Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 159-167 recibido 1-05-2009; aceptado 5-06-2009

# Describing Biomedical Document Sets in Terms of its Most Distinctive Facts

## *Descripción de conjuntos de documentos biomédicos a través de sus hechos más distintivos*

**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

**Aurora Pons-Porrata** Center for Pattern Recognition and Data Mining Universidad de Oriente Santiago de Cuba, Cuba aurora@cerpamid.co.cu

**Resumen:** En este artículo proponemos un método para describir un conjunto de documentos biomédicos, conceptualmente indexados, a través de sus hechos más distintivos. Estos documentos han sido recuperados como soporte de un concepto foco, el cual representa una necesidad de información. Los hechos utilizados para la descripción son unidades de información concisas, representadas mediante tripletas con la forma entidad-verbo-entidad. Éstos se presentan ordenados por su relevancia con respecto al concepto foco, la cual se calcula usando modelos de lenguajes. Los resultados experimentales, obtenidos sobre tres conjuntos de documentos de una colección extraída de MEDLINE, son prometedores.
**Palabras clave:** minería de textos, recuperación de información, aplicaciones biomédicas.

**Abstract:** In this paper, we propose a method to describe a set of conceptually indexed biomedical documents in terms of its most distinctive facts. These documents are retrieved to support the occurrence of a focus concept, which expresses an information need. The facts used for description are concise information units, represented as triples of the form entity-verb-entity. These are presented as a ranked list, ordered by their relevance with respect to the focus concept, which is determined using a language modeling approach. Experimental results, obtained on three document sets over a collection extracted from MEDLINE, are promising.
**Keywords:** text mining, information retrieval, biomedical applications.

## *1 Introduction*

Information retrieval is systematically used by clinicians and researchers to find evidences that give 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. For example, PubGene[2] generates a gene network from the gene relations found in the sentences where the keywords fetched by the user appear. Similarly, iHOP[3] identifies the sentences where the keyword given by the user (i.e. a gene name) co-occurs with other genes or chemical compounds. Afterwards, the user can build a gene model by selecting the sentences deemed as relevant. Both systems are only focused to gene entities, which limits their range of application. For example, it is not possible to extract re-

[1] www.pubmed.org

[2] www.pubgene.org

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

ISSN: 1135-5948 

lations between genes and other medical concepts such as diseases, anatomical parts, etc. EBIMed[4] is aimed at finding richer relations between biomedical entities other than gene and proteins. EBIMed semantically annotates abstracts with a series of ontologies and dictionaries (e.g. Gene Ontology, Drug Bank, UniProt, etc.). Then, EBIMed extracts statistically significant co-occurrences between annotated entities. The semantic entities regarded by EBIMed are: Protein/Gene, Cellular component, Biological process, Molecular function, Drug and Species.

A limitation of all the previous approaches is that they do not provide clues about the true relation that is behind the found co-occurrences. This kind of information requires a deeper analysis of the sentences where the identified entities participate. For example, the system 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 this paper, we present a first approximation to the discovery of relevant biomedical information for a specific concept: the *focus* (e.g. a given disease, a given gene). Unlike previous approaches, we are aimed at providing *a ranked list* of facts, which are extracted from the context of the focus concept and are relevant to it. We use the concepts from the Unified Medical Language System (UMLS) (Bodenreider, 2006) for semantically annotating the document collection. UMLS regards a much wider range of biomedical entities (more than 100) than previous approaches, thus providing a richer set of relations to the users. Facts are represented as triples of the form entity-verb-entity. They are extracted using a simple heuristic, which does not rely on syntactic analysis, and ranked according to their relevance with respect to the focus concept using a language modeling approach. The distinctiveness of facts in the context of the focus concept and its hyponyms, both at document and sentence level, is used as a measure of their relevance.

Additionally, in order to increase understandability and provide hints for extra information, relevant facts extracted by our method are contextualized by the set of sentences where they occur.

Our method for extracting facts is similar to that proposed by Filatova and Hatzivassiloglou (2003). However, while they consider unnormalized named entities (e.g. persons, organizations, etc.) and a few very frequent nouns, we consider all non stopword nouns and instances of UMLS concepts. Besides, we only consider verbs, whereas they also consider *action nouns*, as defined by WordNet (Miller, 1995). On the other hand, Filatova and Hatzivassiloglou (2004) propose to use triples as features for other tasks (e.g. calculating a global score in a sentence extraction method), while we treat triples as the basic information-conveying unit, in terms of which the document set is described.

The rest of the paper is organized as follows. In Section 2 we describe our proposal in detail, whereas in Section 3 we describe the experiments carried out to evaluate the validity of our method. Finally, we expose our conclusions in Section 4.

## 2 Our proposal

Given a document collection and a focus concept representing an information need, our proposed method allows to describe the set of documents where this concept is mentioned. This description consists in a ranked list of facts, triples of the form entity-verb-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".

In Figure 1, we depict the overall workflow of our proposal. 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 method works on the collection $C$ and uses this conceptual inverted file. For a given focus concept, we retrieve the set $S$ of documents from $C$ where it is mentioned, which we call support set. In order to obtain a description of $S$, we extract the set of facts that occur in it. These facts are ranked according

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

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

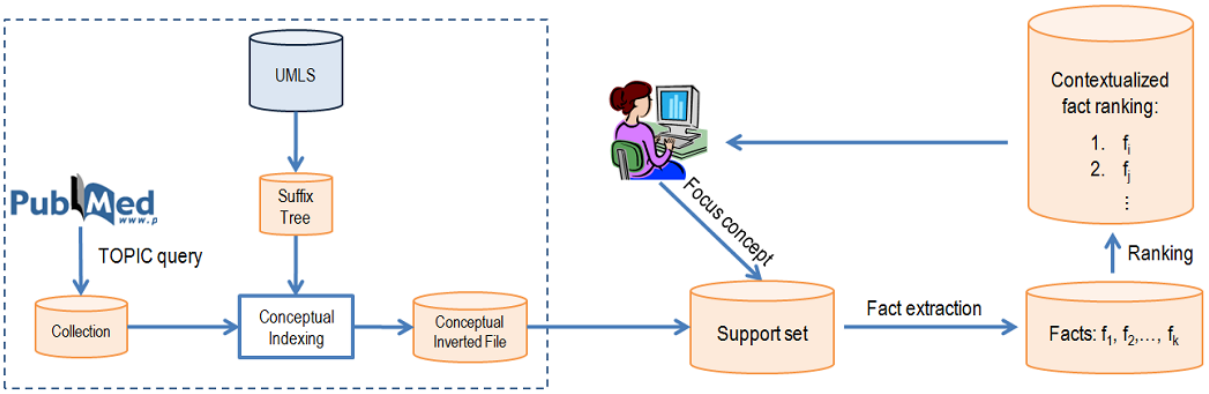


Figure 1: General architecture of our proposal.

to their distinctiveness, giving more importance to those that are specific to the support set of the focus concept. In order to increase understandability and explainability, each fact in the ranking is contextualized by the set of sentences that support it.

## 2.1 Building a Conceptually Indexed Collection

Conceptually indexing a collection consists on finding the set of concepts that describe its contents. Usually, concepts are taken from well established external knowledge sources such as WordNet or specific-purpose thesauri. External knowledge sources usually consist of two parts, namely: a very large lexicon with the different lexical variants of the concepts and, a set of semantic relationships between concepts (e.g. "is a").

In the medical domain, conceptual indexing is critical for managing the huge volumes of scientific documents stored in publicly available repositories. For example, MEDLINE, the world's largest repository of medical publications, is indexed through the Medical Subject Heading (MeSH), which allows users to search and browse the publications more effectively. Since MeSH concepts are assigned by authors, the resulting annotations may either not cover properly the document content or be biased to some aspect. For this reason, it is preferable to automate as much as possible the conceptual indexing of the document 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 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, version UMLS2008AC, as the external 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.

Unfortunately, current approaches for automatically identifying concepts from medical documents, like MetaMap (Aronson, 2001), are not scalable for large document collections. Alternative methods recently proposed by Rebholz-Schuhmann et al. (2008), called MWT (MultiWord Tagger), consist on applying pure dictionary look-up techniques based on finite state automata. This approach is very efficient, more precise than MetaMap but with lower recall values (Jimeno-Yepes et al., 2008). However, its main limitation is the enormous space required to allocate large lexicons.

In this work, we have approached the problem of tagging a large collection by taking a global strategy. That is, instead of tagging one by one the documents of a collection, we index concepts in the whole collection by merging both the vocabulary of the collection and the external lexicon.

We start with the inverted representation of the document collection. Thus, now each collection item is a single term (e.g. lemmatized word present in the collection) which

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

```
joint|C0444497,C0555829,C0022417,C1706309,C1269611,C0558540
        spherical|C0224504
        pastern|C1279617
        Jaw|C0039493
        zygapophyseal|C0224521
                entire|C1267117
                thoracic|C0504605
        tibiofemoral|C1269072,C0447795
                right|C0834358
                left|C0834359
        xiphisternal|C0447790,C1280647
```

Figure 2: Fragment of the suffix tree of the UMLS lexicon.

has associated the set of its document hits. Obtaining the inverted representation of a document collection is straightforward, and usually is a previous step required for information retrieval tasks.

On the other hand, the UMLS lexicon is organized into a suffix tree as follows. First, each lexicon string is processed to identify its head noun. This is done with a few simple rules that detect prepositions. Meaningless words are removed from the lexicon strings. Then, the string tokens are ordered so that the head appears first and its modifiers appear at the back. The resulting list is inserted into the suffix tree, where each token is associated to a tree node. Finally, concepts associated to each string are attached to the node of its last token. Figure 2 presents a fragment of the resulting suffix tree. In this case, concepts are expressed as CUIs (Concept Unique Identifiers) of UMLS. Notice that some paths of the tree lead to ambiguous conceptual representations (they are associated to more than one concept). Currently, ambiguity issues are not addressed in our system, leaving them for future work.

The process of merging both the inverted file and the suffix tree is described in Algorithm 1. This basically follows a greedy strategy so that paths of the suffix tree are transformed into queries for the inverted file. As a result, each concept accessed through a path is associated to the documents retrieved with its query. More specifically, the algorithm takes each maximal path of the suffix tree and produces a query for each of its subpaths with length greater than one. If such a query succeeds, the retrieved documents are associated to the concepts reached with the corresponding path. Consequently, the references of the retrieved documents are removed from the inverted file entries associated to the query. In this way, these references are not regarded again when checking other subpaths. It is worth mentioning that inverted file queries are evaluated as boolean and expressions, taking into account plural/singular forms and proximity contraints between query terms.

## 2.2 Fact extraction and ranking

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 an action. Thus, a fact is a triple of the form (*entity*, *verb*, *entity*). Here, by *entity* we mean any non-stop word noun or a phrase which is a lexical variant of a concept.

Facts are extracted using a simple mechanism which does not rely on complex syntactic analysis. Documents were POS-tagged and lemmatized in order to identify verbs and nouns and collapse words to their canonical forms. Additionally, all occurrences of lexical variants of an UMLS concept are also collapsed to a single term representing the concept. For example, the phrases “uveitis” and “intraocular inflammation” are both lexical variants of the UMLS concept C0042164, so occurrences of any of them are treated as occurrences of the concept. As previously mentioned, no disambiguation is performed on lexical variants, so if a phrase may be a lexical variant of several concepts, it will be treated simultaneously as an instance of every concept.

Every triple formed by a pair of entities cooccurring in a sentence and a verb that occurs between them is extracted. This shallow heuristic is motivated by the common subject-verb-object (SVO) phrase order of English. For example, the triples “children”-“develop”-“rash”,

**Algorithm 1** Algorithm for merging lexicons and inverted files

**Input:** Suffix Tree $FP$; Inverted File $FI$.
**Output:** Conceptual Inverted File $FCUI$.

```
1. for all maximal path p in FP (ordered by length) do
     Normalize p according to the rules applied to the inverted file FI
     Revise.append(p[0])
     while |p| > 1 do
       if p has concepts in FP then
         R=FI.query(p)
         if |R| > 0 then
           update FCUI with FCUI[FP.concepts(p)]=R
           update FI removing the elements of R associated to p
         end if
         p.removeLastToken()
       end if
     end while
   end for
2. for all p in Revise do
     R=FI.query(p)
     if |R| > 0 then
       update FCUI with FCUI[FP.concepts(p)]=R
     end if
   end for
```

"children"-"develop"-"polyarthritis" and "children"-"develop"-"uveitis" are among those that may be extracted from the sentence "*All children developed the typical symptom triad of rash, polyarthritis and uveitis, with onset before their 4th birthday*".

In order to create a ranking of the most salient facts, which are additionally relevant to the focus concept according to which the support set was constructed, 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 $t$:

$$P(t \mid M_C) = \frac{count(t)}{\sum_{t' \in V} count(t')} \qquad (1)$$

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 $S$. Because of this, the occurrences of terms in sentences containing lexical variants of the focus concept will be taken into account for the three components of the mixture, whereas their occurrences in sentences containing lexical variants of immediate hyponyms of the focus concept but no lexical variants of the focus concept itself will be taken into account for estimating $M_{exp}$ and $M_S$, but not for $M_{focus}$.

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

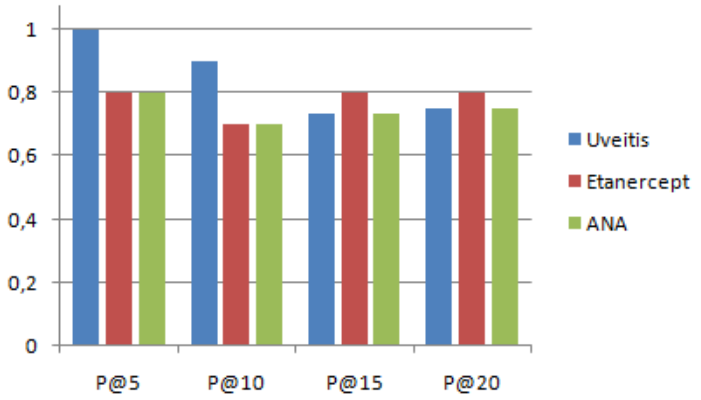


Figure 3: Precision at top elements for the three support sets.

tences 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 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 $f = (e_1, v, e_2)$ 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 $f = (e_1, v, e_2)$ as

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

Since no disambiguation is being performed on lexical variants, lexically redundant triples may be obtained. In order to prune the ranking, every fact $f = (e_1, v, e_2)$ is compared to all those facts outranking it, and it is eliminated if an outranking fact is found such that its three components, pairwisely, are identical or share lexical variants.

Once the ranking has been obtained, in order to improve understandability and explainability, each fact is contextualized by its supporting sentences, i.e, those sentences in which its representing triple occurs.

## 3 Experiments

Our experiments aim to asses the quality of the fact rankings obtained. We additionally

```
Score:  8.21e-09
[C0030705(patient/patients) -- develop -- C0042164(inflammation
intraocular/uveitis)]
        About 20% of patients with juvenile chronic arthritis develop uveitis
which is frequently bilateral
        We describe a patient with adult onset Still's disease AOSD who developed
meningoencephalitis sensorimotor peripheral neuropathy and uveitis during the
course of disease
[...]
Score:  3.55e-09
[C0030705(patient/patients) -- C0011900(diagnoses/diagnosed/diagnosis) --
C0042164(inflammation intraocular/uveitis)]
        Between January 1989 and December 1999 in the Department of Ophthalmology
of Hacettepe University School of Medicine 219 patients were diagnosed or
observed as having pediatric uveitis
        Patients diagnosed with uveitis before or within 1 year from the onset
of arthritis required longer treatment and suffered more episodes than those with
uveitis found later on
[...]
Score:  1.87e-09
[C0042164(inflammation intraocular/uveitis) -- develop --
C0030705(patient/patients)]
        Uveitis developed in the patients at a mean age of 9.0 years
[...]
```

Figure 4: Fragment of the ranking obtained for the support set associated to concept C0042164.

conducted a qualitative evaluation of the results, which allowed us to draw some conclusions and clues for future improvement.

We constructed a conceptually indexed collection by retrieving the documents from MEDLINE 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 C0042164 ("uveitis", "intraocular inflamation"), a complication of JIA; C0177758 ("etanercept"), a drug used for treating JIA; and C0003243 ("antinuclear antibody"), an indicator of the presence of the disease.

The parameters in Equation 2 were empirically set to $\lambda_0 = 0.6$, $\lambda_1 = 0.3$ 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 top ranking elements ($P@k$), which is a usual measure for rankings in Information Retrieval. This measure is defined as:

$$P@k = \frac{\#\text{ of relevant facts in the top } k}{k} \quad (5)$$

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 $k \in \{5, 10, 15, 20\}$.

As it may be noticed, the highest results over all three support sets are obtained for $k = 5$, indicating that top ranking facts are determined more accurately, which corresponds with the behavior typically expected by users.

As $k$ grows, the support set corresponding to concept C0042164 ("uveitis", "intraocular inflamation") behaves differently from the other two support sets. While precision values for "uveitis" decrease uniformly up to $k = 15$, "etanercept" and "ANA" show a minimum precision for $k = 10$ and grow slightly for $k = 15$ and $k = 20$.

The decreasing behavior for "uveitis" is caused by the fact that, in this case, the Kullback-Leibler divergence contribution value for the focus concept is considerably

| a) | b) |
|---|---|
| intraocular inflammation / uveitis | diagnose |
| complication / complications | develop |
| patient / patients | treat |
| visual / visualized | occur |
| jia-associated | mean |

Figure 5: Top-ranking terms according to KLC: a) entities, b) verbs.

greater than those for triples representing facts and dominates the score in Equation 4. In the future, we will evaluate modifications to Equation 4 to take this effect into account.

Regarding the minimum precision obtained for $k = 10$ for the support sets of "etanercept" and "ANA", we verified that they were caused by POS tagging errors which caused nouns, e.g. *tests*, to be tagged as verbs, forming incorrect triples.

In general, we consider these results promising, specially given the simplicity of the heuristic used for extracting facts.

In order to ilustrate the obtained results, in Figure 4 we show a fragment of the ranking obtained for the support set associated to concept C0042164 ("uveitis", "intraocular inflammation"). As it may be observed, facts are arranged by score and the triples are shown including all possible lexical variants of UMLS concepts. Due to space limitations, we show only a subset of the supporting sentences for each fact. Notice that supporting sentences facilitate the interpretation of facts that might be confusing otherwise. For instance, the third fact might be misleading about the entity that performs the action and the one that receives the result (interpretable as *a disease developing a patient*), but the supporting sentence clarifies the situation. Additionally, supporting sentences may allow users to find extra information (e.g. the age of patients who have developed a disease) investing a considerably smaller amount of effort than would have been required if these sentences had not been hinted by the facts they support.

In Figure 5 we show the five top-ranking entities and verbs according to their KLC values. Notice that top-ranking triples shown in Figure 4 are composed by the top-ranked entities and verbs.

## 4 Conclusions

In this paper, we have proposed a method for describing a set of documents in the biomedical domain, which is known to be constructed to support the occurrence of a concept, in terms of its most distinctive facts. Facts provide concise information about relations held between concepts, which are useful for physicians and other specialists. Additionally, facts are contextualized by sentences, which facilitate their interpretation and may provide extra information.

Facts are represented by triples of the form entity-verb-entity, which are extracted by a simple mechanism which does not utilize syntactic analysis. Language modeling is used to model the set of extracted facts in order to create a ranking ordered by their distinctiveness in the context of the document set.

Despite the simplicity of the fact extraction procedure, experimental results, obtained over three different document sets from a subcollection of MEDLINE, are promising.

While our method has been initially proposed for the biomedical domain, we consider that it may be ported to other domains for which external knowledge sources providing concepts and relationships are available.

Attractive directions for future work include improving fact extraction mechanisms in such a way that it becomes possible to extract relations that are being missed currently. Besides, we intend to use semantic relations contained in the concept hierarchies to constrain and/or generalize the set of facts to be considered for the ranking. Finally, we intend to use the set of top-ranking facts and their supporting sentences for constructing summaries of the document sets associated to the relevant concepts.

**Acknowledgements**

This work has been partially funded by the CICYT Project TIN2008-01825/TIN and the Research Promotion Program 2008 of Universitat Jaume I, Spain.

## *References*

Aronson, A. R.: 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the 2001 AMIA Symposium*, pages 17–21, Washington DC, USA.

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.

Jimeno-Yepes, A., E. Jimenez-Ruiz, V: Lee, S. Gaudan, R. Berlanga-Llavori and D. Rebholz-Schuhmann: 2008. Assessment of diseases named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(Suppl 3):S3.

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

Rebholz-Schuhmann, D., M. Arregui, S. Gaudan, H. Kirsch and A. Jimeno-Yepes: 2008. Text processing through Web services: Calling Whatizit. *Bioinformatics*, 24(2): 296–298.