# Identifying Relations between Medical Concepts by Parsing UMLS® Definitions

Ivelina Nikolova and Galia Angelova

Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences
25A, Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
{iva,galia}@lml.bas.bg
http://lml.bas.bg/

**Abstract.** To automatically analyse medical narratives, one needs linguistic and conceptual resources which support capturing of important information from texts and its representation in a structured way. Thus the conceptual structures encoding domain concepts and relations are crucial for the development of reliable and high-performance information extraction system. We present research work enabling automatic extraction of relations between medical concepts. The lack of conceptual resources with Bulgarian ontological vocabulary provoked us to reuse already existing resources with English labels, more especially the UMLS® Metathesaurus®. We form a terminological dictionary of the Bulgarian terms of interest, translate them to English and extract their UMLS definitions which are short English statements in free text. These definitions are processed automatically by a semantic parser; afterwards we apply additional extraction, alternation and validation rules and built a set of new relations to be inserted in our conceptual resource. The article presents the input data and available tools, the knowledge chunks extracted from UMLS and their processing, as well as a discussion of the present results.

**Keywords:** relation extraction, clinical terms, biomedical NLP.

## 1 Introduction

Secondary use of medical health records is an important research trend which has been given high priority especially in the recent years. Processing patient data by all possible means will enable the improvement of the clinical decision-support systems, health management and treatment through provision of personalised healthcare services. There are various ways to reuse the information stored in the electronic health records; an important activity is to automatically analyse the free text paragraphs which present the most important findings regarding the case history. The records contain focused information and differ from patient to patient, from author to author. In order to unify the structure of the extracted descriptions and produce a common template where each patient's feature with

S. Andrews et al. (Eds.): ICCS 2011, LNAI 6828, pp. 173–186, 2011.

its attributes and values has a separate slot, we need a kind of conceptual framework which supports the segmentation and extraction of focal information from the raw text. Such a framework could be provided by an ontological resource covering the medical concepts/terms used in the patient records and the relations between them. Thus knowledge about concepts and relations is crucial for building reliable and high-performance information extraction (IE) systems.

Analysing automatically Patient Records (PRs) in Bulgarian language, we need underlying conceptual resources labeled by Bulgarian medical terms (to be able to map the text units onto conceptual entities). Unfortunately no medical ontologies with Bulgarian vocabulary exist (except the International Classification of Diseases ICD) but there are large resources available for other languages. The UMLS Metathesaurus [1], which is widely used in biomedical Natural Language Processing (NLP), contains information about biomedical and health related concepts, their various names in several languages, and some labeled relationships among them. One finds their descriptions of various depth and quality: comprehensive, systematic representation of concepts and relations in some areas and schematic or poor coverage in other areas.

We present here research efforts and experiments done in order to enrich a domain model supporting IE from hospital PRs in Bulgarian language. In general we aim at the construction of a task-specific Bulgarian terminological bank enriched with relations that are obtained automatically by reusing existing resources in English. We have built a terminological dictionary containing Bulgarian terms of interest found in the training corpus of PRs, translated these terms to English and extracted their UMLS definitions which are short English statements in free text. These definitions were processed automatically by a semantic parser, that transforms them to dependency and samantic framing structures which easily map to concept graphs (CGs). Afterwards we have applied additional extraction, alternation and validation rules and have built a set of new relations to be inserted in our conceptual resource. The article will present the work done so far with focus on the conceptual processing: the input data, the knowledge chunks extracted from the UMLS system using the available tools, the procedures for automatic processing of the free text definitions in order to extract relations between clinical terms, and the evaluation of the present results. To the best of authors knowledge this task has been explored for Bulgarian only under the current project [2].

The article is structured as follows: section 2 summarises related work regarding automatic relation acquisition; section 3 considers the background resources and tools; section 4 presents our relation extraction approach; section 5 discusses the results and section 6 contains the conclusion.

## 2 Related Work

Defining relations between concepts is a puzzling knowledge representation exercise since the relations in general reflect implicit and task-dependent connections between entities. All attempts to unify the AI approach to relation elicitation have failed; for instance we still see practical solutions where relations between concepts are labeled by verbs and the concept labels are interpreted as verb role fillers, e.g. *subject-object*; on the other hand using the verb thematic roles (e.g. *agent, object, instrument*) as conceptual relations is considered a good style of conceptual design because it enables to address systematically most domain entities of interest [3]. Given some application area, e.g. medicine and healthcare, the natural choice for acquisition of *concepts* is to juxtapose concepts to important terms (most often nouns); however, there could be a variety of approaches and application-dependent considerations regarding the definition of *relations*. This is clearly seen in the largest collection of medical terms UMLS: it comprises more than 100 nomenclatures, controlled vocabularies and terminology systems with over 5 million concept names; the UMLS Metathesaurus is organised by concepts but the most often relations are the 'classical' *IS_A* and *part-of* (even these relations are not always encoded). Many available 'properties' convey either very general relationships or relationships that are hard to interpret in the NLP context [4]. In this way automatic acquisition of relations is a hot research task, especially in large domains where manual elicitation is almost impossible, and there is a variety of application-specific solutions to explicate some of the numerous relations existing between the concepts.

Usually we assume that conceptual relations between entities can be (semi-) automatically acquired by *(i)* automatic identification of *linguistic* relations between the corresponding terms in some text descriptions, and *(ii)* filtering and refinement of the linguistic relations in the process of their interpretation as *conceptual* relations. Actually the text-based acquisition of domain entities is applied for concept elicitation as well, e.g. [5] describes a plug-in OntoLT for the widely used Protégé ontology development tool that supports the interactive extraction and/or extension of ontologies from text. The linguistic analysis is integrated with ontology engineering through the definition of mapping rules that map linguistic entities in annotated text collections to concept and attribute candidates (i.e. Protégé classes and slots). In this way a shallow ontology for the neurology domain was derived from a corresponding collection of neurological scientific abstracts. Focusing on relation extraction, the system RelExt extends an ontology by automatically identifying highly relevant triples (pairs of ontology concepts connected by a relation) from a domain-specific text collection. RelExt works by extracting relevant verbs and their grammatical arguments (i.e. terms) and computing corresponding relations through a combination of linguistic and statistical processing [6]. A system with similar name - RelEx - supports relation extraction from free text [7]. RelEx is based on NL preprocessing producing dependency parse trees and applies a small number of simple rules to these trees. RelEx was evaluated on a comprehensive set of one million Medline abstracts dealing with gene and protein relations and extracted approximately 150,000 relations with an estimated performance of both 80% precision and 80% recall [8].

Regarding the automatic processing of relations in UMLS, [9] analyses the potential of using ontological relations to produce correct semantic structures for a medical document automatically. Presenting a method called SeReMeD, the article discusses an approach to generate representations of unstructured medical narratives. The method makes use of UMLS concept relations and UMLS Semantic Network (SN) semantic types to acquire additional semantic relations and support the structuring process. The results show that the relations can enhance and ameliorate the automatically generated semantic structures.

There are a lot of studies for relation extraction but we are focused on extraction from short medical definitions and elaboration of constraints (e.g. filtering rules) that might help to refine and interpret the discovered relations. Therefore, we have studied approaches for relation processing as well. The article [10] presents a method for relation filtering and a method to discover new relation instances that were developed in the context of cross-language information retrieval (CLIR) and exploit semantic annotation, particularly semantic relations, in the medical domain. As the baseline for automatic semantic annotation [10] uses the existing semantic relations between medical concepts in UMLS. Both methods were applied to a corpus of English and German medical abstracts and evaluated for their efficiency in CLIR. Results show that filtering reduces recall without significant increase in precision, while discovery of new relation instances indeed proved a successful method to improve retrieval. Another article suggesting helpful hints is [11]; it reports about experiments for identifying and evaluating context features and machine learning methods to identify medical semantic relations in texts (more precisely Medline abstracts). Using hierarchical clustering the authors compare and evaluate the linguistic aspects of relation context and different data representations. Through feature selection on a small data set they show that relations are characterised by typical context words, and by isolating these they can construct a more robust language model representing the target relation.

As shown above, to accomplish the task of relation extraction from biomedical texts, researchers use wide-ranging techniques that take advantage of domain, statistical, and linguistic information. Although some studies focus on only one technique, the majority integrate multiple methods to accomplish their aims. These experiments leverage the power of statistical pattern matching but also integrate linguistic characteristics and expert knowledge; this combination is essential to the holy grail of biomedical natural language understanding [12].

The environment OntoLT [5], dealing with ontology extraction from text, proposes a precondition language for defining *mapping rules*. Preconditions are implemented as XPATH expressions over the XML-based linguistic annotation. If all constraints are satisfied, the mapping rule activates one or more operators that describe in which way the ontology should be extended if a candidate is found. We find this idea useful for the elaboration of our approach to relation extraction.

## 3 Prerequisites

We apply a pipeline of existing open source NLP tools and new software components to process automatically various linguistic and conceptual resources. The background is presented here.

As a terminological framework to support our relation extraction task we employed the UMLS Metathesaurus. The advantage of UMLS as a terminological bank is that it contains well documented, consistently structured information, moreover all resources have the same internal representation in RRF text format and are easy to process. In addition there are tools which support the browsing and extraction of term-related information. We used the Metathesaurus for two purposes: *(i)* to filter more carefully the terms that we want to process and to select a concept label for them, and *(ii)* to extract term definitions, which are short paragraphs written in a domain-specific language (in contrast to longer Wikipedia articles where the definitions are written in a popular style).

Another tool we applied out of the box is the relation extractor RelEx [7]. RelEx is a syntactic dependency extractor and semantic framing generator; it parses English language sentences and returns the dependency relationships between different parts of the sentence, and also provides semantic framing tags based on syntax and semantic categories. The core component extracts the dependency relationships. Additional modules perform functions such as anaphora resolution and provide semantic frame output. The Link Grammar Parser[15] is the underlying engine, providing the core sentence parsing ability. Wordnet [18],[19] is used to provide basic English morphology, such as singular versions of (plural) nouns, base forms (lemmas) of adjectives, adverbs and infinitive forms of verbs. Dependency grammar, as formulated by Lucien Tesnière, was one of the influences on the development of conceptual graphs and related versions of semantic networks and similar graph representations for syntax and semantics. Dependency parsers and the related link parsers are useful for generating graphs that have a simple mapping to CGs. In the next section we show examples of parse trees produced in LinkGrammar. We stress that any parser labels the links between the sentence objects only by names of syntactico-semantic linguistic relationships which reflect the sentence structure and general linguistic knowledge about semantic connections between sentence phrases; in other words no domain-specific relations appear as tags in any parse tree.

Our vocabulary of interest is constructed by a bottom up approach: we have analysed automatically the patient status in a corpus of 1200 hospital PRs of diabetic patients and have extracted important clinical terms. The extraction was done semi-automatically starting with an initial list of often used terms and augmenting it iteratively by an expert. The final set of terms was translated to English (where possible more than one translation was given) and the English nouns were lemmatised. The terms denoting diabetes complications were justified using the diabetes ontology at the Biomedical portal [13]. Please note that extraction of UMLS definitions is only possible if the exact (up to lemmatisation) UMLS entry is specified, therefore the English terms have to be represented in their UMLS format.

Hence, the next preprocessing step includes defining the final term list in a sense of UMLS semantic concepts - the terms were turned to single words or phrases of 2-3 words. This task was supported by continuous browsing of the UMLS resources. By extracting the concept identifiers from UMLS we actually checked the availability of the term in the Metathesaurus. Once a term was found in the Metathesaurus, an expert manually selected which of the corresponding concepts are of interest for our study.

In our opinion many important domain-specific relations are rarely declared explicitly in single, well structured sentences (to be automatically extracted from there with the algorithms of the present NLP tools). Therefore we also looked for sources providing hints and insights about human experts' perspective to possible relationships between medical concepts. One such source is the list of relations in the UMLS Semantic Network [16]. These 52 relations are not explicitly encoded between most UMLS concepts but they present an instance of an expert perspective how medical concepts might be interconnected. Having such a list at hand simplifies a bit the relation extraction task because one knows what should be extracted; e.g. we can aim at the extraction of *affect* or *cause* from the definition sentence.

# 4 Extracting Relations from UMLS Definitions

## 4.1 Definition Extraction

The text corpus in our experiment is a collection of term definitions extracted from the UMLS Metathesaurus. Each term in UMLS can name one or more concepts. Each concept is given a unique identifier for the entire UMLS database - the so-called Concept Unique Identifier (CUI). Each concept encodes a different meaning for certain term and the concept definition is given in some source vocabulary (within the hundred resources integrated under UMLS). Sometimes a concept can be encoded without any text definition. Often terms name more than 2 concepts (up to 5) and some concepts might have up to 8 text definitions in the various resources; obviously not all of them are subjects of our study. We used remote access to the UMLS servers provided by the UMLS Terminology Services (UTS) API [17] and extracted all possible definitions for the terms of interest. These are standard procedures provided by the UTS.

Table 1 presents sample definitions of concepts obtained when submitting the term "pulse" to the UTS service. The result contains the concepts' CUIs, in this case *C0391850* and *C0034107*, their corresponding labels "*Physiologic pulse*" and "*Pulse taking*" and the definitions. The first concept has only one definition and the second one has two definitions, extracted from various UMLS vocabularies. Queries to the UTS were sent for all terms of interest and after the definition extraction, an expert filtered out only the definitions of interest which are suitable for the analised conceptual subset. For instance two definitions were selected from Table 1: *Def1* for "*Physiologic pulse*" and *Def1* for "*Pulse taking*" but *Def2* is removed because it is irrelevant for our experiment. Where definitions contain more than one sentence we analyse only the first one of them as our assumption is that it carries the most important information defining the concept.

**Table 1.** Extracted definitions for the term "pulse" from UTS

| Input term | CUI | Output term | Definitions |
|---|---|---|---|
| pulse | C0391850 | Physiologic pulse | **Def 1.** The rhythmic wave within the arteries occurring with each contraction of the left ventricle. |
| pulse | C0034107 | Pulse taking | **Def 1.** The rhythmical expansion and contraction of an ARTERY produced by waves of pressure caused by the ejection of BLOOD from the left ventricle of the HEART as it contracts.<br><br>**Def 2.** Actions performed to measure rhythmical beats of the heart. |

### 4.2 Definition Parsing and Relation Explication

The system RelEx, which is employed as a basic tool at this stage, works in the following way:

*Step 1:* Executes the Link Parser and converts the Link Parser output to a feature structure representation;
*Step 2:* Executes a series of Sentence Algorithms which modify the feature structure;
*Step 3:* Extracts the final output representation by traversing the feature structure.

From the RelEx output we use both the dependency parse of the definition sentence and the semantic framing relations shown on figure 2 and combine them in a conceptual structure which is further used for inference of new relations. We apply further application-specific rules on the text, which improve the relation elicitation process. The rules are developed by studying the available corpus from task-specific perspective. The parser performance at Step 1 is much better after the following transformations are done (examples are shown on table 2.):

*(i)* First letter of each sentence was capitalised;
*(ii)* Dot was put in the end of each definition which lacked it;
*(iii)* Examples were removed from the definitions;
*(iv)* Special mark-ups pointing to source vocabularies were removed;
*(v)* Pronunciation transcriptions were removed;

**Transformations of the syntactic structures.** At step 1 RelEx came across few obstacles, mostly due to the very complex or too simple syntax of the analysed definitions. Only 34% of the definitions were completely parsed and the rest 66% failed in recognising at least one of the words. Since parsing is build

**Table 2.** Normalisation of the extracted UMLS definitions

| Original Definition | Corrected Definition |
|---|---|
| **(eh-DEE-ma)** Swelling caused by excess fluid in body tissues. | Swelling caused by excess fluid in body tissues. |
| A blood vessel that carries blood away from the heart. **(NCI)** | A blood vessel that carries blood away from the heart. |
| **s**welling from excessive accumulation of serous fluid in tissue. | **S**welling from excessive accumulation of serous fluid in tissue. |
| The controlled release of a substance by a cell. [**GOC:mah**] | The controlled release of a substance by a cell. |

around the sentence verb, definitions which are lacking a verb are not juxtaposed a parse tree and respectively no semantic framing was performed(see example 1).

*Example 1.* (LIMB) A body region referring to an upper or lower extremity.
(MUSCLE) One of the contractile organs of the body.
(PHALANX OF HAND) A bone of the hand.

We transformed these phrases to sentences by adding in the beginning the term which is defined followed by the verb "to be". Thus the sentences from example 1 were changed to:

Limb is a body region referring to an upper or lower extremity.
Muscle is one of the contractile organs of the body.
Phalanx of hand is a bone of the hand.

Parsing failures are also due to improper punctuation and structure like in example 2. We also noticed that the parser sometimes fails to disambiguate verbs/nouns for words like e.g. measure, joint, etc.; sometimes nouns are parsed as verbs which is the case with the sentences in table 3.

*Example 2.* Touch; the faculty of touch, the sensation produced by pressure receptors in the skin.

**Table 3.** Wrong disambiguation of verbnoun POS tags

| Sentence | Constituency Parse Tree |
|---|---|
| Touch; the faculty of touch, the sensation produced by pressure receptors in the skin. | (S (VP touch [;] (NP (NP the faculty) (PP (NP of touch) , (NP (NP the sensation) (VP produced (PP by (NP pressure receptors)) (PP in (NP the skin))) .))))) |
| A joint connecting the lower part of the femur with the upper part of the tibia. | (S [a] (S (VP joint [connecting] (NP (NP the lower part) (PP of (NP the femur))) (PP with (NP (NP the upper part) (PP of (NP the tibia)))))) .) |

The final output of RelEx includes constituency and dependency parse trees (including the set of all dependency relations and features), a link grammar parse tree and relations determined by semantic framing rules, such as for identifying the discourse entities:

1_Entity:Entity(Diabetic_Cataract, Diabetic_Cataract)

The upper part of figure 2 contains two RelEx outputs for the sentence *Diabetic Cataract is a rare, usually bilateral, opacity shaped like a snowflake, affecting the anterior and posterior cortices of young diabetics.*

### 4.3 Adding New Relations to the Original Relation Set

#### IS_A Relation

Using the dependency graph, the regular structure of the narrative sentence - Subject Verb Object (SVO), the fact that the definitions are written in a very short form, and following a commonly accepted rules about meaning composition, we are able to infer the "IS_A" relation out of some parsed definitions. After transforming to sentences the definitions-phrases like the ones in example 1, some 10% of the definitions have exactly this structure.

*Rule 1.* If we have the dependency relations corresponding to subject, object, and an object and/or subject modifier in the main sentence marked as:

_obj(be, N3)
_subj(be, N1)
_tense(be, present)
_nn(N1, N0)
_nn(N3, N2)

then we can infer that N1 IS_A N3. The last two relations are optional and if present N0 would be modifier to N1 and N2 to N3 correspondingly.

*Example 3.* Given the definition: "*Artery is a blood vessel that carries blood away from the heart.*" The extracted relations of interest are shown on the graph on figure 1 as solid and the inferred IS_A relation is marked dashed.

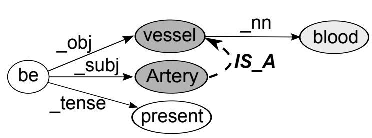


**Fig. 1.** Dependency graph explicating the inference of IS_A relation by Rule 1

#### AFFECTS Relation

Another important relation is the AFFECTS relation, especially between complications/diseases/symptoms and the organs they affect. We consider examples extracted from definitions of Diabetes Mellitus complications. The RelEx syntactic and semantic analysis delivers the entities occurring in the definitions for further processing of each definition. The lexico-semantic structure, needed

***Constituency parse***

(S (NP Diabetic Cataract) (VP is [a] (NP (NP (ADJP rare ,)
(ADJP (ADVP usually) bilateral ,) opacity) (VP shaped (PP like (NP a snowflake
(VP , affecting (NP (NP the (ADJP anterior and posterior) cortices)
(PP of (NP young diabetics))) .)))))))

***LinkGrammar parse***

```
                    +-----------------------Out------------------+
                    |        +-----------------A----------------+
                    |        |                      +-------A------+
  +-----G----+--Ss--+        +-Xc+      +--Ec--+--Xc--+       +------M
  |          |      |        |   |      |      |      |       |
Diabetic Cataract is.v [a] rare.a , usually bilateral.a , opacity.n-u


                                        +---------------------------------
                                        +---------------------Op--------
                                        |       +----------------Dmc---
               +-----Js-----+-----MXsp----+      |                   +----
v-----+--MVp--+    +--Ds--+       +--Xd--+       |          +--AJla--+---A
      |       |    |      |       |      |       |          |        |
shaped.v-d like.p a snowflake.n , affecting.g the anterior.a and.j-a


---------------Xc----------------------------------------+
-----------------------+                                 |
-----------------------+                                 |
----------A------------+       +-------Jp--------+       |
Jra---+                +-Mp-+      +-----A----+       |
      |                |    |      |          |       |
posterior.a cortices[!].n of young.a diabetics.n .
```

***Dependency tree***

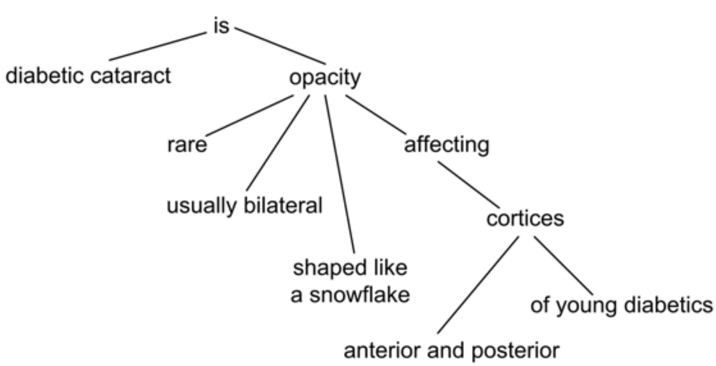


**Fig. 2.** RelEx outputs and illustration of the dependency skeleton supporting relation elicitation

for relation extraction, is elaborated in five steps which are illustrated here for the sentence *Diabetic Cataract is a rare, usually bilateral, opacity shaped like a snowflake, affecting the anterior and posterior cortices of young diabetics.*

**Step 1:** Construct a list of the entities obtained from RelEx.
**Entities** {Diabetic_Cataract, opacity, snowflake, cortex, affecting, Diabetic, diabetic, like}

**Step 2:** Augment each entity with its possible modifiers collected at the parsing stage. These could be adjectives or nouns in the role of adjectives, forming a noun phrase, or compounds connected by preposition extracted from the parser (e.g. *accumulation of amount*). Skip the modifiers which are stop words (e.g. *and, or, due etc.*).
**Entities extended with modifiers** {Diabetic_Cataract, opacity, bilateral opacity, rare opacity, snowflake, shape like snowflake, cortex, *and cortex*, affecting, *Diabetic*, diabetic, young diabetic, like}
At this step the following terms were subtracted: *and cortex*, because *and* is a stop word; *Diabetic*, because it is already in the list and *Diabetic_Cataract* is transformed to *Diabetic Cataract*.

**Step 3:** Search in UMLS for the words and compounds from *Step 1* and *Step 2* in order to prove which of them are medical terms. Extract the terms which have the least Hemming distance to the initial term.
**Entities recognised by UMLS as medical terms** {Diabetic Cataract (Diabetic Cataract), Retinal opacity (opacity), Snowflake retinal degeneration (snowflake), Visual Cortex (cortex), affecting (affecting), diabetic (diabetic)}

**Step 4:** Assign the UMLS semantic type to the terms extracted at *Step 3* given by UMLS and obtain the following figure.
**Semantic type attachment** (semantic types are in the brackets): {Diabetic Cataract (*Disease or Syndrome*), Retinal opacity (*Finding*), Visual Cortex (*Body Part, Organ, or Organ Component*), snowflake (*Acquired Abnormality*), affecting (*Functional concept*), diabetic (*Finding*)}

**Step 5:** Construct a new lexico-semantic representation corresponding to the initial definition and including the semantic types.
**Lexico-semantic representation**: (Disease or Syndrome) is a rare, usually bilateral, (Finding) shaped like a (Acquired Abnormality), affecting the anterior and posterior (Body Part, Organ, or Organ Component) of young (Finding).

**Step 6:** Infer the AFFECTS relation and its arguments given the features collected on the previous steps and the dependency information available from the RelEx output.

In contrast to the "bag of words" approach, these steps give us the advantage to deal at *Step 6* with concrete medical terms, to have their possible modifiers,

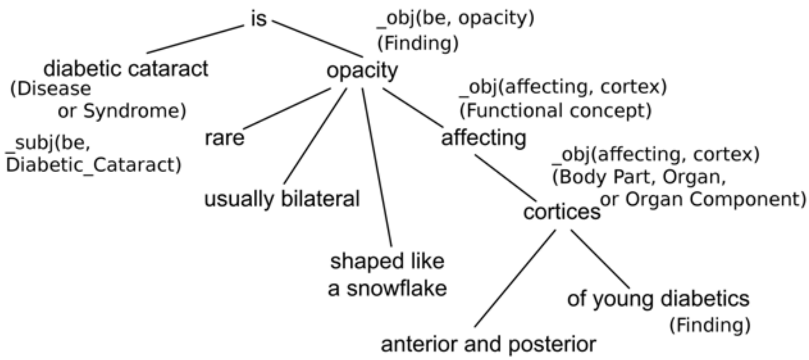


**Fig. 3.** Feature integration and relation selection

and a dependency representation helping to resolve the concept which is being affected. Since our approach is bottom-up and we are looking for concrete relations between the diseases and the organs they affect, and given the fact that the analysed text is a definition, the dependency relation *subject* is of crucial importance to us. That is why at this final step we organise all features selected in the previous step as shown on figure 3.

*Rule 2*. If we have dependency and semantic structures matching the following conditions:

_subj(be, N1) AND N1 has Semantic Type *Disease or Syndrome*
_obj(be, N2) AND N2 has Semantic Type *Disease or Syndrome* or *Finding*
A verb V singling AFFECT relation
_obj(V, N3) AND ( N3 has Semantic Type (*Body Part, Organ, or Organ Component*) OR the augmented entity of N3 identifies has Semantic Type *Body Part, Organ, or Organ Component*

Then we infer by this rule that N1 AFFECTS N3
and by *Rule 1* that N1 IS-A N2.

E.g. *Diabetic Cataract* IS-A *opacity* AND *Diabetic Cataract* AFFECTS *cortices.*

## 5 Discussion

In this experiment we parsed 194 definitions corresponding to 129 terms. Lexical constructions referring to Body parts or Organs were found and further analysed in 57% of the definitions. The analysis of the constituency parsing proved that the focal terms (organs/body parts) are always located in a prepositional phrase attached to the verb of interest. We studied the patterns matching the AFFECTS relation and made a list of verbs expressing the availability of this relation.

The lexical expressings occur in active or passive voice. Most frequent lexical patterns are: *affecting*, *consisting of* and *characterized by*. The phrase "resulting from" signals the availability of another disease or condition, which triggers the disease of interest.

The extraction of the IS-A relation was done with 81% precision, whereas in the subset of definitions with automatically transformed syntactic structure this result is as high as 89%. There are several reasons for the lower performance over the whole collection; the first one is wrong parse trees, due to the complicated syntactic structure; another one is the partial recognition of compound terms which were mapped to subject and/or object which revents the relation recognition. The recall for definitions with explicitly stated IS-A relations is 86%.

The extraction of the AFFECTS relation performed, as expected, worse than IS-A extraction, because the variety of affect expressions is much higher. Another complication is due to the fact that the arguments of AFFECTS may be positioned in the sentence in longer distance from each other, thus the parsing errors imply incorrect inferences by Rule 2. Some of the words signaling AFFECTS are ambiguous and lead to a different relation.

## 6 Conclusion

Our goal in this study was to prove the availability of NLP tools and resources as well as their readiness to serve for developing of new conceptual resources. We presented an approach enabling automatic extraction of relations between medical concepts by reusing existing resources and NLP tools and applying additional transformation rules. By analysing the output in the intermediate steps we notice that often the failure of our algorithm is due to wrong parsing trees.

The IS-A relations we extracted were often available in the UMLS Metathesaurus, but some 45% are newly created. As for AFFECTS, only 3 of the extracted relations were available in the Metathesaurus and they were not concretely specified, but available as concept relations *has relationship other than synonymous, narrower, or broader*. While analysing the results in the intermediate steps we noticed that often the failure of the algorithm was due to an error in the parsing stage. Therefore a better parsing would lead to improvement in the relation extraction algorithm as well.

In the future we plan to do a more thorough evaluation of the learned rules and go into detail of the extraction relations towards specifying their different subtypes such as *functionally_related_to manages*, *treats*, *disrupts*, *complicates*, *interacts_with*, *prevents*.

**Acknowledgments.** The research work presented in this paper is partly supported by grant DO 02-292/December 2008 "Effective search of conceptual information with applications in medical informatics", funded by the Bulgarian National Science Fund in 2009-2012 [2].

## References


1. Terminological Services of Unified Medical Language System (UMLS), https://uts.nlm.nih.gov/home.html
2. Project Effective search of conceptual information with applications in medical informatics, http://www.lml.bas.bg/evtima
3. Sowa, J.: Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, Reading (1984)
4. Nirenburg, S., McShane, M., Zabludowski, M., Beale, S., Pfeifer, C.: Ontological Semantic Text Processing in the Biomedical Domain. University of Maryland Baltimore County, Institute for Language and Information Technologies, Working Paper, 3–5, http://naboo.ilit.umbc.edu/ILIT_Working_Papers/ILIT_WP_03-05_Biomed_Mesh.pdf (last visited February 2011)
5. Buitelaar, P., Olejnik, D., Sintek, M.: A Protg Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 31–44. Springer, Heidelberg (2004)
6. Schutz, A., Buitelaar, P.: RelExt: A Tool for Relation Extraction from Text in Ontology Extension. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., et al. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 593–606. Springer, Heidelberg (2005)
7. RelEx, https://launchpad.net/relex
8. Fundel, K., Kffner, R., Zimmer, R.: RelEx - Relation extraction using dependency parse trees. Journal of Bioinformatics 23(3), 365–371 (2007)
9. Denecke, K.: Enchancing Knowledge Representations by Ontological Relations. In: Andersen, K., et al. (eds.) Proc. of MIE 2008, eHealth Beyond the Horizon - Get IT There, Goteborg. Studies in Health Technology and Informatics, vol. 136, pp. 791–796. IOS Press, Amsterdam (2008)
10. Vintar, S., Buitelaar, P., Volk, M.: Semantic Relations in Concept-based Cross-language Medical Information Retrieval. In: Adaptive Text Extraction and Mining (ATEM), Cavtat-Dubrovnik (2003)
11. Vintar, S., Todorovski, L., Sonntag, D., Buitelaar, P.: Evaluating context features for medical relation mining. In: ECML/PKDD Workshop on Data Mining and Text Mining for Bioinformatics (2003)
12. Chapman, W., Cohen, K.B.: Current issues in biomedical text mining and natural language processing. Journal of Biomedical Informatics 42(5), 757–759 (2009)
13. Diabetes Ontology at the BioPortal provided by the Medical Dictionary for Regulatory Activities Terminology (MedDRA), http://bioportal.bioontology.org/visualize/42280/?conceptid=10012614 (last visted February 2011)
14. Browne, A.C., Guy, D., Aronson, A., McCray, A.: UMLS language and vocabulary tools. In: Proceedings Annual Symposium AMIA p. 798 (2003)
15. Link Grammar Parser, http://www.link.cs.cmu.edu/link/
16. Current Relations in the UMLS Semantic Network, http://www.nlm.nih.gov/research/umls/META3_current_relations.html
17. Terminological Services of Unified Medical Language System (UMLS), Developer's Guide, https://uts.nlm.nih.gov//doc/devGuide/index.html
18. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 38(11), 39–41 (1995)
19. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)