Medical Relation Extraction with Semantic Grammars

Thomas B. Røst, MSc; Saiful Akbar, PhD; Øystein Nytrø, PhD; Márcio Basgalupp, PhD
Norwegian University of Science and Technology, Trondheim, Norway

Abstract

For the 2010 i2b2/VA relation extraction challenge a semantic grammar-based system was developed. Precision and recall for identifying relations between concepts was 67.8% and 56.1%, and 60.6% and 50.2% for also recognizing the correct relation type.

Introduction

The 4th i2b2/VA challenge was split into three sub tasks: (1) Extraction of medical problems, tests and treatments, (2) classification of assertions made on medical problems, and (3) relations of medical problems, tests and treatments[1]. This paper describes the participating entry from the Norwegian University of Science and Technology (NTNU) team, which focused on the relation extraction task.

The relation extraction task assumed that concepts and assertions about medical problem concepts were known in advance. Concepts were defined as belonging to three general semantic categories: (1) *medical problems*, (2) *treatments*, and (3) *tests*. *Medical problems* include observations about the patient that are related to a disease or some abnormality; *treatments* are all interventions done in order to alleviate a problem; *tests* are all procedures and observations performed to determine if a problem exists.

For each concept belonging to the medical problem category, an assertion about the type of problem was available. The assertion categories were (1) *present*, (2) *absent*, (3) *uncertain*, (4) *conditional*, and (5) *not associated with the patient*.

For the relation extraction task, the first challenge was to determine the concept pairs in a sentence that constituted a relationship. The second challenge was to determine which of 8 relationship types the relation belonged to: (1) *TrIP* (treatment improves medical problem; (2) *TrWP* (treatment worsens medical problem); (3) *TrCP* (treatment causes medical problem); (4) *TrAP* (treatment is administered for medical problem); (5) *TrNAP* (treatment is not administered because of medical problem); (6) *PIP* (medical problem indicates medical problem; (7) *TeRP* (test reveals medical problem); and (8) *TeCP* (test conducted to investigate medical problem).

As part of the challenge description, a detailed set of guidelines described the full set of inclusion and exclusion criteria for concepts, assertions and relations.

Related Work

Burton claimed to introduce the concept of semantic grammars in a 1976 technical report[2]. Haug et al. used a semantic parser to extract concepts from events described in chest radiographs[3]. They report an accuracy of 81% in finding primary concepts from a set of 10 radiograph reports, but their system tended to generate many false positives.

Methods

After considering and testing various approaches towards finding relations between medical concepts, we decided on a purely rule-based method. For each relationship class a number of grammar rules were created manually based on the given inclusion/exclusion criteria and examples. These rules described the different ways of expressing a particular relation between a concept pair. The top-level, non-terminal categories were semantic or conceptual rather than syntactic, meaning that they focus on the type of relationship rather than syntactic features. In this sense, our grammars may be described as *semantic grammars*.

Semantic grammars were introduced by Burton[2] as a means of organizing knowledge when doing parsing of natural language. The criteria for using semantic grammars are a limited domain, a limited number of activities within the domain, and that the conceptualizations of the domain are or can be known. It is also assumed that there is a benefit to including semantic knowledge into the parsing process.

For the relation extraction challenge, it can be argued that these criteria are partially met. Although the number of possible relations between medical problems, treatments and tests is likely to be very large when dealing with specific instances of medical concepts, we assumed that the number of ways of describing relations between concepts (e.g. how one concept influences another concept) will in many cases be independent of the specific concepts. Since the concepts were known this was considered a viable

approach. Moreover, through previous work we had observed that there were recurring syntactic patterns when patient treatment was documented[4].

Another key characteristic of our solution is that we did not use any medical knowledge or resources beyond our assumptions about the semantics of what is being communicated. All concepts in the text were abstracted to a small number of general concept types. As will be shown, this made it possible to represent a large number of relationship types with a fairly limited set of rules.

Note that there was also a pragmatic side to this decision. From a medical point of view, it is reasonable to assume that some medical concept pairs are more likely to occur as part of a relation than other. Such knowledge could prove valuable for e.g. detecting obviously false relations. However, the team did not include members with medical expertise. Time constraints also prevented us from incorporating this dimension into the system through the use of resources such as UMLS. In spite of this limitation, it was nonetheless of interest to see how well such a naïve solution would perform.

Step 1: Data transformation and analysis

At the core of the training data was a set of anonymized discharge and progress notes from Partners Healthcare, Beth Israel Deaconess Medical Center and University of Pittsburgh Medical Center[1]. Each discharge note was contained in a single text file with one sentence per line. Each sentence had been tokenized so that whitespace was inserted between each token. For each discharge note there were three additional files containing the concepts, the assertions and the relations, all with references to the line and token positions in the original text.

To learn about the different types of relations and to understand language use and sentence structure for the relation types, we aggregated the information available to us in two new representations. The first representation was a new set of discharge notes where all the relationship spans were annotated. This was useful for group discussions and for understanding the choices made by the human annotators. The second representation was a set of eight files, with one file for each relationship type. Each file contained all text segments with the context of a relation pair for the given relation, sorted according to frequency of occurrence. For this to be useful, all concepts were transformed into a set of concept tokens for all possible concept/assertion types: *<TE>* for tests, *<TR>* for treatments, and *<PR-PRE>*, *<PR-ABS>*, *<PR-CON>*, *<PR-HYP>* and *<PR-POS>* for each problem/assertion tuple.

For the *TeRP* class, the top five segments were as follows:

```
75:    <TE> showed <PR-PRE>
45:    <TE> revealed <PR-PRE>
24:    <TE> was <PR-PRE>
24:    <PR-PRE> with <TE>
24:    <PR-PRE> on <TE>
```

From this representation an initial set of semantic type candidates were manually created for each relationship type. As an example, the two top rows tells us that there is a semantic rule

(<TE>, is_indicating, <PR-PRE>)

that holds true for a number of *test* and *present problem* relations.

It would later turn out that a representation where only the text between concepts was used was too simplistic in the sense that many important qualifiers (e.g. negation indicators and action verbs) were lost. To resolve this, a similar set of aggregated, frequency-sorted files were created, only this time with a segment window that included up to three words before and after the starting and ending concept. The corresponding first five segments for the *TeRP* class are shown below:

```
15:    <TE> showed <PR-PRE> .
13:    <TE> revealed <PR-PRE> .
9:     <TE> showed no <PR-ABS> .
5:     <TE> showed no evidence of
            <PR-ABS> .
5:     <TE> demonstrated no evidence of
            <PR-ABS> .
```

The first number of each line corresponds to the number of times that the segment appears. As expected, the frequency counts are considerably lower. For both aggregations, there was a long tail of segments occurring only once. This indicated that further simplification of the source text was necessary in order for manual grammar construction to be a viable strategy.

We observed that there were a number of words occurring between relations that could be described as more representative than others when determining the type of relationship, and that these words tend to occur more frequently. Correspondingly, other words appear to add little value to the relationship description, and are as such candidates for removal. Similar observations have been made from research on unsupervised relation discovery[5]. From a combination of manual review, prior assumptions and iterative refinement from classification results, a

vocabulary containing words thought to be most relevant for the set of medical relations was created. The text in the training data was then filtered to only allow vocabulary words and concept tags.

As an example, the original sentence (concepts are highlighted)

"**On physical examination** , patient is in no **acute distress** , **afebrile** , **blood pressure** 134/80 , **heart rate** 80 and regular , no **bruits** ."

was transformed into the representation

"<TE> in no <PR-ABS> <PR-ABS> <TE> <TE> and no <PR-ABS>"

We observed that many traditional stop words turned out to be relevant for distinguishing between relationship classes, so our vocabulary would include many words typically removed when doing text classification. We also refrained from stemming and lemmatization, as it was not known at the time if word inflection would influence the semantics.

After simplification, all concept pair permutations were extracted from each sentence. A pair consisted of the start and end concept and up to four words before and after the concepts. This was assumed to be a sufficient context for determining the relationship type. Each pair would then be tested on the implemented semantic grammars.

Step 2: Iterative grammar implementation

Eight semantic grammars—one for each relationship type—were built manually. Grammar rules followed a simple two-level hierarchy. The first level specified the concepts and a mixture of semantic and syntactic rules necessary to bridge the concepts. The second level contained the lexical information necessary to match the semantic rules with the allowed vocabulary.

An example of two top-level rules, in this case for the *TeRP* category, is shown below:

```
(test_list, ws,
       is_because_of, ws,
       problem_present_list,
       (ws, is_showing,
        ws, problem_present_list)?)

(test_list, ws,
       is_done_for, ws,
       problem_absent_list, ws,
       is_not_showing)
```

The token `ws` indicates whitespace, while the question mark follows a single token or a group of tokens that may or may not occur. For the sake of reference, 11 top-level rules such as these were implemented for the *TeRP* class in time for the competition due date.

We focused on implementing support for the simplest and most frequently occurring relation semantics, and iteratively added support for the more complex statements. All grammars were implemented as standard EBNF (Extended Backus-Naur Form).

The second-level lexical rules were deliberately kept very simple. Constructing exhaustive, specific grammars for natural language by hand would quickly lead to complex, convoluted, multi-level hierarchies that would be difficult to both understand and modify without breaking previous functionality. We approached this problem by using an extended bag of words model where some words were defined as must-occur, or primary, and the rest as may-occur, or secondary. A simple example for the *is_because_of* semantic relation is as follows:

```
is_because_of := (
       (is_because_of_sec, ws)*,
        is_because_of_pri,
        (ws, (is_because_of_sec /
              is_because_of_pri))*)

is_because_of_pri :=
       'in light of' / 'due to' /
       'because' / 'while'

is_because_of_sec :=
       'continued' / 'increase' /
       'likely' / 'to' / 'of' /
       'performed'
```

This rule says that a primary word must occur, that it may be preceded by an arbitrary number of secondary words, and that it may be succeeded by an arbitrary number of primary or secondary words. Similar rules were defined for all top-level relations.

Step 3: Grammar application

All of the extracted and simplified concept pair permutation strings were sent to each relation class grammar. If a grammar rule matched a string, the relation class was recorded as a candidate class. The grammars were greedy by default, meaning that the longest possible match would be returned.

For cases where multiple candidate classes are returned, a number of optional rules were applied to attempt a possible class resolution. These rules would either take into account the specificity of the classes (e.g. *TrWP* is more specific than *TrAP*) or the class distribution (e.g. *TeRP* is more common than *TeCP*). If no resolution was possible, the match was ignored. Finally, the results were output as a relations file in the i2b2 challenge format.

Results

The size of the allowed vocabulary was 276 words. A total number of 80 top-level semantic rules were

crafted. The reference data set consisted of 477 discharge notes with a total of 9,070 relations. Our system extracted 7,507 relations. Table 1 shows the overall precision, recall and F-measure for detection of exact relation spans and detection of exact spans with a matching relation, while Table 2 and Table 3 show the corresponding results for each relationship class.

	Exact span w/ matching concept	*Concept w/ matching span and relation*
True positive	5,089	4,551
False negative	3,981	4,519
False positive	2,418	2,956
Recall	0.561	0.502
Precision	0.678	0.606
F-measure	0.614	0.549

Table 1. Exact span for all relationships.

	Recall	*Precision*	*F-measure*
TrIP	0.379	0.893	0.532
TrWP	0.343	0.942	0.503
TrCP	0.367	0.558	0.443
TrAP	0.653	0.723	0.686
TrNAP	0.340	0.833	0.483
PIP	0.513	0.501	0.507
TeRP	0.603	0.767	0.676
TeCP	0.452	0.773	0.571

Table 2. Exact span for separate relationships (exact span true positive).

	Recall	*Precision*	*F-measure*
TrIP	0.207	0.631	0.312
TrWP	0.007	0.125	0.013
TrCP	0.279	0.308	0.293
TrAP	0.583	0.656	0.617
TrNAP	0.120	0.348	0.179
PIP	0.513	0.501	0.507
TeRP	0.597	0.712	0.650
TeCP	0.139	0.463	0.214

Table 3. Exact span for separate relationships (exact span and matching relations).

The full run time was 36 minutes on Ubuntu Linux version 9.04 running in a virtual VMware instance with 2GB of dedicated memory. The host operating system was Windows 7 64-bit running on a Lenovo ThinkPad X200s with 4GB of total memory and an Intel Core2 Duo CPU running at 1.86GHz. All software was implemented and run on Python version 2.6.2.

Discussion

The precision for finding correct relationship span varies between 50.1% for the *PIP* class to 94.2% for the *TrWP* class, while the recall is in the range of 34.0% to 65.3%. The variation in precision and recall is considerably higher when requiring correct relation matches as well, with precision ranging from 12.5% to 71.2% and recall from 0.7% to 59.7%. In this case, the relatively small *TrWP* class has particularly bad performance.

With the exception of the *PIP* and *TrCP* classes, the precision when extracting relationship spans seems fairly good. The low recall can partly be explained by time constraints; there appears to be a positive correlation between the grammars that were the most refined at the time of results submission and the corresponding class recall, as shown in Figure 1. To get good overall results, focus was given to the classes that were most prevalent in the training data. In addition, the lack of training material for some of the classes meant that it became more difficult to identify which rules would have the most impact.

Precision and recall for having both the correct relation span and class is somewhat lower. We suspect that this is partly related to the crude way of resolving multiple grammar matches. The original plan was to implement a second layer of classification using machine learning approaches that would take the information from the grammars into account, but we did not have time to implement this before the deadline.

We observed that very simple grammar rules tended to generate too many false positives, and therefore tried to keep the rules very specific. For this reason, our system did not fare well with the some of the "simpler" ways of documenting relations, such as a problem followed by a list of treatments, with little or no intervening filler text. It was also evident that we oversimplified our language representation by excluding all punctuation, thereby missing important sentence markup such as commas. An effect of this was that rules would sometimes get subordinate clauses mixed up with each other.

Stemming and lemmatization was not done as it was not known if inflectional forms would be necessary to distinguish between semantic types and classes. The presence of inflectional forms turned out to be unnecessary, and in retrospect it seems that stemming would have simplified the vocabulary and grammar rule creation a great deal.

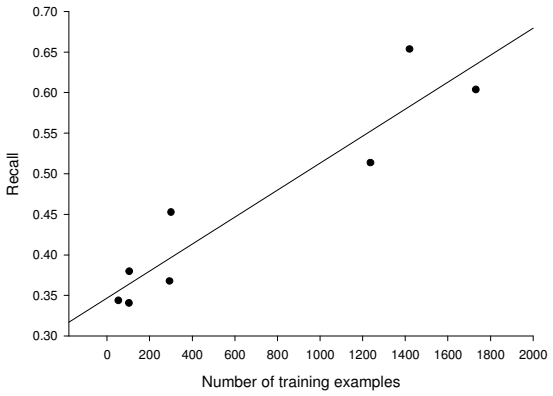

Figure 1. Relation class recall vs. number of training examples.

Future Work

It seems reasonable that including more information about the medical concepts into our rules and vocabulary, either verbatim or as higher-order concept hierarchies, could be a worthwhile effort. As noted, we do not take advantage of facts about e.g. concepts that normally occur together for given medical problems.

Our method of building semantic grammars can be time-consuming, although not prohibitively so given a structured approach towards the task. Even so, investigating ways of automatically augmenting both grammar rules and vocabulary, such as automated grammar induction, relation discovery[5] and latent semantic analysis[6], should be investigated.

As mentioned, another improvement would be to extend our symbolic approach with subsymbolic techniques so that the information provided by the grammars is used as features for automated learning and classification of relations, especially since the current approach seems to work fairly well for isolating the relationship spans.

We did not take part in the concepts extraction challenge, but an interesting extension of our system would be to use it in "reverse" in order to identify concepts by assuming relations.

Finally, implementing support for parallelizing the classification would be an important measure to reduce the run-time of the classification task. Rapid feedback is a key requisite of manual grammar-building.

References

1. Fourth i2b2/VA Shared-Task and Workshop: Challenges in Natural Language Processing for Clinical Data. Available at: https://www.i2b2.org/NLP/Relations/Main.php.

2. Burton RR. Semantic Grammar: An Engineering Technique for Constructing Natural Language Understanding Systems. 1976.

3. Haug P, Koehler S, Lau LM, et al. A natural language understanding system combining syntactic and semantic techniques. *Proceedings of the Annual Symposium on Computer Applications in Medical Care*. 1994:247-51. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2247803&tool=pmcentrez&rendertype=abstract.

4. Røst TB, Edsberg O, Grimsmo A, Nytrø Ø. Comparing medical code usage with the compression-based dissimilarity measure. *Studies in health technology and informatics*. 2007;129(Pt 1):684-8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17911804.

5. Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Barcelona, Spain: Association for Computational Linguistics; 2004:415-422.

6. Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*. 1997;104(2):211-240. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.104.2.211.