Pattern Mining for Information Extraction Using Lexical, Syntactic and Semantic Information: Preliminary Results

Christopher S.G. Khoo, Jin-Cheon Na, and Wei Wang

Division of Information Studies, Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore 637718
{assgkhoo,tjcna,w060001}@ntu.edu.sg

Abstract. A method is being developed to mine a text corpus for candidate linguistic patterns for information extraction. The candidate patterns can be used to improve the quality of extraction patterns constructed by a pseudo-supervised learning method—an automated method in which the system is provided with a high quality seed pattern or clue, which is used to generate a training set automatically. The study is carried out in the context of developing a system to extract disease-treatment information from medical abstracts retrieved from the Medline database. In an earlier study, the Apriori algorithm had been used to mine a sample of sentences containing a disease concept and a drug concept, to identify frequently occurring word patterns to see if these patterns could be used to identify treatment relations in text. Word patterns and statistical association measures alone were found to be insufficient for generating good extraction patterns, and need to be combined with syntactic and semantic constraints. In this study, we explore the use of syntactic, semantic and lexical constraints to improve the quality of extraction patterns.

Keywords: Information Extraction, Pattern Mining, Apriori Algorithm.

1 Introduction

Information extraction systems use automated methods to extract from natural-language text facts or pieces of information related to a particular topic or event. The facts are used to fill pre-defined templates or to populate a database for various purposes. Information extraction is usually performed using pattern matching—searching for certain linguistic patterns in the text that indicate the presence of the desired information. These extraction patterns can be constructed automatically or semi-automatically by the system by analyzing sample relevant text and the associated answer key (the training corpus) that is usually constructed by human analysts. An information extraction system requires an extensive training corpus or review of the extraction patterns by a human expert to achieve good accuracy.

The challenge is to develop user-friendly personalizable information extraction systems that can be trained by end-users to give reasonable accuracy with a small training set. Since the training set is small, the system needs to use other sources of information to supplement the small amount of information provided by the user

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 676–681, 2008.

in the construction of extraction patterns. The text corpus itself represents the most conveniently available source of supplementary information to exploit to improve the extraction patterns.

In this study, we attempt to develop a method to mine candidate linguistic patterns from the text corpus for information extraction. The candidate patterns mined from the corpus can be used in two ways in the development of extraction patterns: (Not clear about the following two approaches)

1. Machine-assisted pattern construction: given a sentence containing a target piece of information to extract, the system can present the most promising candidate patterns (this term is not clear) for the user to select and customize to form an extraction pattern.
2. Pseudo-supervised learning method: an automated method in which the system is provided with a high quality seed pattern or clue, which is used to automatically generate a training set. Since the training set generated by the seed pattern is not as good as a manually constructed training set, the patterns learnt will include a higher proportion of erroneous patterns. Limiting the patterns learnt to those in the set of candidate patterns will serve to filter out the more promising patterns.

Though we are interested in both these uses of pattern mining, this report focuses on the second application of mining patterns that can be used in pseudo-supervised learning. We retrieved a sample of medical abstracts from the Medline database on the topic of colon cancer therapy and attempted to develop patterns for extracting treatment relations from the abstracts. Instead of manually constructing a training set, we assumed that any sentence that contains a treatment concept (e.g. drug) and a disease concept expresses a treatment relation between the treatment and disease. In other words, we used a semantic pattern to retrieve all sentences containing a treatment and a disease, and assumed that the treatment and disease are to be extracted. These sentences then represent the training set, and extraction patterns are constructed to represent the linguistic context of *treatment* and *disease*. The extraction patterns can later be applied to other sentences to extract new treatments, new diseases and new relations. Figure 1 shows the overall process for mining information extraction patterns; detailed explanations follow in later sections.

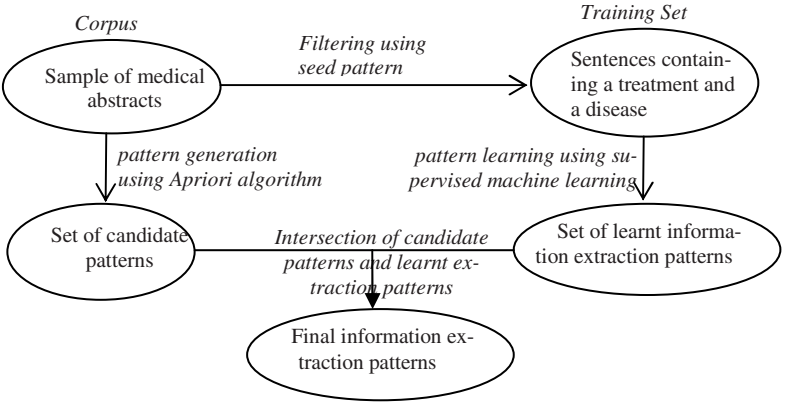

Fig. 1. The overall process of generating information extraction patterns

2 Previous Studies

In an earlier study [1], we had found that mining patterns to extract drug and disease in a sentence is a non-trivial task. We used the Apriori algorithm [2] to mine sample sentences containing a disease concept and a drug concept, to identify frequently occurring word patterns to see if these patterns could be used to identify treatment relations in sentences. Various measures were used to rank the rules, such as Rule Confidence, Normalized Chi Square, Confidence Difference and Confidence Ratio. The results were not convincing as the rules contained few terms that signified a treatment relation.

Word patterns and statistical association measures alone were not good enough to construct extraction patterns. Statistical association measures need to be combined with syntactic and semantic constraints. To obtain some insights into what kind of syntactic and semantic constraints might be helpful, we manually constructed extraction patterns for identifying sentences containing drug-disease relations based on 100 abstracts. We found that the patterns could be grouped into the following domain-specific semantic categories:

- Administration of treatment, e.g. *exposure to, use of, using, clinically used, administered,* and *receiving treatment with.*
- Treatment dosage, e.g. *low-dose, dose of,* and *dosage schedule.*
- Mortality and survival, e.g. *mortality, death rate, survival benefit,* and *extends the survival.*
- Therapy, e.g. *chemotherapy, treatment, regimen, adjuvant, drug,* and *prodrug.*
- Clinical trial, e.g. *tested on, feasibility trial,* and *clinical trial.*
- Effect, e.g. *outcome, responsive, influence, results, sensitivity,* and *effective*. Words referring to an effect can be subdivided into 11 subtypes, including agent of effect (e.g. *anti-cancer agent*), target of effect (e.g. *targeting*), effect action (e.g. *anti-tumor activity*), effect against something (e.g. *anti-cancer, anti-tumor,* and *antagonist*), etc.

From this, we compiled a dictionary of words belonging to these semantic categories.

We continue to investigate what kind of syntactic and semantic constraints can be imposed on linguistic patterns mined from a text corpus to generate good quality extraction patterns. In particular, we wanted to find out to what extent adding the constraints from the domain-specific semantic categories would improve the extraction patterns.

3 Method for Generating the Extraction Patterns

1570 abstracts were downloaded from the MEDLINE database [3] via the PubMed interface using "colon cancer/therapy" as query. These articles were then parsed using the MMTx (MetaMap Transfer) program developed by the National Library of Medicine [4] to produce an output text file. MMTx is a part-of-speech and semantic tagger which takes biomedical text as input and identifies Unified Medical Language System (UMLS) concepts in the text by mapping relevant phrases to the UMLS Metathesaurus

[5]. The output was further processed to tag the tokens (either words or phrase chunks) in the sentence with part-of-speech and semantic tags (if available) and stored the information in a relational database.

The Apriori Algorithm was then used to generate all possible 2, 3, 4 and 5-token patterns that occur at least 5 times in the corpus. A token can be represented by either of the following attributes:

- Lexical token (L), i.e. word or phrase chunk,
- Part-of-speech (P), or
- Semantic concept (C).

Example candidate patterns generated are shown in Table 1. Note that there is an implied wildcard (representing 0 to 3 tokens) between the tokens in each pattern—i.e. the patterns are sequential patterns but not adjacent patterns.

Table 1. Example candidate patterns

Pattern Type	Example candidate patterns
CLC	[Neoplastic Process] treats [Therapeutic or Preventive Procedure]
CPLC	[Neoplastic Process] *noun* underwent [Therapeutic or Preventive Procedure]
CLCPC	[Patient or Disabled Group] underwent [Therapeutic or Preventive Procedure] *prep* [Neoplastic Process]

Note: Terms in square brackets represent UMLS Metathesaurus concept.
Terms in italics represent part-of-speech tag.

Without any lexical, syntactic or semantic constraints, a large number of candidate patterns were generated from the 1570 abstracts. For any particular sentence in the training set, there were on average more than 1000 candidate patterns that match parts of the sentence and have to reviewed by a human analyst to select an extraction pattern. Furthermore, the very few useful patterns were often buried deep in the set of candidate patterns.

Based on an informal error analysis, we introduced the following constraints to improve the quality of the candidate patterns:

- There must be at least 1 lexical item that is not a stopword in the candidate pattern
- The pattern must not contain a preposition as the first or last token

We then filtered out the patterns containing a treatment concept and a disease concept. These are candidate patterns for identifying disease-treatment relations in sentences. The treatment and disease concepts were identified using the UMLS semantic types annotated by the MMTx program.

On examining the candidate patterns, we found some useless patterns that contain the treatment concept and disease concept in adjacent positions (with no tokens between them). These were eliminated since the tokens between the treatment and disease concepts seem to be the most useful for identifying disease-treatment relations.

Next, we separated the candidate patterns into 2 subsets:

- Subset 1 contains a word token that matches an entry in the domain-specific semantic dictionary described in the last section. These are words associated with the semantic categories of *treatment administration*, *dosage*, *mortality and survival*, *therapy*, *clinical trial* and *effect*.
- Subset 2 does not contain a word in the dictionary.

Subset 1 contains 62 candidate patterns, some of which are listed in Table 2.

Table 2. Some candidate patterns

Token1	Token2	Token3	Token4	Token5	Type
[Therapeutic or Preventive Procedure]	Patients	[Neoplastic Process]			CLC
[Neoplastic Process]	noun	underwent	[Therapeutic or Preventive Procedure]		CPLC
[Therapeutic or Preventive Procedure]	in	[Neoplastic Process]	cells		CLCL
[Neoplastic Process]	prep	[Therapeutic or Preventive Procedure]	prep	Apoptosis	CPCPL

The candidate patterns were converted to final extraction patterns. Converting a candidate pattern to an extraction pattern involves indicating where the extraction slots are in the pattern—the placeholders for the information of interest to extract. To extract a disease-treatment relation, two slots—a disease slot and a treatment slot need to be created, and this can easily be accomplished by converting the disease concept token and treatment concept token to slot tokens. Three types of extraction patterns can be constructed:

- Type 1: patterns with a disease slot only
- Type 2: patterns with a treatment slot only
- Type 3: patterns with a disease and a treatment slot.

We have investigated only the last two types of patterns.

4 Extraction Results

114 extraction patterns (Type 2 and Type 3) were derived from the 61 candidate patterns. The patterns were applied back to the corpus, i.e. the 1570 medical abstracts, to extract *disease* and *treatment* from each matched sentence through pattern matching. We computed the estimated precision measure for each pattern based on the first 20 extractions by the pattern.

The average precision for the two-slot (*disease+treatment*) patterns (Type 3) were:

- 47% for sentences containing both *treatment* and *disease* concepts
- 47% for sentences containing only *treatment* concepts
- 54% for sentences containing only *disease* concepts.

The average precision for single-slot (*treatment* only) patterns (Type 2) were:

- 58% for sentences containing both *treatment* and *disease* concepts

Wrong extractions were mainly due to six causes:

1. Erroneous Parsing. The MMTx parser and an auxiliary annotation module failed to tag noun phrases with correct part of speech class and hence incurred missing hits.
2. Inadequate chunking. The current preprocessing is unable to recognize major phrasal units such as noun phrases and hence causes a lot of inaccurate extractions.
3. Complex entity names. Names of most drugs or therapies, especially their abbreviations and acronyms, often could not be tagged as noun or adjective and hence ended up with "unknown" as their part-of-speech tag.
4. Complex syntactic structures. Coordination, relative clauses, prepositional phrases, etc. reduce the extraction accuracy of a pattern.
5. Coreference problem. Pronouns and definite references are common in medical articles probably because of the complex names of many medical terms. Another type of reference is an is-a reference, e.g. "TS-1 is expected to be an effective agent for the treatment of colon cancer with peritoneal dissemination." The extracted treatment was "agent" instead of "TS-1".
6. Semantic uncertainty. This refers to sentences expressing relations like "A has/causes C which cures B". "C" can be extracted instead of "A". However, this is not necessarily true when "C" is just a chemical or biological function or reaction, but not a therapy or drug. Similarly, relations like "A inhibits/causes C of B" would not always be extracted correctly.

Solutions for these six issues such as adding a second layer parsing and a phrase identifier, designing domain-specific patterns to identify entities and so on will be investigated and implemented in the future work. We are currently analyzing each extraction pattern and the text extracted by the pattern to see what further constraints can be added to improve their accuracy. We have also noticed that out of the 176 entries in the domain-specific dictionary, only 22 matched with patterns mined from the corpus. We are investigating the usefulness of the other 154 entries to see in what context they appear in the text.

References

1. Lee, C.H., Khoo, C., Na, J.-C.: Automatic Identification of Treatment Relations for Medical Ontology Learning: an Exploratory Study. In: McIlwaine, I.C. (ed.) Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference, pp. 245–250. Ergon Verlag, Wurzburg, Germany (2004)
2. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216. ACM Press, New York (1993)
3. National Library of Medicine. MEDLINE Fact Sheet (Retrieved November 14 2007), http://www.nlm.nih.gov/pubs/factsheets/medline.html
4. National Library of Medicine. MetaMap Transfer (MMTx): Documentation (Retrieved November 14 2007), http://mmtx.nlm.nih.gov/docs.shtml
5. National Library of Medicine. Unified Medical Language System Fact Sheet (Retrieved November 14 2007), http://www.nlm.nih.gov/pubs/factsheets/umls.html