Abstracts Booklet

Doctoral Consortium
at the 2005 Semantic Mining Summer School

Tihany, Hungary

June 30 & July 1, 2005

Organized by Marian Petre and Paul Piwek

The Mobility Programme (WP6) of the EU Network of Excellence in Semantic Interoperability and Data Mining

Contact: Mcs-Semantic-Mining-WP6@open.ac.uk

Contents

1 Hybrid Mappings of Complex Questions over an Integrated Semantic Space

Gaston Burek

Institution The Open University

Email G.G.Burek@open.ac.uk

Stage of Studies PhD project in its second year within a three years program

Abstract In order to facilitate sharing and processing of online information sources researchers have created the vision of the Semantic Web (Hendler et al, 2002). This initiative formalises knowledge by marking up documents using ontologies (Noy and McGuiness, 2001) that provide semantics and structure to data. Integrating all that information requires making sense of the different terminology used within the various sources and exposes all the problems related to the terminology gap: the fact that concepts and their semantic relations can be realised in different ways. In our work we seek to address those problems by calculating the semantic similarity between ontologies and text within a narrow domain Question Answering framework.

Our research explores term-concept dimension for solving the problem of mapping between binary semantic relations integrated within a semantic space and a set of questions and answers. Those relations can be formalised by ontologies as attributes of classes and by natural language expressions describing an interaction between two or more concepts.

To integrate the relations within the semantic space we can quantify their similarity by comparing terms associated to the relation extensions but the problem with this approach is that two ontologies may use different name and extensions to describe the same relation. Thus to quantify similarity we need to reason about the uncertainty in the similarity.

We use Latent Semantic Analysis (LSA) together with the Cosine Similarity (Deerwester et al, 1990) to measure similarity between term frequency vector representations of queries derived from complex questions, extensional representations of semantic relations belonging to ontologies (i.e. classes , relations, and instances) and text. This methodology for measuring semantic similarity calculates distance between the vectors by means of finding different orders of term co-occurrence. We also design a series of experiments to investigate whether the mapping method can be used to a) relate ontologies to queries and text, b) to measure the distance between ontologies describing the same domain and c) to represent relation directionality by means of asymmetrical term weighting. Results of our experiments show that LSA can indeed be used in that way.

Our work makes several research contributions. First, the use of LSA to measure semantic similarity between structured information sources (i.e. ontologies)

and no-structured (i.e. text documents) ones. Second, the introduction of uncertainty and directionality when mapping concepts and their semantic relations by means of using a probabilistic method such as LSA that takes into account frequencies of terms. Finally, the third contribution is the possibility of capturing implicit similarity between concepts or semantic relations; this capability is provided by the Single Value Decomposition (SVD) (Press et al, 1992) method incorporated within LSA.

References

Deerwester,S.C., Dumais,S.T., Landauer, T. K., Furnas, G. W., Harshman, R.A. (1990). Indexing by Latent Semantic Analysis. JASIS 41(6): 391-407.

Hendler J., Berners-Lee T. and Miller E., Integrating Applications on the Semantic Web. (2002). English version. Journal of the Institute of Electrical Engineers of Japan, Vol 122(10), October, 2002, p. 676-680.

Noy, N.F. and McGuiness, D.L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March.

Press,W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1992). Singular Value Decomposition. 2.6 in Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd ed. Cambridge, England: Cambridge University Press, pp. 51-63.

2 Identifying Concepts For Semi-automatically Building Ontologies From Natural Language Text

Dileep G. Damle

Institution The Open University, UK

Email d.g.damle@open.ac.uk

Stage of Studies Halfway

Abstract significant words and multi-word terms is widely regarded as an important first step in extracting an ontology from a domain specific collection of documents. Different statistical approaches validating extracted terms are examined over three different document collections within scientific and technical

domains. Distributions of lemmas over each target corpus are compared with the distributions of those lemmas over a general balanced corpus, the British National Corpus. The relevance of document length in calculating relative frequencies used as the statistic in such comparisons is compared with an alternative way of defining relative frequencies. Results indicate that The difference between Poisson and Binomial distributions make only slight differences in just one corpus, while the new method provides significantly different results for all experiments.

After examining the appropriateness of three different approaches to term extraction, further experiments are carried out using WordNet senses as proxy for concepts. It has been said that (Morris and Hirst) that lexical chains (Halliday and Hasan) may be useful for word sense disambiguation. However, it is proposed to use a network of WordNet senses and their various relations (lexical, semantic and ontological) to identify the ontology structure. This process would be considerably aided if the sense each word is previously identified accurately. This seems a chicken and egg situation, so it is proposed to attempt to both disambiguation and ontology extraction as mutually reinforcing processes.

3 A CLIR interace to a Web Search Engine

Philipp Daumke

Institution Freiburg University Hospital, Germany

Email daumkep@web.de

Stage of Studies Second year of two years (PhD)

Abstract I develop a query construction tool by which a standard Web search engine (e.g. Google) is enabled to process native language queries from the biomedical domain in order to retrieve documents in another, user-defined target language (Cross-Language Information Retrieval).

The underlying methodology uses a special type of dictionary which entries consist of subwords. Subwords are grouped into equivalence classes (represented by Morpheme identifiers (MIDs)) which capture intralinugal as well as interlingual synonymy. Based on subword lexicons in six different languages, a morphosemantic indexing procedure (MSI) extracts subwords from textual input and maps them to their associated MIDs.

In the training phase, words, word bigrams and trigrams from domain and language specific corpora are extracted and subsequently processed by MSI (target data). Afterwards, a user can enter queries on a web interface (www.morphosaurus.net) and specify his query language and favoured target language. Again, these queries are transformed to a set of corresponding MIDs. These representations are matched against the language-specific target data in order to get (the most

frequent) word translations. These matching records are used to generate target queries by applying several combination heuristics.

Finally, these queries are sent to a standard Web search engine. The claim that the interlingual approach is useful for the purpose of cross-language information retrieval and text categorization has already been experimentally supported.

4 Identifying Concepts For Semi-automatically Building Ontologies From Natural Language Text

Natalia Grabar, Magali Sillam, Marie-Christine Jaulent

Institution SPIM/U729, Inserm, Paris

Email natalia.grabar@spim.jussieu.fr

Stage of Studies Natalia Grabar: PhD 2004, first months of bioinformatic project

Abstract The aim of our project is to extract automatically, from the scientific literature and existing databases, semantic annotation for Drosophila fly and C. elegans worm genes.

Biological databases (WormBase, FlyBase, GeneOntology) and expert knowledge are used to extract this semantic annotation. In particular, GeneOntology is used as source for normalized description of genic expressions. The annotation corresponds to molecular function, biological process or cellular component. The objective is then classify the genes of these two species into homogeneous clusters. In this way, we can observe genetic conservation and developmental diversity of the studied species. The classes of genes obtained from semantic annotation from texts will be compared with clusters obtained from DNA microarrays analysis.

To handle the different issues, we adopt the following methodology:

1. corpora building and pretreatments (lemmatizing, term normalization, ...)
2. annotating corpora with entities and terms (gene, patterns, functions, ...)
3. mining the annotated corpora in order to find genes and their expressions. The result of this step is a set of genes associated to specific information
4. weighting informations for each gene

5. classification of genes according to the information they are associated with

6. comparing the clusters with already existing ones

The first step of the work is to produce a textual corpora relevant for the domain. Existing databases are already rich in information, but not exhaustive. We make use of different database fields in order to collect more scientific literature sources: - Gene names and their synonyms are the key-words for collecting more literature. - Some fields contain already encoded expressions and characteristics about genes and their allels. This information is directly useful for the semantic annotation of genes. - The summary of already compiled bibliographic references is contained in free-text fields. - And finally, already compiled bibliographical references provide us with additional key-words (author names and year of publication), which can be useful for the collection of more scientific literature.

The first survey of databases and corpora shows that all the genes are not investigated with the same interest. For some of them dozens of references are available, and none for others. For instance, among 96 genes studied, 48 Drosophila and 48 C. elegans genes, 8 exist in FlyBase (CoVa, l(2)35Di, CG3560, CG4692, ct, bic, da, E2f) but no worm genes from our set are known in WormBase.

Quering the PubMed with gene names, and their synonyms when they are available, for scientific abstracts gives information about 43 genes, mainly Drosophila genes. But these abstracts must be filtered before processing because of the ambiguity of certain gene names.

During the second step, we recognize entities (gene names) and their relations (genic expression). GeneOntology provides hierarchies of normalized terms for the description of genic expressions: molecular functions, biological processes and cellular components. But authors do not use currently the same expressions. We apply a set of rules performing on character, morphological and lexical levels to normalize author expressions and match them with GeneOntology terms.

In the same way, gene names provided by databases may vary in scientific literature: conventions about these names seem not to be stable and not really respected. Rules needed for their recognition are different from those needed for GeneOntology terms recognition. They perform on character and punctuation levels and handle abbreviations.

In the third step, all the additional literature and free-text fields are analyzed to extract gene names and their expressions. We combine two semantic mining methods: symbolic (association rules, likelihood ratio) and syntactic (lexical and syntactic patterns, syntactic parsing).

Biological texts and especially entities are known to be exceedingly ambiguous. The two methods we apply have different requirements. Syntactic approaches rely on the contextual analysis of gene names and their expressions. With these approaches it is possible to use all the gene names and their synonyms. But when using symbolic methods it is preferable to use only non am-

biguous gene names: those which do not appear in a general language dictionary and which are not used to name genes from different species.

Applied extraction methods propose informations about the gene profile:

[species, gene, typeOfExpression, expression]

This profile is used to calculate semantic proximity between genes and to classify them.

We are currently working on the three first steps.

5 Selecting target words in direct machine translation

Maria Holmqvist

Institution Linköping Universitet, Sweden

Email marho@ida.liu.se

Stage of Studies First year in a five-year Ph.D. programme

Abstract A fundamental issue in machine translation (MT) is to use the appropriate target word when translating a source word with several equivalents in the target language. For example, the English word "space" can be translated to Swedish as either "utrymme" (as in "disk space"), "rymden" (outer space) or "blanksteg" (space key). Depending on the context of a particular instance of "space", one of these translations will be more appropriate than the others. My research focuses on how the correct translation of lexical units (words and multi-word compounds) can be decided from contextual information.

At Linkping University, we are currently working on a corpus-based direct machine translation system, called T4F that translates English texts to Swedish (Ahrenberg and Holmqvist, 2004). It's a direct translation system, meaning that the translation takes place on the level of word and multi-word units. It's entirely corpus-based; the dictionary is extracted from word-aligned parallel corpora from a suitable domain e.g. the medical domain or database manuals. The hypothesis is that by training the system on corpus data from a particular domain, domain specific terminology and syntactical constructions should be adequately captured in order to translate new texts from the same domain.

Each entry in the T4F dictionary consists of aligned source and target segments from a syntactically annotated parallel corpus. The dictionary includes word correspondences as well as some linguistic and contextual information to ensure that a source word will be translated correctly according to the context it occurs in. Currently, we apply hand-written rules to ensure that the right

kind of linguistic and contextual information is added to the words in the dictionary. Therefore a human annotator must decide what linguistic and contextual features are necessary for correct translation of each word.

In my research I intend to investigate methods of automatically identifying the contextual features that should be included in the lexicon in order to properly translate words in their context. In the initial phase of the project, I will investigate a selection of corpus-based methods for word sense disambiguation that might also be used on the analogous problem of deciding the correct translation of a word, e.g. the method used by Brown et al. (1991) to improve target word selection in statistical MT.

References

Ahrenberg, L. and M. Holmqvist (2004). Back to the future? The case for direct English - Swedish MT. To appear in Proceedings of RASMAT'04. Uppsala.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer (1991). Word Sense Disambiguation using Statistical Methods. Berkeley, University of California.

6 Cross-Language Information Retrieval (CLIR) for Biomedical Content

Kornel Marko

Institution Freiburg University Hospital

Email marko@coling.uni-freiburg.de

Stage of Studies Last year in three-year programme

Abstract The project is based upon an interlingua-based approach to cross-language information retrieval (CLIR) in the medical domain, in which queries, as well as documents, are mapped onto a language-independent concept layer on which retrieval operations are performed.

The interlingua is based upon equivalence classes of subwords, which capture intralingual synonymy, as well as interlingual translations. Furthermore, machine-learning algorithms are adapted to the needs of cross-lingual document classification (MeSH, ICD, ...). The research implies the elaboration of differences in language-specific medical sublanguages (German, English, Portuguese, French, Spanish, Swedish), the automatic acquisition of multilingual (medical) lexica at the level of subwords, and statistical word sense disambiguation.

The claim that the interlingual approach is useful for the purpose of cross-language information retrieval and text categorization has already been experimentally supported.

7 Thesaurus-Based Index Term Extraction

Olena Medelyan

Institution Freiburg University, Germany

Email medelyan@gmail.com

Stage of Studies Completing Master's project

Abstract Keyphrases are widely used in both physical and digital libraries as a brief but precise summary of documents. They help organize material based on content, provide thematic access, represent search results, and assist with navigation. In some contexts keyphrases are freely chosen; in others they are restricted to terms that occur in a controlled vocabulary or thesaurus. In either case assigning high-quality keyphrases to documents manually is expensive, because for best results professional human indexers must read the full document and select appropriate descriptors according to defined cataloguing rules. There are two approaches to automating the identification of keyphrases. In keyphrase extraction, the phrases that occur in the document are analyzed to identify apparently significant ones, on the basis of intrinsic properties such as frequency and length. In keyphrase assignment, which might more properly be called index term assignment because keyphrases are not constructed as free text but are chosen from a controlled vocabulary of terms, documents are classified according to their content into classes that correspond to elements of the vocabulary. One serious disadvantage of the extraction approach is that the extracted phrases are often ill formed or inappropriate. The assignment approach circumvents this problem, but for satisfactory results a very large and accurate corpus of training material is needed, which must be produced by manually classifying training documents. The conducted research describes a new approach called index term extraction, which is intermediate between keyphrase extraction and term assignment. It combines the advantages of both, while avoiding their shortcomings. A thesaurus is used as the controlled vocabulary, and the relationship between manually assigned index terms and the phrases that actually appear in the document text are explored using the semantics implicit in the thesaurus. A machine-learning model is trained based on features from this relationship, along with other characteristics of extracted phrases that are used in conventional automatic keyphrase extraction. Given a test document, vocabulary terms are assigned to it by mapping all the documents phrases to those in the thesaurus and using the learned model to decide which ones are

significant descriptors of the documents content. The resulting set contains only well-formed phrases from the thesaurus that are strongly related to the given document. Currently an algorithm for index term extraction was tested on agricultural documents by using the domain-specific thesaurus Agrovoc. The evaluation on a 200-items document collection that considered semantic relatedness between index terms revealed that this approach is promising: 50% of automatically assigned terms match conceptually phrases assigned by professional indexers. A comparison to inter-indexer consistency achieved by human indexers on a smaller document set showed that the developed system is only 10-15% less consistent with humans than they among each other.

For the PhD thesis, under supervision of Ian H. Witten from the University of Waikato, New Zealand, an extension of the current system to other domains (and maybe languages) is planned. In particular, the Library of Congress Subject Headings (LCSH) should be explored as controlled vocabulary and thesaurus. Further improvements, such as better topic covering and lexical attraction for word sense disambiguation are planned to be incorporated in the final system.

8 Translation of Medical Terminology Systems and Their Use in Electronic Healthcare

Mikael Nyström

Institution Linköping University, Sweden

Email mikny@imt.liu.se

Stage of Studies In the middle of five-years PhD studies

Abstract The area of the Ph.D. studies is medical terminology systems. My current research question is how semi-automatic methods can be used to extract an English-Swedish medical dictionary from medical terminology systems existing in parallel in both English and Swedish. The subsequent research question is how the generated dictionary can be used to semi-automatically translate large medical terminology system such as SNOMED CT. A later research question is an evaluation of a translated SNOMED CT in the Swedish healthcare system.

The research relates to the knowledge area around medical terminology systems. The works include preparation and evaluation of the systems to be able to use them during the lexicon generation process, and will also later include evaluation of medical terminology system's use in Swedish healthcare. The research also relates to natural language processing for the semi-automatic lexicon generation and semi-automatic translation.

Until now a collection called TermColl has been compiled containing both English and Swedish versions of the medical terminologies ICD 10, ICF, MeSH,

NCSP and KSH97-P. TermColl has been used as a source for semi-automatic creation of a first version of an English-Swedish dictionary. A suite of programs called I*Tools has been used. The I*Tools combine static resources (such as bilingual dictionaries, part of speech patterns across languages, functions for utilising string similarity [cognates] and handling interpunction) with creation of new dynamic domain- and application-specific data through interactive training and machine learning. The I*Tools create links between the two language's words or small phrases. The I*Tools has been interactively trained on 4,200 rubrics and a manual verification of I*Tools proposed word links has been done. The result is a first version of an English-Swedish medical dictionary containing 31,000 entries.

The next step is trying to improve the results of the word alignment. One of the improvements is to filter and split TermColl in partitions to better support the word alignment process before the I*Tools are used. Another improvement is to test different modes on the part of speech tagger.

A further step is to use the generated dictionary with semi-automatic methods for translation of medical terminology systems. The most prioritised system to translate from English to Swedish is SNOMED CT, because with a Swedish version of it we are able to evaluate SNOMED CT under Swedish circumstances.

The methods used in my research are relevant for other west European language pairs than English and Swedish as well.

9 Semantic Mining of Prepositional Phrases in Biomedicine

Michael Poprat

Institution Freiburg University Hospital and Jena University, Germany

Email poprat@coling.uni-freiburg.de

Stage of Studies First year in three-year programme

Abstract Prepositional phrases (PP) usually consist of a preposition (of, from, since ...) and a succeeding nominal phrase (NP). Prepositions relate NPs to preceding nouns, verbs or adjectives causing different semantic interpretations (spatial, temporal, causal etc.). For a natural language processing system, PPs often evoke ambiguities insofar as several possibilities occur, where the PP can be related to.

> Example: '[the removal of the octamer-binding site]/NP reduces/VBZ [transcription]/NP to/TO [levels]/NP below/IN [detection]/NP in/IN [both cell types]/NP'

where the PP 'to/TO [levels]' could be related to the NP '[transcription]' or to the verb 'reduces/VBZ'. Also for the attachment point for the PP 'below/IN [detection]' it is unclear if the PP refers to the NP '[detection]', to the NP '[levels]' or to the verb 'reduces/VBZ'. The same holds for the PP 'in/IN [both cell types]'.

The problem of PP ambiguities is a frequently discussed subject and a lot of solutions both knowledge-based and statistical have been proposed. None of them has been evaluated for the biomedical domain yet. In preliminary studies, I have shown that statistical approaches on PP disambiguation in biomdedical texts yield promising results (about 85% accuracy and 86% coverage). Their performance depends on various parameters such as statistical model, morphological normalization, etc.

One assumption that I follow up in my thesis dealing with text mining in biomedical texts is the following: Before mining the semantic relations carried by prepostional phrases, their disambiguation is an essential prerequiste. Only after the disambiguation, it makes sense to apply further processing steps to discover and interprete PP-related concepts.

To assure the concept-based treatment of PPs I propose to map text entities to concepts encoded in the UMLS vocabulary. One of the aspects of my thesis is to investigate the usefulness of the UMLS for the biomedical domain.

10 Determining semantic interoperability between archetypes and clinical terminologies represented as ontologies

Rahil Qamar

Institution University of Manchester

Email qamarr@cs.man.ac.uk

Stage of Studies Middle (completed 1.5 years of 3 year PhD)

Abstract An unusual feature of health informatics is that clinical terminology models and clinical data models are developed by separate groups which work independently of each other. There are, therefore, inherent differences in the principles on which each of them are based as well disparities in the semantics of representation.

In addition, the number of different clinical situations requiring specialisations of the data model is very large. The model, therefore, has to be flexible enough to avoid writing tens of thousands of specialised bits of code to cope with every data-entry situation as it would become unmanageable. Dealing

with this issue requires that the data models should be able to interact with the underlying clinical terminologies as per requirements in a principled way.

The research is aimed at determining the syntactic and semantic constraints involved when integrating these independently developed clinical terminologies and data models and their resolution.

Research Question What are the syntactic and semantic constraints that need to be resolved to enable the integration of archetypes with ontologies that model clinical terminologies logically?

Research Contribution The research forms part of the Semantic Mining WP26 and CEN 13606 initiatives to address the issue of integrating archetypes with external terminologies such as SNOMED-CT, GALEN, and others.

It researches the heuristics of resolving the integration issues by determining rules and strategies for establishing some sort of semantic correlation between archetypes and ontologies. The research is at the heart of one of the core problem areas highlighted by the WP26 proposal. Likewise, one of the tasks of CEN 13606 and the HL7 Templates community to be resolved is similar to the objective of this research providing more platforms to test out the research outcomes.

Research Methods Some of the methods that have been applied to achieve the objectives of the research are as follows:

- Selected the Headache archetype and GALEN ontology for carrying out the research.
- Ensuring that the archetype and the ontology to which it binds are in the same context of use. This was done by matching the top-level definition of the archetype with its existence in the preferred ontology.
- Parsing the archetype and the ontology to a common language to facilitate semantic interoperability. XML was the chosen language to which the ADL archetype and the OWL ontology were parsed.
- Generating a list of semantic mappings of archetype elements with the ontologies as well as logging its subsumption relation in the ontology hierarchy. Results indicated whether the archetype element was present as a Class, Sub Class, Property or Restriction value within the ontology.
- Using lexical analysis to help with semantic mapping of concept names. Also includes splitting combinatorial terms, ontology mining and ontology mapping techniques.
- Other suitable techniques will be gradually adopted to help approach the research question of semantic integration of archetypes with ontologies by resolving mappings and inconsistencies.

Preliminary Findings

- Concepts like 'diagnosis' might not be present as such in terminologies so might need to match values for probable existence. They should be treated as roles.
- Archetype elements such as 'description' might not exist in ontologies. They should be treated as part of data structures.
- Ontology segmentation might not be a good approach especially when elements not directly related to the archetype concept are used for recording data. In such a case the entire ontology might need to be mined for results. This approach might involve high processing time when very large ontologies with several million concepts and relationships are to be mined.
- Replacement of earlier results found when better results are obtained later on.

Conclusion More specific findings have been logged as a result of Phase 1 of the research process. A demonstration accompanied by explanation of the results might prove beneficial to the audience.

11 The Semantic Synapse Project: Accelerating Neuroscientific Research by Semantic Web Technologies

Matthias Samwald

Institution University of Vienna

Email samwald@neuroscientific.net

Stage of Studies -

Abstract

Introduction The proposed project aims to adapt new Semantic Web technologies for the use in neuroscience, biomedicine and bioinformatics. Specifically, these innovative technologies will be used to build an internet portal through which molecular interactions of the synapse can be explored.

Rationale Neuroscientific research is increasingly hampered by the way scientific data and information is stored and communicated. Information is often presented in forms that are hardly machine processable, databases are poorly connected to each other. As a consequence, researchers in data - intensive disciplines like biomedicine or neuroscience find it increasingly difficult to keep track of relevant information from different system levels. The project will point out ways in which these problems and limitations can be overcome by Semantic Web technology, based on new World Wide Web data standards. These standards allow making information and its logical interconnections explicit and machine processable.

Aims The interactome of the synapse is a subject that is of great importance for the understanding of the nervous system in health and disease, as well as drug development. To make efficient research and development possible, a lot of heterogeneous information has to be combined and integrated. This project will apply Semantic Web technologies to solve this task. Data and Information relevant to the subject will be converted to formats that are compatible with Semantic Web standards. This information will be made openly accessible through a user - friendly portal that allows querying, visualizing and exploring the underlying information. However, the portal will not be designed as a closed database, but as a web search engine: A web crawler will actively search for new information published on the Internet. Due to the flexibility of the Semantic Web, newly discovered information from disparate sources can be seamlessly integrated and queried through a single interface. In the long run, the portal should develop into a highly frequented search engine for all fields of systems biology and neuroscience, with applications in basic and preclinical research, as well as in education.

12 Degrees of Certainty in Protein-Protein Interactions mined from text

Olivia Sanchez-Graillet

Institution University of Essex, UK

Email `osanch@essex.ac.uk`

Stage of Studies Second year Phd. Student in a three-year programme

Abstract

Motivation and Objectives Protein interactions are one of the most relevant information that Biologist look for in the literature. Because the amount of literature is available in large sources of information like MEDLINE, tools that help Biologist to discover new or not obvious knowledge are required.

There is already work done on the extraction of protein interactions (Yakushiji et al., 2004). However, in our study we take into account other factors like degree of certainty given by linguistic clues (i.e. modality) as well as the inconsistencies that may be found in the literature. With these considerations we want to give more certainty to the results obtained.

Methods We combine protein information resources with NLP extraction methods in order to extract the protein interactions from text. A first stage for the extraction of the interactions is the identification and annotation of protein names. For this purpose we have developed a system that uses a simple heuristics to find candidate terms and verifies that they belong to the class protein in different sources of protein knowledge (e.g. UniProt, UMLS).

The following stage is the proper extraction of the protein interactions. We have developed a system that performs this task by using a full syntactic parser and pattern matching. We have made a pre-evaluation of this method. We used two different ways of evaluation. The first one is based on the rouge metric. We obtained precision of 0.93 and recall of 83.29. A second evaluation was done by using a scale based on Rinaldis (2004) classification. We obtained recall of 83.35 and precision of 93.03. However, this approach only considers explicit types of interactions.

In the next stage we want to implement a machine learning approach that uses a maximum entropy approach, similar to the one developed by the Singapores group (Xiao et al., 2004) but adding other features that we believe can consider a wider context for the identification of protein interactions.

Remaining Work We are constructing the training data as well as to implementing the machine learning approach. Once we have the protein interactions we will analyse them and consider negation cues (not), coordination, correference with demonstratives (e.g. This protein) and modality.

References

Rinaldi, F., Schneider,G., Kaljurand, K., Dowdall, J., Andronis, Ch., Persidis, A., Konstanti, O. (2004). Mining Relations in the GENIA corpus. In Proceedings of the second workshop on Data Mining and Text Mining for Bioinformatics, Pisa, Italy.

Xiao, J., Su J., Tan G. Z. C., (2004). Protein-Protein Interaction Extraction: A Supervised Learning Approach. In the First International Symposium on Semantic Mining in Biomedicine, Cambridge.

Yakushiji, A.,Miyao, Y., Tateisi, Y., Tsujii, J.,(2004), Biomedical Information Extraction with Predicate-Argument Structure Patterns. In the First International Symposium on Semantic Mining in Biomedicine, Cambridge.

13 Conjunctive Query Answering over Conceptual Schema of Relational Database

Manta Simkus

Institution Free University of Bolzano, Italy

Email mantas.simkus@stud-inf.unibz.it

Stage of Studies First year student European Master in Computational Logic

Abstract The research we are currently carrying out is aimed at the development of a query answering system that enables the users to pose queries over the conceptual schema of a database. Such a system provides added value against conventional DBMSs, where the users are exposed the relation schema only. At the core of our work there is the idea of query answering by rewriting.

In general, query answering by rewriting is divided into two phases. The first one re-expresses a user query posed over the conceptual schema in terms of the relations at the underlying database, and the second evaluates the rewriting over the underlying database (see, e.g., [1]).

Our approach uses a formalism based on Description Logics (DLs) (see [2,3]) to formalize the conceptual schema of the database. Specifically, we have extended the DL DL-Lite (see [5]) with the ability to support n-ary relations, obtaining the DL DLR-Lite. Such a formalism is expressive enough to capture basic Entity-Relationship or UML Class diagrams, while allowing query answering that fully takes into account the constraints in the conceptual schema and is still tractable (i.e., polynomial) in the size of the data (see [4,5]).

We have devised a flexible way of mapping the conceptual level to the underlying relational level, which provides the users an SQL-like query language over the conceptual schema. Queries at the conceptual level are first translated into the relational level queries by taking into account the mapping of entities and relationships to the actual database relations. To provide a complete answer to the query, the system then uses the developed query rewriting technique to take into account the constraints expressed in the conceptual schema. The initial user query is thus translated to a set of SQL queries that are evaluated by the DBMS.

This rewriting technique adds value to conventional query answering techniques. Firstly, the user is allowed to formulate more simple queries using terms

defined in the conceptual schema only, without taking into account some relational database related details (e.g., join attributes). Moreover, the query rewriting technique allows one to infer additional information that was not stated explicitly in the user query but is implied by the constraints at the conceptual level. Last but not least, the formalization of the conceptual schema and the developed reasoning technique allow checking the consistency of the underlying database against the conceptual schema, therefore, the trustiness of the information system is improved.

References

[1] Domenico Lembo, Maurizio Lenzerini, Riccardo Rosati. Methods and techniques for query rewriting. Technical report D5.2, Infomix Consortium, 2004.

[2] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, 2003.

[3] A. Borgida and R. J. Brachman. Conceptual modeling with description logics. In Baarder et al.[2], Chapter 10.

[4] A. Cali, D. Lembo, and R. Rosati. Query rewriting and answering under constraints in data integration systems. In Proc. of the 18th Int. Joint Conf. on Artificial Intelligence (IJCAI 2003), pages 16-21, 2003.

[5] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. DL-Lite: Tractable Description Logics for Ontologies. In Proc. of the 20th Nat. Conf. on Artificial Intelligence (AAAI 2005), 2005.

14 Question Classification in Question Answering Systems

Håkan Sundblad

Institution Linköpings universitet, Sweden

Email `hakjo@ida.liu.se`

Stage of Studies Halfway through PhD

Abstract The ultimate goal of open-domain question answering systems is to provide correct and succinct answers to questions posed in natural language. One of the most crucial aspects for this venture to succeed is to identify the answer type, i.e. the sematic category that the answer should belong to. Moldovan et al. (2002) showed that in cases where "the derivation of the expected answer type [...] fails, the set of candidate answers identified in the retrieved passages is either empty in 28.2% of the cases (when the answer type is unknown) or contains the wrong entities for 8.2% (when the answer type is incorrect)". The problem of correctly assigning a category to a question is referred to as question classification.

The problem of question classification is in many respects similar to text categorization. Therefore, to formalize the problem, we can adopt the definition of text categorization (Sebastiani, 2002). Question classification is the task of assigning a boolean value to each pair $< q_j, c_i > \in Q \times C$, where Q is the domain of questions and $C = c_1, c_2, ..., c_{|c|}$ is a set of predefined categories. Assigning $< q_j, c_i >$ to the value T indicates that q_j is judged to belong to the category c_i, while an assigment to the value F indicates that q_j is not judged as belonging to the category c_i.

In a series of experiments I have examined how various machine learning approaches perform (k Nearest Neighbors, Nave Bayes, Decision Trees, Sparse Network of Winnows, Support Vector Machines) for question classification. Two different coprora have been used for the task: one prepared by the Cognitive Computation Group at the University of Illinois, and one based on questions posed to the AnswerBus system. A taxonomy of 50 distinct answer types has also been used. Results from the experiments are in some respects similar to results from previous research (Li and Roth, 2002; Zhang and Lee, 2003; Suzuki, Taira, Sasaki and Maeda, 2003; Hacioglu and Ward, 2003), but also modifiy the picture to some extent.

Future work will include looking at question classification for restricted-domain question answering systems (taxonomies and classification), and also at question processing at large in question answering systems.

15 Patient Record Overviews and Navigation

Erik Sundvall

Institution Linköpings universitet, Sweden

Email erisu@imt.liu.se

Stage of Studies first year PhD student in a five-year programme

Abstract How can patient information in an electronic health record (EHR) be displayed best in order to gain a quick overview of a patients history and

other important facts? The thesis will explore different facets of this question. Requirements vary depending on the users role and current task, the overview should adjust accordingly.

The need for more structured, computer interpretable and semantically well defined EHR content has this far to a great deal been driven by the desire to use decision support systems. The thesis will also use this improved structure to enhance overview, presentation and navigation of the EHR using various visualization techniques. A hypothesis is that automated reasoning may also be used to better filter/summarize available information before presentation.

The structure of the information in an EHR, and the possibilities to aggregate data for statistical purposes are dependent on data models, terminologies etc. used. Good overview and navigation of those is a related important task that is also explored in the thesis. The first paper (submitted to AMIA2005) thus presents a prototype applying well known methods like focus+context and self-organizing layouts from the fields of Information Visualization and Graph Drawing to terminologies like SNOMED CT and ICD-10. The aim was to simultaneously focus on several nodes in the terminologies and then use interactive animated graph navigation and semantic zooming to further explore the terminology systems without loosing context. The prototype, based on Open Source Java components, demonstrated how a number of information visualization methods can aid the exploration of medical terminologies with millions of elements and can serve as a base for further development. Evaluation of the solution is expected to start after the summer.

Display of interactive graphical timelines showing patient history is an obvious and important part of a patient overview. Work along these lines has already been conducted by others (e.g. the LifeLines project) and will be explored and expanded in papers and software planned for submission and release in the beginning of 2006 (including a brief evaluation study).

The continued exploration and development of patient overviews depends on what can be concluded by the initial findings. One thing that stands out as important is the need to work closely with currently available and possible future EHR platforms. Involvement in the openEHR project and archetype based systems has begun as a part of this. The use of graphic languages (pictograms etc.) for quick information overview is another thing expected to be further explored in the thesis.

The main method used in the thesis is work pretty much just a common software development cycle starting by identifying problems/tasks followed by literature studies and exploration of currently available solutions and technologies. Then design and prototyping (currently java-implementations) follow, including a lot of redesign and refactoring. At the end of the cycle the prototype should be evaluated using user studies and other methods.

16 The Spock System: Developing a Runtime Application Engine for Hybrid-Asbru Guidelines

Ohad Young and Yuval Shahar

Institution Ben Gurion University, Israel

Email {ohadyn , yshahar}@bgumail.bgu.ac.il

Stage of Studies -

Abstract Clinical Guidelines are a major tool for improving the quality of medical care. However, most guidelines are available only in free text. A major current research direction is automating the application of guidelines at the point of care. To support that automation, several requirements must be fulfilled, such as guidelines specification in a machine-interpretable format, and provision of a connection to an electronic patent record. We propose an innovative, comprehensive approach for guideline application, which capitalizes on our extensive work on development of the Digital electronic Guidelines Library (DeGeL). The DeGeL framework includes a new hybrid model for incremental conversion of free-text guidelines through several intermediate representations into a machine-interpretable format. The new approach was implemented, in the case of the Asbru guideline ontology, as the Spock system. Spock's hybrid execution engine supports to some extent the automation of guideline application represented at an intermediate format such as semi-structured text. Spock uses the IDAN mediator for answering complex queries referred to heterogeneous clinical data repositories and the KAVE-II system for intelligent and interactive exploration and visualization of patient data. Spock was evaluated in a preliminary fashion by applying several guidelines such as COPD, PID and Hypothyroidism to sample simulated patient data.

Keywords Guideline-Based care; Protocols and guidelines; Knowledge representation; Clinical decision support systems; Guideline execution engines, Therapy planning; Temporal mediation; Medical informatics