiments. In biomedicine, PubMed[1] is the main entry point for either users and text-mining applications. Starting from a free-text query, PubMed efficiently returns a list of titles or abstracts in XML format. Unfortunatelly, PubMed relies on boolean queries and results are just ordered by publication date (alternatively by journal, authors and title), which makes it difficult for users to explore the resulting document set.

One of the main retrieval goals of these users is to find relational information about the main entities they handle in their research tasks (e.g. gene, proteins, disease, etc.). Thus, there has been a great interest in developing tools aimed at extracting entity-based relations from the abstracts returned by PubMed.

These efforts may be divided into several classes. First, a number of systems obtain predefined relations between a given

[1] www.pubmed.org

type of entities, for example, PubGene[2] for gene-gene relations. Other systems focus on finding co-occurrences between several types of entities, for example, iHOP[3] for co-occurrences between genes and other chemical compounds and EBIMed[4] for co-occurrences between genes, proteins, cellular components, biological processes, molecular functions, drugs and species, which are semantically annotated using ontologies and dictionaries.

These approaches are limited either by the restrictions on the types of entities they handle or the difficulty at extracting the semantics behind relations inferred by co-occurrence statistics, as this kind of information requires a deeper analysis of the sentences where the identified entities participate.

A group of systems apply deeper analysis techniques. For example, MEDIE[5] applies a deep parsing to the abstracts and performs a semantic annotation, which allows users to pose queries on either the subject, the verb and/or the object.

In a previous paper, we presented a first approximation to the extraction of relevant biomedical information in a document set treating a specific focus concept, which consisted on the obtention of a ranked list of *facts*, triples of the form entity - relation - entity, relevant with respect to that focus concept in a document collection conceptually annotated using terminology from the Unified Medical Language System (UMLS) (Bodenreider, 2006). In this paper, we build on that initial approach to propose a more general framework which includes the generation of this ranked fact list and, additionally, includes methods for building a graph-based structure representing the information contained in the entire document set and determining an appropriate navigation strategy within this structure through link analysis algorithms.

Our fact extraction method differs from related approaches in the nature of the information units used for constructing facts, the way they are extracted, and the way they are used. For example, Filatova and Hatzivassiloglou (2003) consider unnormalized named entities (e.g. persons, organizations, etc.) and a few very frequent nouns, whereas we focus mainly on UMLS concepts. For relations, we only consider verbs, whereas they also consider action nouns, as defined by WordNet (Miller, 1995). Finally, we use dependency parsing-based patterns to extract facts, while they use a named-entity tagger and a position-based event extraction heuristics. Filatova and Hatzivassiloglou (2004) extract triples in order to use them as features for other tasks (e.g. calculating a global score in a sentence extraction method), whereas we treat the graph containing the aggregation of the most relevant and distinctive triples as the information-conveying structure on which further tasks will rely.

The rest of the paper is organized as follows. In Section 2 we describe our framework in detail, whereas in Section 3 we experimentally evaluate its components. Finally, we expose our conclusions in Section 4.

2 Framework description

Given a document collection and a focus concept representing an information need, our framework obtains a representation of the set of documents where this concept is mentioned.

In order to obtain this representation, we first obtain a ranked list of facts, triples of the form entity - relation - entity, which describe events that are distinctive of this document set with respect to the collection and relevant with respect to the focus concept. Every fact conveys a very concise piece of information, e.g. *children - develop- uveitis*. Once the ranking has been obtained, a subset of the best ranked facts is selected for constructing the graph structure that represents the most important information extracted from the document set.

In Figure 1, we depict the overall workflow of our framework. As an offline previous step, we construct a document collection C, which is the result of a topic-based query on MEDLINE (e.g. a specific disease). This collection is conceptually indexed using the concepts from UMLS. The result of this step is a conceptual inverted file where each UMLS concept is mapped to the positions in documents where it is mentioned.

Our framework works on the collection C and uses this conceptual inverted file. For

[2] www.pubgene.org

[3] http://www.ihop-net.org/UniPub/iHOP/

[4] http://www.ebi.ac.uk/Rebholz/srv/ebimed/

[5] http://www-tsujii.is.s.u-tokyo.ac.jp/medie/search.cgi

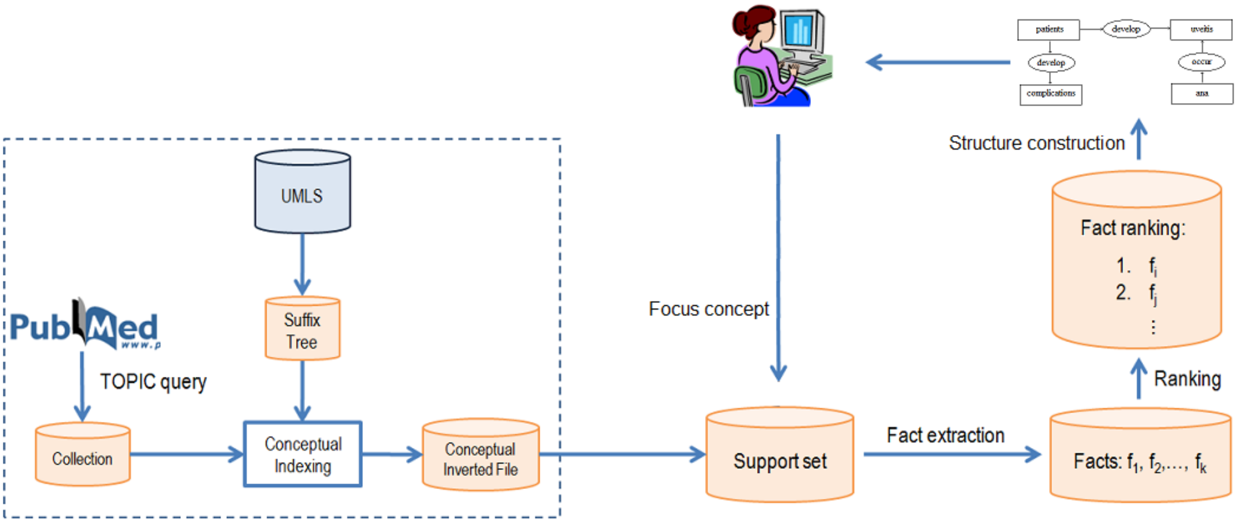

Figure 1: General architecture of our proposal.

a given focus concept, we retrieve the set S of documents from C where it is mentioned, which we call *support set*. The previously described steps are then performed on this support set.

2.1 Building a Conceptually Indexed Collection

Conceptually indexing a collection consists on determining the set of concepts that must be used to describe its contents. In the context of this paper, conceptual indexing allows us to homogenize the terminology used in the medical documents. Additionally, the conceptual index guides the selection of the document sets to be described and semantic relationships may allow to build concept hierarchies to enhance fact extraction with extra knowledge which is not explicitly stated in texts.

As we mentioned before, in this work we use UMLS, specifically version UMLS2008AC, as our knowledge source. The UMLS Metathesaurus is one of the three components of the UMLS Project and comprises many different controlled and well-known vocabularies[6]. Each UMLS concept is linked to a set of synonyms available in the associated vocabularies. In addition, UMLS provides taxonomic relations between concepts.

In order to avoid tagging the entire collection for the occurence of lexical variants of all concepts, the initial, non conceptual inverted index is merged with the lexicon containing the terminology. First, a suffix tree is created containing the entire lexicon. Then, the phrases defined by its paths are used as queries on the collection's single-term inverted index. Thus, a new inverted index is constructed, where entries are Concept Unique Identifiers (CUIs) associated to all documents retrieved by the constructed queries, i.e. those documents that contain some lexical variant of the concept represented by that CUI.

2.2 Fact extraction

As we mentioned previously, facts are simplified representations of the events described in the document set. We consider a fact as a relation between two entities, which is characterized by a verb. Thus, a fact is a triple of the form entity - relation - entity. Here, by *entity* we mean either a lexical variant of a UMLS concept or a non-stopword noun.

Documents are POS-tagged and lemmatized in order to identify verbs and nouns and normalize words into their canonical forms. All occurrences of lexical variants of UMLS concepts are also normalized into the corresponding CUI. For example, *uveitis* and *intraocular inflammation* are both lexical variants of CUI C0042164, so all occurrences of any of them are treated uniformly. No semantic disambiguation is performed on lexical variants, so if a phrase turns out to be a lexical variant of several concepts, it is treated simultaneously as an instance of every concept. In further studies, we will assess

[6] UMLS Source Vocabularies: http://www.nlm.nih.gov/research/umls/metaa1.html

the convenience of applying semantic disambiguation.

Fact extraction is performed in a sentence-by-sentence basis, using the dependencies obtained by a dependency parser in combination with a set of pattern-based extraction heuristics. In this work, we used the Stanford Parser (Klein and Manning, 2003) dependency analysis module (de Marneffe et al., 2006) to obtain dependencies. The following patterns are used:

- **subject - verb - direct complement**
- **subject - verb - indirect complement**
- **subject - verb - prepositional complement**
- **agent complement - verb - passive voice subject**

Since the use of passive voice is very common in scientific literature, the simultaneous use of patterns **subject - verb - direct complement** and **agent complement - verb - passive voice subject** implicitly introduces a simple, partial semantic role labeling-based heuristics allowing to extract facts that follow a general pattern of the form **agent - action - patient**.

For example, the triple *patients - tolerate - etanercept* may be extracted from the sentence *All patients tolerated etanercept with no side effects* as well as from the sentence *Etanercept is well tolerated by pediatric patients*.

When applying the extraction patterns, multi-word lexical variants of UMLS concepts are considered to be good matches for a member of the triple if any of its constituent words is labeled with one of the syntactic dependency tags used in the pattern.

If the subject or any of the used complements is a coordination of several noun phrases linked by the conjunctions *and* or *or*, as many facts are extracted as members of the coordination.

For example, the triples *etanercept - demonstrate - safety* and *etanercept - demonstrate - efficacy* are both extracted from the sentence *Etanercept has demonstrated excellent safety and efficacy in large scale randomised double blind placebo controlled trials.*

Finally, a verb and its negation are treated as a different relation so different facts will be extracted from dependencies where they occur, even if the same entities are involved.

2.3 Initial fact ranking

Two criteria are to be considered in creating a ranked fact list. They must be both relevant to the focus concept according to which the support set was constructed and distinctive with respect to the collection. In order to create a ranking where both criteria are simultaneously considered, we follow a language modeling approach. We construct the unigram models of the set of terms (entities and verbs) in both the support set S and the collection C, as well as the language models of the facts in the support set and the collection.

The unigram model of the collection, M_C, is estimated by maximum likelihood (ML). Thus, for a term t:

$$P(t \mid M_C) = \frac{count(t)}{\sum\limits_{t' \in V} count(t')} \tag{1}$$

where V is the vocabulary of the collection and $count(t)$ indicates the number of occurrences of t in the collection.

Since the support set being described is focused on a concept, we take this into account for estimating its unigram model in such a way that it is *biased* towards the focus concept. We express the biased unigram model of the support set, $M_{S_{biased}}$, as a mixture of three components: the ML unigram model of the set of sentences in S where some lexical variant of the focus concept occurs, M_{focus}, the ML unigram model of the set of sentences in S where some lexical variant of either the focus concept or its immediate hyponyms in the UMLS concept hierarchy occur, M_{exp}, and the ML unigram model of the support set S itself, M_S.

Unlike common language modeling approaches, where mixture models are used for smoothing or modeling the presence of several underlying topics in the documents, in our approach the mixture is used as a mechanism to favor the selection of terms coocurring with lexical variants of the focus concept and/or its related concepts. Notice that the sentence set from which M_{focus} is estimated is a subset of the sentence set from which M_{exp} is estimated, which is in turn a subset of S.

For instance, if the focus concept is C0042164, the sentences containing *uveitis* or *intraocular inflammation* will be considered for estimating M_{focus}, whereas the sentences

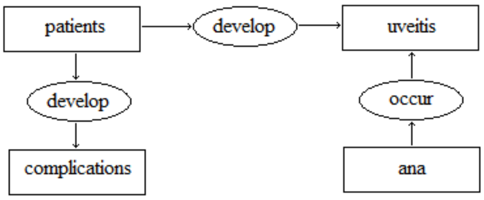

Figure 2: Example of the graph generation method.

containing *uveitis*, *intraocular inflammation*, *anterior uveitis*, *intermediate uveitis*, *posterior uveitis*, *panuveitis* or *diffuse uveitis* will be considered for estimating M_{exp}. The latter may be seen as a form of concept hierarchy-based query expansion.

Finally, the occurrences of terms in sentences not containing lexical variants of neither the focus concept nor any of its immediate hyponyms will only be accounted for when estimating M_S. Since the three components contribute to the focus concept-biased model $M_{S_{biased}}$, the estimated probability of terms in the context of the focus concept and/or its immediate hyponyms will be increased at expense of the estimated probabilities of non coocurring terms.

Thus, the probability of a term t in $M_{S_{biased}}$ is calculated as:

$$\begin{aligned} P(t|M_{S_{biased}}) &= \lambda_0 P(t|M_{focus}) &+ \\ &+ \lambda_1 P(t|M_{exp}) &+ \\ &+ \lambda_2 P(t|M_S) & \end{aligned} \quad (2)$$

where $\lambda_0 + \lambda_1 + \lambda_2 = 1$.

The language models of the set of facts in the collection and the support set, M'_C and $M'_{S_{biased}}$, are estimated in a similar way.

Two criteria are considered when ranking facts: first, the triple representing the fact must be distinctive as a whole; second, the three terms composing the triple must be distinctive as well.

For a term, or a triple representing a fact, we use its contribution to the Kullback-Leibler (KL) divergence between the language model of the support set and that of the collection as a measure of how distinctive the term or triple is. The contribution of a term t to the KL divergence between $M_{S_{biased}}$ and M_C is defined as:

$$KLC(t) = P(t|M_{S_{biased}}) \log \frac{P(t|M_{S_{biased}})}{P(t|M_C)} \quad (3)$$

Notice that KLC values above zero characterize terms that are more frequent according to $M_{S_{biased}}$ than according to M_C, thus being distinctive terms of the support set. Also notice that as KLC values grow, terms may be considered more distinctive.

The contribution of a fact $f = (e_1, r, e_2)$ to the KL divergence between $M'_{S_{biased}}$ and M'_C is calculated similarly.

Since we intend to rank facts according to the distinctiveness of both the triples by which they are represented and that of the terms conforming these triples, we calculate the score of a fact $f = (e_1, r, e_2)$ as

$$\begin{aligned} score(f) &= KLC(f) * KLC(e_1) & * \\ &* KLC(r) * KLC(e_2) & \end{aligned} \quad (4)$$

2.4 Constructing and traversing a global structure

In order to make the extracted information navigable, as well as facilitating further tasks, such as summarization, complex question answering, etc., we construct a structure where all relevant information is aggregated, thus allowing to consider global scale interactions between entities and relations.

Graphs have been widely used for representing entity-relation information in a structured way. Following this line of thought, we aggregate all information in the ranked fact list into a bipartite graph. In this graph, a first set of nodes represents the entities and a second set represents the relations.

Every entity occurring in the ranked fact list is represented by one node in the graph. Relations are not treated in the same way. For every fact a relation is involved in, a new node representing this occurrence of the relation is added. Finally, for every fact (e_1, r, e_2), edges are included linking the node representing e_1 to the node representing the corresponding occurrence of r and this node to the node representing e_2. Both edges are weighted by the score of the fact. Notice that no edge links e_1 and e_2 directly. Adding a different node to represent every occurrence of a relation prevents the structure from encoding inconsistent information. For example, if a single node is used to represent every relation, the subgraph obtained by adding the facts (e_1, r, e_2) and (e_3, r, e_4) would be the same as the one obtained by adding the facts (e_1, r, e_4) and (e_3, r, e_2), which is not desirable. To better illustrate the graph construction process, Figure 2 shows an example of the graph obtained for the set of facts

patients - *develop* - *uveitis*
ana - *occur* - (*in*) *uveitis*
patients - *develop* - *complications*

In order to determine a convenient presentation order, we take into account both the scores obtained at creating the original ranked fact list and the sctructural importance of nodes in the graph.

It is important to notice that, while the first ranking aims at obtaining the most relevant and distinctive facts, i.e. determining which information to include in the representation; this second ranking aims at determining a convenient order for presenting and navigating the information conveyed by these facts. Since fact scores are used for weighting the edges in the graph, the second ranking is not unaware of informational relevance when determining structural importance.

Structural importance of entities and relations is assessed via a link analysis algorithm on the graph. In our framework, we use a variation of PageRank for weighted graphs, which is defined as follows (Mihalcea, 2004):

$$\begin{aligned} PR(v_i) &= (1-\alpha) + \\ &+ \alpha \sum_{v_j \in In(v_i)} w_{ji} \frac{PR(v_j)}{\sum\limits_{v_k \in Out(v_j)} w_{kj}} \end{aligned} \qquad (5)$$

where $In(v)$ represents the set of nodes v_i such that there exists and edge (v_i, v), $Out(v)$ represents the set of nodes v_i such that there exists and edge (v, v_i), w_{ij} represents the weight of the edge linking node v_i to v_j, and parameter α expresses how much importance is given to the graph structure and is normally set to 0.85.

Adding nodes to represent both entities and relations allows us to obtain scores for all of them, not only for entities. Although a single relation may be represented by several nodes in the graph, the final measure we use for determining its structural importance is the sum of PageRank values over all the nodes representing it, which we will refer to as *aggregated* PageRank.

Once final scores have been obtained, the presentation order to be used is determined by a breadth-first traversal of the graph in the following manner. First, the entity-representing node having the greatest PageRank value is selected as the starting point v_{e_s}. Let $v_{r_1}, v_{r_2}, \ldots, v_{r_k}$ be the relation-representing nodes linked to v_{e_s} and $v_{e_1}, v_{e_2}, \ldots, v_{e_k}$ the corresponding entity-representing nodes linked to them. For every pair v_{r_i}, v_{e_i}, if $AggrPR(v_{r_i})$ or $PR(v_{e_i})$ are below given thresholds, the fact (e_s, r_i, e_i) is discarded for presentation. For every remaining fact (e_s, r_i, e_i) to be considered, a new score $score_{gr}$ is calculated as follows:

$$score_{gr}(f) = AggrPR(v_{r_i}) * PR(v_{e_i}) \qquad (6)$$

Facts are ordered for presentation in descending order of this score. Notice that $PR(v_{e_s})$ is not considered since it does not affect the ordering. Following this order, every newly reached entity-representing node is then taken as starting point and the process is repeated recursively until all includible facts have been added to the final ordering.

3 Evaluation

There are different considerations to take into account at evaluating the components of our framework. In the case of the initial ranked fact list, traditional Information Retrieval quality measures may be used as good indicators of the performance of the method. On the other hand, evaluating navigability or appropriateness of a given presentation order is not trivial since these notions are not well specified and are difficult to quantify.

In our experimental setting, we constructed a conceptually indexed collection by retrieving from MEDLINE documents that satisfy the query *juvenile idiopathic arthritis* (*JIA*). This collection is composed by 7654 documents (45672 sentences), which are described by 32350 terms, out of which 12572 represent lexical variants of UMLS concepts found during conceptual indexing.

Three support sets were retrieved according to the focus concepts C0177758 (*etaner-*

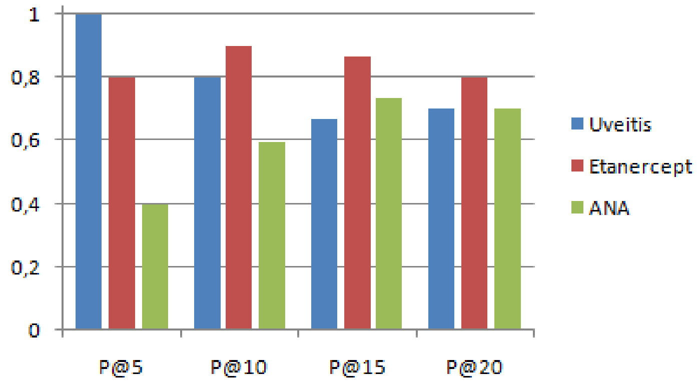

Figure 3: Precision at top elements for the three support sets.

cept), a drug used for treating JIA; C0042164 (*uveitis*, *intraocular inflamation*), a complication of JIA; and C0003243 (*antinuclear antibody*), an indicator of the presence of the disease.

The parameters in Equation 2 were empirically set to $\lambda_0 = 0.7$, $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$. After fact rankings for each support set were constructed according to the proposed method, the 20 top-ranking facts in each case were manually evaluated, labeling them as relevant or not relevant.

The quality of the rankings was measured in terms of precision at k top ranking elements ($P@k$), a typical IR measure, which is defined as:

$$P@k = \frac{\#\text{ of relevant facts in top } k}{k} \tag{7}$$

The nature of the problem makes it impossible to define the entire set of relevant facts, which prevents us from using metrics depending on it, such as recall or average precision.

Figure 3 shows the results obtained for the three support sets for $k \in \{5, 10, 15, 20\}$.

A manual inspection of the rankings allowed us to determine that the main cause for the extraction of incorrect facts was the effect of dependency parsing errors, which mislead the extraction rules. The error that most commonly affected fact extraction was the incorrect attachment of prepositional phrases modifying a noun phrase, which were instead attached to the clause main verb as a prepositional complement. Some of these erroneous facts reached a high position in the ranking because of two reasons. First, their noisy nature, which makes them extremely unfrequent in the entire collection, thus obtaining high KLC values. Second, the occurrence of high KLC-valued entities in the fact. The combination of both circumstances is likely to make these facts obtain high scores. Although we observed cases of this situation for all three focus concepts, it occurred particularly often in the fact ranking obtained for focus concept *antinuclear antibody*.

In our opinion, the values at which $P@k$ appears to stabilize, around 0.7, are reasonably good, although there is still room for improvement in the fact extraction patterns and the ranking score formula.

As we mentioned previously, it is hard to define measures of how navigable or purpose-fit a particular fact presentation order is. In order to illustrate the performance of the graph-based structuration method, in Figure 4 we show a fragment of the ordering obtained for focus concept C0717758 (*etanercept*), setting the fact discarding thresholds at 0.01. As it may be observed, the presentation order first provides all facts having as subject the entity that emerges as the most important in the graph structure (which does not necessarilly mean that it is the most relevant with respect to the focus), then provides this information for the entities that are introduced by facts linking them to the chosen entity, and so on. We consider this behavior to be useful, as it may provide a good paragraph structure for future text generation methods.

4 Conclusions

In this paper, we have presented a framework for obtaining structurally complex condensed representations of document sets in the biomedical domain. Facts, concise information units conveying information about

```
patients (C0030705) -- tolerate -- etanercept (C0717758)
patients (C0030705) -- tolerate -- treatment (C0087111)
patients (C0030705) -- tolerate -- (with no) side effects (C0001688) (incorrect fact)
etanercept (C0717758) -- demonstrate -- efficacy (C1707887)
etanercept (C0717758) -- demonstrate -- safety (C1705187)
etanercept (C0717758) -- demonstrate -- beneficial activity (C0600075)
etanercept (C0717758) -- approved -- (in) europe (C0015176)
etanercept (C0717758) -- approved -- (in) united states (C0041703)
[...]
```

Figure 4: Fragment of the ordering obtained on the graph generated for focus concept *etanercept.*

relations held between entities, are the base from which a graph structure representing the document set is constructed.

Facts are extracted by simple dependency parsing-based patterns and ranked by their relevance and distinctiveness in the document set using a Language Modeling approach. Link analysis algorithms are used in order to determine a presentation order over this graph, which arguably facilitates tasks such as summarization, complex question answering, etc.

Despite the simplicity of the fact extraction procedure, experimental results, obtained over three different document sets from a subcollection of MEDLINE, are encouraging. We have presented a case study of the graph construction method and the obtained ordering, which we intuitively consider to be sound and useful. However, a principled evaluation criterion for the quality of the proposed presentation order is still required.

While our method has been initially proposed for the biomedical domain, we consider that it may be ported to other domains for which rich knowledge resources are available.

In addition to the previously mentioned need for an evaluation criterion for the presentation order, other attractive directions for future work include improving fact extraction mechanisms, mainly by taking into account the semantic nuances introduced by the use of different syntactic patterns, leading prepositions in prepositional phrases, etc. Besides, we intend to use semantic relations contained in the concept hierarchies to constrain and/or generalize the initial set of facts to be considered and/or enrich it with non explicit information.

References

Bodenreider, O.: 2006. Lexical, Terminological, and Ontological Resources for Biological Text Mining. In *Text Mining for Biology and Biomedicine*. Artech House.

Filatova, E. and V. Hatzivassiloglou: 2003. Domain-Independent Detection, Extraction, and Labeling of Atomic Events. In *Proceedings of RANLP 2003*, pages 145–152, Borovets, Bulgaria.

Filatova, E. and V. Hatzivassiloglou: 2004. Event-Based Extractive Summarization. In *Proceedings of the ACL 2004 Workshop "Text Summarization Branches Out"*, pages 104–111, Barcelona, Spain.

Klein, D. and C. D. Manning: 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

de Marneffe, M. C., B. MacCartney and C. D. Manning: 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC 2006.*

Mihalcea, R.: 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Proceedings of the ACL 2004 Interactive Poster and Demonstration Sessions.*

Miller G. A.: 1995. WordNet: a Lexical Database for English. *Communications of the ACM*, 38(11): 39–41.