# Automatic Detection of Treatment Relationships for Patent Retrieval

Aaron Chu
UCLA Computer Science
Department
Los Angeles, California 90024
aaronc@ucla.edu

Shigeyuki Sakurai[1]
4th Patent Examination Department
Japan Patent Office
Tokyo, Japan
sakurai-shigeyuki@jpo.go.jp

Alfonso F. Cardenas
UCLA Computer Science
Department
Los Angeles, California 90024
cardenas@cs.ucla.edu

## ABSTRACT

We devise a method for automatically detecting treatment relationships using lexico-syntactic patterns and its application to medical-oriented patent retrieval. This process for detecting treatment relationships involves finding lexico-syntactic patterns that are highly indicative of treatment relationships and also producing classification rules for those patterns.

This treatment relationship detection process is then used in a system to find treatment relationships based on a user query in a medical patent source. The query will consist of terms that the user wants to find in the subject or object of a treatment relationship. This is of great interest to both patent examiners and patent applicants as they search for prior art. Through the use of classification rules, this system was able to achieve a precision of 85.81% on a set of 20 test queries.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Linguistic processing.*

## General Terms

Experimentation, Languages

## 1. INTRODUCTION

Treatment relationships refer to any case where A (the subject of the relationship) can be used in the treatment of B (the object of the relationship) to lessen the adverse effects of B. In most of the cases, B will be some sort of negative disease or condition state such as depression, arthritis, or fever. On the other hand, A may be a drug such as Tylenol, an activity such as surgery, or something else that can be used to treat B.

The invention of drugs and other methodologies to treat various diseases and conditions is an ever expanding patent classification that has the interest of patent examiners and applicants. Our system can be used as an efficient way to search these classifications of patents for relationships instead of using only traditional keyword searches.

In section 2, we review previous work done in the area of patent and relationship search. Section 3 discusses the process of automatically detecting treatment relationships using lexico-syntactic patterns. The results of this process are presented in section 4. Lastly, we present our conclusions in section 5.

## 2. BACKGROUND

The number of patent applications is steadily increasing all over the world. The USPTO (United States Patent and Trademark Office) received approximately 445,000 and 467,000 patent applications in 2006 and 2007 respectively as noted in [5]. Thus, the demand for a powerful patent search system and an effective information retrieval method for patent documents is growing.

[2] and [3] presented practical and basic patent search systems which retrieve patent documents according to keywords provided as a user query. [2] established an actual patent search and classification system using tf-idf scoring to retrieve and rank patent documents. [3] presented collections of patent documents organized by topic and patent retrieval with collection selection.

There has also been research in trying to retrieve semantic information from patents. [7] showed how to generate a regular expression for patent documents, which matches and identifies the problems solved in the patents. In an integrated and modern manner, [6] is creating PATExpert, an integrated environment for storing, viewing, and searching patents. PATExpert is based on recent ontology technology. It can store concepts and relations between the concepts in a patent document.

Content-oriented semantic processing of patent documents is more powerful than keyword search or statistical keyword search, especially for patent researchers to find out prior-art patents and for patent applicants to find out contending technologies of other companies. Semantic search allows users to filter out irrelevant patents that would be returned through keyword searches by narrowing down the concepts of interest. Nevertheless, in order to store semantic information in a patent document, we have to prepare it manually or extract it from the existing text of patents.

In the area of relationship retrieval, Girju proposes in [1] a method of automatically detecting causal relationships using lexico-syntactic patterns of the form $NP_1$ VP $NP_2$ (where NP means noun phrase and VP means verb phrase) for question answering purposes. In this paper, we adapt Girju's approach to automatically detect treatment relationships for the purpose of a semantic patent search engine.

*PaIR'08*, October 30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-60558- 256-6 /08/10...$5.00.

[1] Work done on this project was during a one year leave at the UCLA Computer Science Department.

# 3. OUR APPROACH

We use lexico-syntactic patterns of the form $NP_1$ VP $NP_2$ to identify treatment relationships. The VP would contain the pattern and the two NP would contain the subject and object. This structure is a very common relationship structure, known to be able to capture a multitude of relationships. These patterns are learned from our patent corpus as described in section 3.1.

The data source we are using to extract these relationships is the U.S. Patents classification "Drug, bio-affecting and body treating compositions." We have collected 50,000 patent documents across this source. This classification deals with medical drugs and other such chemical compositions that are used in the treatment of a particular disease or condition. Naturally, the topic of this data set is highly relevant to the treatment relationship on which we will focus. Section 3.2 will detail how we use this patent corpus for training and testing.

The lexico-syntactic patterns we will obtain from section 3.1 can be used as a good indication of a treatment relationship, but the subject and object of the relationship need to be taken into consideration to be able to more accurately retrieve treatment relationships. For example, the sentence "Many clinicians recommend the use of Zoloft to treat depression." would return "Zoloft" and "depression" as the subject and object of a treatment relationship when using the pattern "to treat." However, the sentence "Currently, Zoloft is prescribed to many subjects to treat depression." would return "subjects" and "depression" as a possible treatment relationship using the same pattern, which would be false. Therefore, the patterns need classification rules to determine the accuracy of the treatment relationships returned. This process of devising classification rules is discussed in section 3.3.

## 3.1 Finding patterns highly associated with treatment relationships

To find patterns associated with treatment relationships, we perform the steps listed as follows:

*Step 1*:
Obtain sentences containing treatment relationships – We first manually construct a set of known treatment relationships in the form of a subject and an object. For example, acceptable treatment relationships are (Zoloft, Depression), (Xanax, Anxiety), and (Ibuprofen, Inflammation). The patent corpus is then searched for sentences containing both terms of the relationship.

In conducting this search, we obtain first the synonyms for the subject and object of the relationship and then use these synonyms in the search as well. We use the UMLS dictionary as described in [10][11] to retrieve synonyms for the terms. UMLS is a medical dictionary created by the National Library of Medicine and contains over 1 million terms.

*Step 2*:
Extract patterns from the treatment relationships – After the sentences are collected from the patent corpus in Step 1, we process these sentences to be able to easily extract the pattern connecting the subject and object of the relationship. In order to process the sentences in our patent corpus, we make use of the PCFG shallow parser in [8]. This shallow parser reduces words to their base forms, assigns part of speech tags to each of the words in a sentence, and attempts to chunk words together in phrases such as noun phrases, verb phrases, and prepositional phrases.

Then, the subject and object terms are located within the sentence and the verb phrase between the subject and object is returned. If more than one verb phrase exists between the subject and object, then we exclude all of the verb phrases because then each of the verb phrases would most likely not link the subject and object of the treatment relationship. For example, if we have the sentence structure $NP_1$ $VP_1$ $NP_2$ $VP_2$ $NP_3$ with $NP_1$ and $NP_3$ containing the subject and object, we can not take into account $VP_1$ and $VP_2$ as possible patterns because both these verb phrases do not link $NP_1$ and $NP_3$ together directly. Consider the sentence "Ibuprofen is taken by many elderly people, whom are more susceptible to arthritis." which has the corresponding shallow parse of [NP ibuprofen/n], [VP be/v take/v], [PP by/i], [NP many/j elderly/j people/n ,/,], [NP whom/w], [VP be/v], [ADJP more/r susceptible/j], [PP to/t], [NP arthritis/n ./.]. If the subject and object of the relationship is "Ibuprofen" and "arthritis" respectively, we cannot use the two VPs between them because [VP be/v take/v] and [VP be/v] do not directly link [NP ibuprofen/n] and [NP arthritis/n ./.] together, but rather involve [NP many/j elderly/j people/n ,/,].

*Step 3*:
Compiling a list of treatment verb patterns – Now that we have a collection of verb phrases linking together known treatment relationships, specific treatment verbs need to be identified from the verb phrases. This is because searching for other treatment relationships using the whole verb phrase is too restrictive, so only individual verbs will be retained. For example, one of the verb phrases returned was [VP significantly/r reduce/v] of which we just retain the verb "reduce." We perform the process of going through all of the verb phrases returned from Step 2 and manually select the key verbs from the verb phrases that would highly indicate a treatment relationship. Some of the treatment verbs obtained during this step include: treat, alleviate, relieve, reduce, prevent, and inhibit.

## 3.2 Constructing the training corpus and test corpus

Once we have a list of highly indicative treatment verbs, we need to construct classification rules that tell us the accuracy of a possible treatment relationship given a subject, treatment verb, and object. In order to devise these classification rules, we use a corpus of positive and negative treatment relationships, created from the U.S. Patents classification "Drug, bio-affecting and body treating compositions" data source. The steps in creating these corpora are detailed as follows:

*Step 1*:
Retrieving possible treatment relationships - We first search the original patent corpus for relevant sentences, which are ones that contain a treatment verb found in section 3.1. Then, we process these sentences using the PCFG shallow parser as presented before. From these sentences, the noun phrases surrounding the verb phrase with the treatment verb are extracted. We also check to see if there are certain prepositional phrases that follow directly after the verb phrase which would indicate that a sentence is passive.

For example, the sentence "Depression is often treated with Zoloft." is passive which is indicated by the prepositional phrase

**Table 1. The final set of classification rules produced in section 3.3 with percentage of accuracy.**

| NP1 Class | Verb | NP2 Class | Accuracy |
|---|---|---|---|
| act | treat | entity | 79.40% |
| entity | inhibit | act | 63.00% |
| entity | * | state | 58.10% |
| * | cure | entity | 50% |
| act | * | state | 59% |
| act | inhibit | act | 61% |
| act | reduce | group | 50% |

"with" following the verb. In passive sentences, the subject will be searched for in noun phrases after the verb and the object will be searched for in noun phrases before the verb. The opposite will be true in the case of active sentences. For each of the noun phrases, only the nouns will be kept as part of the relationship. For example, the noun phrase [NP cancer/n and/c] will have the conjunction "and" dropped and the noun "cancer" retained. After this process is done, we will have a list of possible treatment relationships in the form of a subject term, a treatment verb, and an object term.

Step 2:
Setting up and tagging the corpora - For our purposes, we retain a total of 1250 possible treatment relationships extracted from a random set of sentences in the 50,000 patent documents we collected. We divide this set by using the first 1000 as our training corpus and the last 250 as our test corpus. All of the relationships in the training corpus and test corpus are then manually tagged as positive (a correct treatment relationship) or negative (an incorrect treatment relationship). In determining the tag for a possible treatment relationship, the sentence from which the relationship came from is used for context.

## 3.3 Constructing the classification rules

This section details the process of constructing classification rules for our treatment patterns in order to achieve better precision. We use the C4.5 decision tree learning algorithm [12] developed by Quinlan and the WordNet semantic lexicon [9] to perform this task. The steps involved are:

*Step 1*:
Producing the training data for C4.5 from the training corpus – The C4.5 statistical classifier takes in training data in the form of a set of variables with one designated as the target variable (the variable that the classifier is trying to determine based on the other variables). The output will be a set of rules, each comprising of variable assignments that give percentage accuracy for determining the target variable.

The target variable in this case will be whether the relationship is a treatment relationship or is not. For the other input variables, one will be the treatment verb of the relationship, which is one of the verbs discovered from section 3.1. The subject and object of the relationship will also be input variables to the classifier, but we will use their hierarchical noun class instead of the actual subject/object term. This allows us to have these variables only have a select number of distinct values, which allows for better classification results. The hierarchical noun classes we use are: act, possession, group, event, state, phenomenon, abstraction, and entity. The majority of terms encountered can be categorized into one of these classes, as determined in [1].

An example training data input is: (entity, cure, state, 1). The first and third variable are hierarchical noun classes of the subject and object respectively. The second variable is the treatment verb, and the fourth variable is the target that is 1 for a correct treatment relationship and 0 for an incorrect treatment relationship. The C4.5 classifier will take in a set of training data input in this form and output rules to classify the target variable of other treatment relationships. An example classification rule output is (entity, *, state, 1, 58.10%). The first and third entries denote the hierarchical noun classes for the subject and object respectively. The second variable is again the treatment verb, which is * representing a wildcard. The fourth variable is the target variable and the fifth is a percentage accuracy for classifying the treatment relationship as the target.

The WordNet semantic lexicon is used to find the hierarchical noun classes described above. WordNet contains a database of semantic relations, one of which is the hypernym. A hypernym is defined as a word that is more generic than a given word. For example, the hypernym of a dog is animal and the hypernym of triangle is polygon. Naturally, this WordNet relation can be used to find hierarchical noun classes. The algorithm for doing so iteratively uses the hypernym relation on a noun until it encounters one of the set categories described above, or returns a ? if none of the categories are traversed. C4.5 uses the symbol ? for undefined variable values. For example, when traversing hypernyms of the term "Zoloft," the first hierarchical noun class found that is contained in the set is "entity."

*Step 2*:
Producing the final set of classification rules – We use 4-fold cross validation in combination with the C4.5 statistical classifier to produce and evaluate the classification rules. In other words, we partition the 1000 relationships in the training corpus into 4 separate sets and run the C4.5 algorithm 4 times, once for each set. For each set, we use the other 3 sets for the training data to produce the rules and the current set as the validation data to test the rules. Therefore, each set is used exactly once for validation. 4-fold cross validation allows training data and validation data to be exclusive from one another, which is necessary in order to retrieve unbiased results.

After running C4.5 on each of the 4 sets, we combine all of the rules and select only the rules that meet a certain criteria. We retain the rules that have a 50% or greater accuracy of determining a treatment relationship and appear in at least half of

**Table 2. Examples of queries and retrieved sentences in treatment relationship search.**

| |
|---|
| **Query #1 (One term):**<br>Anxiety |
| **Retrieved sentences with query #1:**<br>**Benzodiazepines** may **relieve anxiety** associated with PTSD.<br>In one of its aspects the invention discloses the use of **FAAH inhibitors** as useful in **treating anxiety** and depression. |
| **Query #2 (One term):**<br>Fever |
| **Retrieved sentences with query #2:**<br>The composition and method utilize a **nonopioid analgesic** and an **endothelin antagonist** as active agents to **treat fever** in mammals, including humans.<br>**Corticosteroids**, such as **prednisone**, **reduce fever** and diarrhea and relieve abdominal pain and tenderness. |
| **Query #3 (Two terms):**<br>Drug, Diabetes |
| **Retrieved sentences with query #3:**<br>Pioglitazone hydrochloride, (ACTOS.RTM.), is an active ingredient for a commercially available **drug** employed to **treat diabetes** mellitus in a host.<br>The resulting compound (1) of the present invention exhibits superior antidiabetic action and lipid reducing action, and is useful as a **drug** for **treating** or **preventing diabetes**, hyperlipidemia, and obesity. |
| **Query #4 (Two terms):**<br>Antibody, Inflammation |
| **Retrieved sentences with query #4:**<br>In this example, **antibody** against MAC-1 to **prevent inflammation** was used, however, similar results can be obtained using substances to block a wide array of these molecules.<br>As a result, it has been confirmed that, like the results of the experiment obtained by using the **antibody**, the erythropoietin receptor protein also **suppresses inflammation**. |

all of the runs. The rules that are kept, as listed in Table 1, will be used as the final set of classification rules.

The rules contained in Table 1 exhibit the hierarchical noun classes act, entity, state, and group. Act refers to an action such as surgery or exercise. Entity contains tangible objects such as items and living things. Drug products are contained in the entity hierarchical noun class. State is defined as the condition of a person or thing, as with respect to circumstances or attributes. State contains most medical diseases and conditions. The last hierarchical noun class used in the rules is group, which refers to a combination of entities.

## 4. RESULTS

We performed two separate evaluation schemes. First, we evaluated the classification rules produced in section 3.3 on our test corpus that we constructed in section 3.2. Then, we evaluated the final system's performance of retrieving treatment relationships from our full patent data source. We elaborate on these two schemes as follows:

*Evaluation 1*:

Evaluating the classification rules - To evaluate the classification rules produced in section 3.3, we apply these rules to the test corpus consisting of 250 relationships, which were previously tagged as to whether they were treatment relationships or not. When applying the classification rules to the 250 relationships, 72 treatment relationships were found. Out of these 72 retrieved relationships, 59 of them were found to be correct as determined by the tagging done in section 3.2, which translates to a precision of 81.94% (59 correctly retrieved relationships / 72 retrieved relationships). From the 250 total relationships in the test corpus, there were a total of 123 correct treatment relationships according to the tagging. Therefore, the recall of the classification rules is 47.97% (59 correctly retrieved relationships / 123 total correct relationships).

Thus, these classification rules have shown to produce high precision of 81.94%, but a lower recall of only 47.97%. We suspect that one of the reasons for the lower recall percentage is that a good number of the subjects and objects of the treatment relationships are highly technical medical terms, for which it is difficult to find hierarchical noun classes for. This is because WordNet only contains a fixed number of terms, so terms that it does not contain will not have hypernym relations associated with them. The subjects and objects in the training set that did not get classified into a hierarchical noun class would have an undefined input variable for the C4.5 classifier. These relationships would then be unable to contribute to rules that don't have wildcard variables, thus leading to low recall due to the lack of classification rules obtained.

One way to improve the recall value would be to increase the size of the training corpus, which would allow more training data for the C4.5 classifier. This would lead to an increase in the number of classification rules obtained since the classifier would have more training data to work with. With an increase in the number of reliable classification rules, the system will be able to retrieve a larger number of correct treatment relationships, and therefore, obtain a higher recall. However, due to the time intensive process of manually tagging the training corpus, we have not yet created a larger corpus.

*Evaluation 2*:
Evaluating the patent retrieval system – This process of finding treatment relationships has been adapted into a system that finds patent documents determined by a user query as part of our R&D project. The user may enter a query in the form of one keyword term or two keyword terms. If the user enters in one keyword term, the system will find and return treatment relationships with either the subject or object containing the term specified. If the user enters in two keyword terms, the system will find and return treatment relationships with one of the terms as the subject and the other term as the object. For example, if the user queries the system with the term "depression," the system will use the treatment verbs to find patterns linking "depression" with other terms that will complete a treatment relationship where one is the subject and the other is the object. In this case, the system would return subjects of the relationship such as "Zoloft" and "Prozac," both of which are antidepressants used in treating clinical depression. The system will also retrieve the sentence and corresponding patent document identification number of the patent from which the relationship is found.

We evaluated this patent retrieval system using our base patent corpus of 50,000 patent documents from the U.S. Patents classification "Drug, bio-affecting and body treating compositions." 20 test queries were used as input to the system, which are a combination of one and two keyword queries. For these queries, we used various drugs and activities as the subject and their corresponding conditions and diseases that they treat. Treatment relationships are most prominent in this form. Examples of queries used in the test set as well as retrieved sentences can be seen in Table 2.

After running the system on this test set of 20 queries, 1825 possible treatment relationships were returned of the form $NP_1$ VP $NP_2$ that were not yet filtered by the classification rules. These relationships were then manually evaluated as to whether they were correct or not. From these 1825 treatment relationships, 987 were determined by the system to be correct based on the final set of classification rules. Out of these 987 relationships, 847 of them were also determined to be correct from the manual evaluation. Therefore, the precision of the system is 85.81% (847 correctly retrieved relationships / 987 retrieved relationships). From the set of 1825 possible treatment relationships that were not yet filtered by the classification rules, manual evaluation determined that 1024 of these were correct. This evaluates to a recall of 82.71% for the system (847 correctly retrieved relationships / 1024 total correct relationships).

The precision of this system achieved a high percentage, which can be attributed to the high precision of the classification rules. The recall is also exceptionally high in this case for two reasons. The first reason is that these queries had the subject, object, or both constrained to be actual terms related to treatment. Due to this, the results naturally conform more to the final set of classification rules than to the relationships used in the classification rules evaluation. We chose to use query terms related to treatment in this evaluation and not random terms because it would most resemble actual queries from users searching for prior art. The second reason is that this recall is not the global recall. This method for detecting treatment relationships focuses exclusively on the structure $NP_1$ VP $NP_2$ as discussed before, so the global recall would be much lower. This is because there are treatment relationships not of the structure we are focused on which are unaccounted for. For example, the sentence "There are many antidepressants that can be used for treating clinical depression, one of which is Zoloft," contains the treatment relationship (Zoloft, Depression), but would not be found because it is not of the structure $NP_1$ VP $NP_2$. However, the high precision of this retrieval system is beneficial for users looking to retrieve results with low amounts of false positives.

One particular side effect of searching with only one keyword query is that either the subject or object of the relationship will not be constrained in the search. This causes the system to retrieve more generic terms such as "method," "agent," and "technique." These terms may not provide the user directly relevant information, although most of these relationships can be still considered treatment relationships. The user is also presented with the patent document identification number so they can inspect nearby sentences for co-referent terms. For example, consider the two adjoining sentences "Selective serotonergic reuptake inhibitors are used in treating a variety of mental disorders. In particular, they are useful as an agent for treating or preventing depression." From the second sentence, the treatment relationship (agent, depression) can be found. The term agent is generic because it does not provide the user with any interesting information directly. Upon inspecting the sentence before it, the user will find out that the agent is referring to selective serotonergic reuptake inhibitors, which would provide more useful information.

## 5. CONCLUSIONS

There has been much work done in relationship detection, especially common relationships like causality. However, more specific relationships such as treatment relationships have not been dealt with as heavily. In this paper, we have shown a reliable process of detecting treatment relationships from a data source using lexico-syntactic patterns. This process was determined to achieve a high precision percentage through the use of learned classification rules. As a major goal, we have also incorporated this process into a patent retrieval system that can find patent documents containing specific treatment relationships determined by the user. This process retrieves documents based on a semantic relationship instead of on just keywords, as many patent retrieval systems are based on. The goal of this system is to retrieve patent documents that are more closely related to what the user is actually searching for, which is of great interest to patent examiners and applicants as they search for prior art. Moreover, the method presented in this paper is highly adaptable to other types of semantic relationships. This adaptability greatly expands the possibility and applicability of semantic patent search.

## 6. REFERENCES

[1] R. Girju, "Automatic detection of causal relations for question answering", Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering, Volume 12, Pages: 76-83, 2003.

[2] Leah S. Larkey, "A patent search and classification system", International Conference on Digital Libraries archive, Proceedings of the fourth ACM conference on Digital libraries table of contents, Berkeley, California, United States, Pages: 179-187, 1999.

[3] L. Larkey, M. Connell, and J. Callan. "Collection selection and results merging with topically organized U.S. patents and TREC data", In Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM), ACM, Pages: 282-289, 2000.

[4] Sougata Mukherjea, Bhuvan Bamba, "BioPatentMiner: an information retrieval system for biomedical patents", Proceedings of the Thirtieth international conference on Very large data bases, Volume 30, Pages: 1066-1077, 2004.

[5] USPTO, Performance and Accountability Report Fiscal Year 2007, Page: 109, 2007.

[6] Leo Wanner, Ricardo Baeza-Yates, Sören Brügmann, Joan Codina, Barrou Diallo, Enric Escorsa, Mark Giereth, Yiannis Kompatsiaris, Symeon Papadopoulos, Emanuele Pianta, Gemma Piella, Ingo Puhlmann, Gautam Rao, Martin Rotard, Pia Schoester, Luciano Serafini, Vasiliki Zervaki, "Towards Content-Oriented Patent Document Processing", World Patent Information, Elsevier, Vol. 30(1), Page: 21-33, March 2008.

[7] Tianhao Wu, Lars E. Holzman, William M. Pottenger, Daniel J. Phelps, "A Supervised Learning Algorithm for Information Extraction from Textual Data", In the proceeding of the workshop on Text Mining, Third SIAM International Conference on Data Mining, 2003.

[8] Xuan-Hieu Phan. Jtextpro: A java-based text processing toolkit.

[9] C. Fellbaum. Wordnet: an electronic lexical database. Mit Pr, 1998.

[10] O. Bodenreider and O. Journals. The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Research, 32(90001):267–270, 2004.

[11] DA Lindberg, BL Humphreys, and AT McCray. The unified medical language system. In Methods of Information in Medicine, volume 32, pages 281–91, 1993.

[12] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.