Citation: Lee, C.H., Khoo, C., & Na, J.C. (2004). Automatic identification of treatment relations for medical ontology learning: An exploratory study. In I.C. McIlwaine (Ed.), *Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference* (pp. 245-250). Wurzburg, Germany: Ergon Verlag.

Manuscript submitted to ISKO 2004

Automatic Identification of Treatment Relations For Medical Ontology Learning: An Exploratory Study

Authors:

Chew-Hung Lee (lchewhun@dso.org.sg)
Christopher Khoo (assgkhoo@ntu.edu.sg)
Jin-Cheon Na (tjcna@ntu.edu.sg)

Authors' address:

Division of Information Studies
School of Communication & Information
Nanyang Technological University
31 Nanyang Link, Singapore 637718
Tel: (65) 6790-5011   Fax: (65) 6792-7526

Chew-Hung Lee, Christopher Khoo, Jin-Cheon Na
Division of Information Studies
School of Communication and Information
Nanyang Technological University, Singapore

Automatic Identification of Treatment Relations For Medical Ontology Learning: An Exploratory Study

Abstract: This study is part of a project to develop an automatic method to build ontologies, especially in a medical domain, from a document collection. An earlier study had investigated an approach to inferring semantic relations between medical concepts using the UMLS (Unified Medical Language System) semantic net. The study found that semantic relations between concepts could be inferred 68% of the time, although the method often could not distinguish between a few possible relation types. Our current research focuses on the use of natural language processing techniques to improve the identification of semantic relations. In particular, we explore both a semi-automatic and manual construction of linguistic patterns for identifying treatment relations in medical abstracts in the domain of colon cancer treatment. Association rule mining was applied to sample sentences containing both a disease concept and a reference to drug, to identify frequently occurring word patterns to see if these patterns could be used to identify treatment relations in sentences. This did not yield many useful patterns, suggesting that statistical association measures have to be complemented with syntactic and semantic constraints to identify useful patterns. In the second part of the study, linguistic patterns were manually constructed based on the same sentences. This yielded promising results. Work is ongoing to improve the manually constructed patterns as well as to identify the syntactic and semantic constraints that can be used to improve the automatic construction of linguistic patterns.

1. INTRODUCTION

This study is part of a project to develop an automatic method to build ontologies, especially in a medical domain, from a document collection. Ontologies are knowledge resources containing the concepts, relations and axioms found in a domain. Ontologies play an important role in the Semantic Web as well as in knowledge managamenet. The creation of ontologies is non-trivial, requiring analysis of domain sources, background knowledge, and obtaining consensus among the users of the ontologies. The conventional approach for constructing an ontology is to manually enumerate the concepts and relations found in a domain from domain sources (Gómez-Pérez, Fernández-López, & Corcho, 2003). This approach is not suitable for developing a large ontology as it is both labour intensive and likely to give rise to inconsistencies. An alternative approach is to use automatic or semi-automatic methods to extract the concepts and relations (Maedche, 2002; Navigli, Velardi, & Gangemi, 2003) through clustering or using frequency-based measures like tf*idf. We have embarked on a project to develop an automatic ontology learning method by enriching existing ontologies, identifying the semantic relations between concepts in the ontology after analyzing domain texts.

In an earlier study (Lee, Na & Khoo, 2003), we carried out a small experiment using a sample of medical article abstracts in the area of colon cancer treatment to identify pairs of related concepts and inferred the semantic relations between the terms in each pair using the

UMLS (Unified Medical Language System) semantic network, an existing medical knowledge base maintained by the National Library of Medicine (URL http:www.nlm.nih.gov/research/umls). Important terms were extracted from the sample medical abstracts and mapped to a medical concept in the UMLS Metathesaurus. The UMLS Metathesaurus contains biomedical concepts and terms from many controlled vocabularies and classification systems used in medical information systems. An association rule mining tool (Borgelt, 2003) was applied to the concepts to find associated concept pairs.

After finding associated concept pairs, we proceeded to identify specific relations between the concepts using the UMLS semantic network. The UMLS semantic network specifies the set of basic semantic types (the nodes in the network) that may be assigned to concepts in the UMLS Metathesaurus, and also defines the set of relationships (the links in the network) that may hold between the semantic types. We used this information to infer the probable semantic relations between the extracted concepts.

We were able to infer semantic relations between concepts automatically from the UMLS semantic network 68% of the time, although the method could not distinguish between a few possible relation types (e.g., treat, cause, etc.). This suggests that it is worthwhile to investigate the use of natural language processing (NLP) techniques to improve accuracy in identifying relations between concepts in the medical ontology.

Our current study seeks to develop automatic methods for improving the identification of semantic relations between ontology concepts using natural language processing—in particular by applying information extraction techniques on a document collection. Information Extraction (IE) refers to the automatic process of extracting, from natural language text, pieces of facts (represented by a word or phrase) relevant to an event or topic to fill slots in a predefined template or fields in a database record. IE typically uses shallow NLP processing to identify the words to extract from the text using a set of extraction patterns. An extraction pattern is typically a sequence of words with slots (or blanks) to indicate the position of the words to be extracted. Our plan for using the IE pattern matching is not to extract concepts but to identify the type of semantic relation between pairs of concepts that have previously been identified.

The strength of the IE engine lies in the set of extraction patterns. The richer the patterns, the better the IE engine can perform. Our current research focuses on the semi-automatic construction of linguistic patterns for identifying semantic relations between a disease and a treatment (e.g. drug). We focus first on relations expressed within sentences. The main objectives of the study are:

- To develop a semi-automatic method of constructing patterns for identifying treatment relations expressed in text, and
- To construct a set of extraction patterns to identify treatment relations.

The approach taken in this study is to first identify sentences containing a reference to a drug as well as to a disease. Most such sentences express a treatment relation between the drug and the disease. Based on these sentences, we explored two methods of constructing the extraction patterns:

- using association rule mining to identify frequently occurring word patterns in the sentences—to see if these frequent patterns can be used as to identify treatment relations in sentences. As reported later in this paper, this was not found to yield good results.
- manual construction of extraction patterns.

In this study, we continue to use medical abstracts in the area of colon cancer treatment as our dataset.

2. RELATED WORKS

Maedche (2002) and Navligli, Velardi & Gangemi (2003) worked on semi-automatic methods to extract the concepts and relations. They investigated building ontologies from broad domain documents, such as travel related documents. Since the target domains were very broad, the generated ontologies did not have deep hierarchies compared to the ones manually generated by domain experts. Some projects use Word Net (Fellbaum, 1998) as an existing domain knowledge base to overcome the problem. However, it could be too general for specific domain documents, such as medical documents. It appears that using existing domain knowledge bases is necessary when building domain specific ontologies.

Blake and Pratt (2001) worked on mining semantic relationships among medical concepts (or terms) from medical texts. They focused on "Breast Cancer Treatment" using association rule mining to find associated concept pairs like magnesium-migraines. They were mainly interested in mining the *existence* of a relationship between medical concepts and not in identifying the specific type of semantic relation for the associated concept pairs. Because identifying specific semantic relations is very important for ontology learning, our work focuses more on finding specific semantic relations.

Khoo, Chan, & Niu (2000) looked for causal relations expressed in medical abstracts by matching graphical patterns in syntactic parse trees They used the FDG parser from Connexor to process medical abstracts from four different disease domains (schizophrenia, depression, AIDS and heart disease) into syntactic parse trees. The parse trees were converted into conceptual graphs, and a graph-matching algorithm was applied to detect causal patterns.

3. DATA PREPARATION

500 records in the area of "colonic neoplasms/drug therapy" were downloaded from the MedLine database through the PubMed interface (National Library of Medicine, 2003). 408 of the records contained abstracts. Each abstract was reduced to a list of sentences, and each sentence passed into the MMTx program that uses the MetaMap algorithm (Aronson, 2001) to extract UMLS concepts. Sentences containing a concept relating to a disease and/or a concept relating to a pharmacologic substance (i.e. drug) were identified.

Of the 408 abstract downloaded, 108 abstracts contained at least one sentence with both a disease concept and a reference to a drug. We shall refer to these abstracts informally as the *good* abstracts. The remainder 300 abstracts are referred to as *bad* abstracts—they do not contain any sentence with both a drug and a disease. The 211 drug+disease sentences in the "good" abstracts constitute our training sentences for constructing extraction patterns. Most of these sentences contain a treatment relation between the drug and the disease. Two example sentences with both a drug and a disease are shown below:

- "We report a case of irinotecan-resistant colon cancer *responding to* chronotherapy with oxaliplatin (L-OHP), 5-FU, l-LV (l-Leucovorin)."
- "These results indicate that chronomodulated 5-FU and LV with L-OHP therapy *could be an effective regimen for* cases of irinotecan-resistant colon cancer."

It can be seen that researchers do not always use the word "treat" or "treatment" to indicate a treatment relation. They may use other terms such as "respond to" or "effective regimen."

The sentences were then divided into individual word tokens, and punctuation marks, prepositions, determiners, conjunctions, disjunctions, pronoun and numbers were removed.

4. EXPERIMENT WITH ASSOCIATION RULE MINING

The Apriori algorithm (Borgelt, 2003) was applied to the dataset using the Clementine data mining software. Using the settings of minimum support 2% and minimum confidence 80%, a total of 281 rules were generated. A variety of statistical measures were used to rank the rules, such as Rule Confidence, Normalized Chi Square, Confidence Difference and Confidence Ratio. Table 1 illustrates the top 10 rules obtained using the Normalized Chi Square measure. The results were not convincing as the rules contained few terms that signified a treatment relation.

Clearly, word patterns and statistical association measures alone cannot be used to construct extraction patterns for the treatment relation. We are exploring how statistical association measures can be combined with syntactic and semantic constraints to construct extraction patterns semi-automatically. To obtain some insights into what kind of syntactic and semantic constraints might be helpful, we manually constructed extraction patterns based on the 211 drug+disease sentences.

Support	Confidence	Consequent	Antecedent 1	Antecedent 2
6.2	92.3	cell	lines	
2.4	92.3	synthase	thymidylate	
2.4	100	thymidylate	synthase	
2.4	100	surgery	alone	compared
2.4	100	compared	surgery	
2.4	100	compared	alone	surgery
4.7	90	leucovorin	5-FU	Dukes
4.3	88.9	cell	colon cancer	lines
12.8	81.5	patients	colon cancer	stage
2.4	100	lines	cell	studied

Table 1. Top 10 rules extracted by the Apriori algorithm using Normalized Chi-Square

5. EXPERIMENT WITH MANUAL CONSTRUCTION OF PATTERNS

A total of 224 extraction patterns were manually constructed by one of the authors. The patterns ranged from single words to phrases with embedded wildcard tokens. The patterns can be grouped into the following semantic categories:

- Administration of treatment, e.g. *exposure to, use of, using, clinically used, administered, receiving treatment with*
- Treatment dosage, e.g. *low-dose, dose of, dosage schedule*
- Mortality and survival, e.g. *mortality, death rate, survival benefit, extends the survival*
- Therapy, e.g. *chemotherapy, treatment, regimen, adjuvant, drug, pro-drug*
- Clinical trial, e.g. *tested on, feasibility trial, clinical trial*
- Effect, e.g. *outcome, responsive, influence, results, sensitivity, effective*. Words referring to an effect can be subdivided into
    - Agent of effect, e.g. *agents, anti-cancer agent*
    - Target of effect, e.g. *targeting, targeted*
    - Effect action, e.g. *active in, anti-tumor activity*
    - Effect against something, e.g. *anti-cancer, anti-tumor, antagonist*
    - Effect in controlling or inhibiting something, e.g. *controlling, inhibition, inhibitor, cytostatic*
    - Effect in decreasing or increasing something, e.g. *impaired, decrease, reduce,*

regression, remission, increase, elevated

- Effect in killing something, e.g. *kill, apoptosis inducing, cytotoxic*
- Good effect, e.g. *beneficial, useful, benefit, improve, promising*
- Therapeutic effect, e.g. *treat, curatively, clinical, clinically*
- Free of disease, e.g. *disease-free, recurrence-free*
- Interaction effect, e.g. *synergistic, modulation*

We carried out a preliminary evaluation of the effectiveness of the patterns for identifying treatment relations, using a convenience sample of 30 "good" abstracts from our dataset. Since the patterns were constructed based on the drug+disease sentences in the abstracts (i.e. sentences containing both a drug and a disease concept), the evaluation was based on the following sets of sentences: drug-only sentences, disease-only sentences and sentences with neither a drug nor a disease concept.

Table 2 summarizes the results. The *recall* evaluation measure was at least 60% for all categories of sentences. *Precision* was low, especially for sentences that do not contain a drug concept.

Sentence Categories	**(1)** **Total # sentences**	**(2)** **# sentences with treatment relation**	**(3)** **# sentence identified by the patterns**	**(4)** **# identified sentences containing a treatment relation**	**Precision (4)/(3)**	**Recall (4)/(2)**
Drug Only	58	21	28	14	50.0%	66.7%
Disease Only	50	11	25	8	32.0%	72.7%
Both	90	55	71	53	74.6%	96.4%
Neither	91	5	38	3	7.9%	60.0%

Table 2. Precision and recall of treatment relations using manually constructed patterns

For comparison, the patterns were also applied to sentences from 30 "bad" abstracts (that do not contain any sentence with both a drug and a disease concept). A precision of 37.3% and a recall of 69.1% were obtained. The low precision suggests that the "bad" abstracts may be quite different in nature from the "good" abstracts. Our impression from a quick scan is that the bad abstracts tend to report more theoretical studies which are not directly focused on treatments.

A preliminary error analysis was carried out with 20 false negatives—sentences containing a treatment relation that were not identified by the patterns—and 20 false positives—sentences identified by the patterns but do not contain a treatment relation.

For the 20 false negative sentences, 16 new patterns were identified, six of which were new patterns not found in the list of original patterns, four were spelling variations of existing patterns (e.g. anti-tumor vs. anti-tumour), four were parts of existing patterns, and the last two were similar to an existing pattern either through rearrangement of word order or having an important word in common.

For the false positives, the majority of the sentences (15 sentences) contained no specific reference to a drug or a treatment. Four of the sentences described the results of a theoretical study. Other sentences described a diagnosis, a target enzyme, a cytotoxic effect and a schedule of chemotherapy. Two of the sentences referred to a treatment but no useful information could be extracted from the sentences:

- "The treatment was *effective* and the lesion disappeared."
- "There was a massive *therapeutic* effect without side effects."

The two example sentences above indicate that in the evaluation, we accepted a sentence as containing a treatment relation only when there was a specific reference to a treatment and a disease. Several false positive sentences contained a reference to a treatment described in other sentences in the abstract. Thus, a less strict definition of a treatment relation would yield a higher precision rate for the constructed patterns.

6. CONCLUSION

We have explored a semi-automatic and a manual approach to constructing linguistic patterns for identifying treatment relations in sentences of medical abstracts. The study was based on a sample of abstracts in the domain of colon cancer treatment. The approach taken is to first identify sentences containing a reference to both a drug and a disease, assuming that such sentences tend to express a treatment relation between the drug and the disease. We explored the use of association rule mining to identify frequently occurring word patterns to see if these patterns could be used to identify treatment relations in sentences. This did not yield useful patterns, suggesting that statistical association measures have to be used in combination with syntactic and semantic constraints.

In the second part of the study, we constructed the extraction patterns manually. This yielded promising results. Work is ongoing to improve the manually constructed patterns based on an error analysis of false positive and false negative sentences. We also intend to develop patterns for identifying treatment relations across two sentences. The extraction patterns developed will also be applied to other medical domains, e.g. breast cancer, heart disease and AIDS, in order to develop a more general set of extraction patterns for the treatment relation.

REFERENCES

Aronson, A.R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap Program. In *Proceedings of the 2001 Symposium of the American Medical Informatics Association* (pp. 17-21).

Blake, C., & Pratt, W. (2001). Better rules, fewer features: A semantic approach to selecting features from text. In *Proceedings of the IEEE Data Mining Conference, held in San Jose, California, 2001* (pp. 59-66).

Borgelt, C. (n.d.). Apriori implementation. Retrieved September 2003 from http://fuzzy.cs.uni-magdeburg.de/~borgelt/doc/apriori/

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2003). *Ontological Engineering*. London: Springer Verlag.

Khoo, C.S.G., Chan, S., & Niu, Y. (2000). Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. In *ACL-2000: 38th Annual Meeting of the Association for Computational Linguistics* (pp. 336-343).

Lee, C.H., Na, J.C., & Khoo, C.S.G. (2003). Ontology learning for medical digital libraries. In *Proceedings of the 6th International Conference of Asian Digital Library* (pp. 302-305).

Maedche, A. (2002). *Ontology learning for the Semantic Web*. Boston: Kluwer Academic Publishers.

Navigli, R., Velardi, P., & Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems,* 18(1), 22-31.