# A Deep NLP System for Extracting Knowledge from Clinical Text: Application to the 2010 i2b2/VA Concept Extraction Task

**Lucian Galescu, PhD, Nate Blaylock, PhD, James F. Allen, PhD, William de Beaumont, Hyuckchul Jung, PhD, Mary Swift, PhD**
**Institute for Human and Machine Cognition, Pensacola, FL**

## Abstract

*We describe recent work on augmenting a general-purpose, deep natural language understanding system with key lexical and ontological medical information in a system for extracting knowledge from clinical text. We describe briefly our base system, and then discuss how we adapted it to compete in the 2010 i2b2/VA concept extraction task.*

## Introduction

We have recently embarked on a project that applies our existing deep natural language processing (NLP) technology to the problem of extracting, from a patient's medical record, events related to the course of their disease and treatment. The ultimate goal is to use the extracted information to support Comparative Effectiveness Research (CER) for cancer treatment.

Given the range of concepts of interest as well as the complexity of extracting various relations  (temporal, causal, agentive, telic, etc.) between them, we believe a deep natural language understanding (NLU) approach is required, where the meaning of the full text is accounted for in a rich semantic representation. In this respect, our approach stands in contrast to most current approaches to information extraction (IE) for the biomedical domain, which typically rely on shallow NLP techniques (pattern matching, chunking, templates, etc.)[1].

In this paper, we first briefly describe our preexisting system. We then describe the specific changes and additions to adapt our system for the concept extraction task included in the 2010 i2b2/VA NLP Challenge. We then discuss results on the task and conclude with areas of future work.

## The TRIPS System

Our system has its roots in the TRIPS general-purpose NLP framework, which has benefitted from over a decade of research in discourse analysis and dialogue systems[2]. In recent years, the NLU components of TRIPS have also been applied to the semantic analysis of text[3,4]. Figure 1 depicts these components, as used in our CER system.

At the core of the system is a packed-forest chart parser; it uses a detailed, hand-built, lexicalized context-free grammar of English, augmented with feature structures and feature unification. The parser draws on a general-purpose semantic lexicon and ontology which define a range of word senses and lexical semantic relations. The core semantic lexicon has been constructed by hand and contains more than 7000 lemmas. The lexicon can be dynamically augmented with new words by consulting WordNet[5]. The lexicon uses the WordNet entries along with mappings we established from WordNet synsets into our semantic language ontology to generate compatible underspecified representations on the fly.

The TRIPS ontology defines a rich set of semantic features that are crucial for constraining ambiguity at multiple levels of language processing. For example, the grammar uses selectional restrictions to guide word sense disambiguation and prepositional phrase attachment during parsing, and reference resolution uses the semantic features to identify valid referents and discard invalid ones. The ontology is designed to be linguistically motivated and domain independent.

Of note, the parser can be easily extended with external NLP tools in its front end. This capability is used to plug in a suite of shallow NLP tools (POS taggers, multiple named entity recognizers, and statistical parsers) that can provide lexical and/or structural advice during the construction of the parse. In particular, this is the locus where medical lexical and ontological information is inserted into the system to support the processing of clinical text (more on this in the next section).

The parser outputs a deep semantic representation of the input text called a Logical Form (LF), which captures the meaning of all the words in the input text (utterance, paragraph, or document). This representation is rich enough to support subsequent use of the information to produce knowledge and then for inference/reasoning.

One of the often repeated claims against using full, general parsers for handling biomedical text is that their performance is not robust. Robustness is a feature of the TRIPS parser, as it has been developed to handle ungrammatical input such as appears in spontaneous speech, as well as input affected by speech recognition errors. Specifically, when it cannot build an interpretation spanning the full

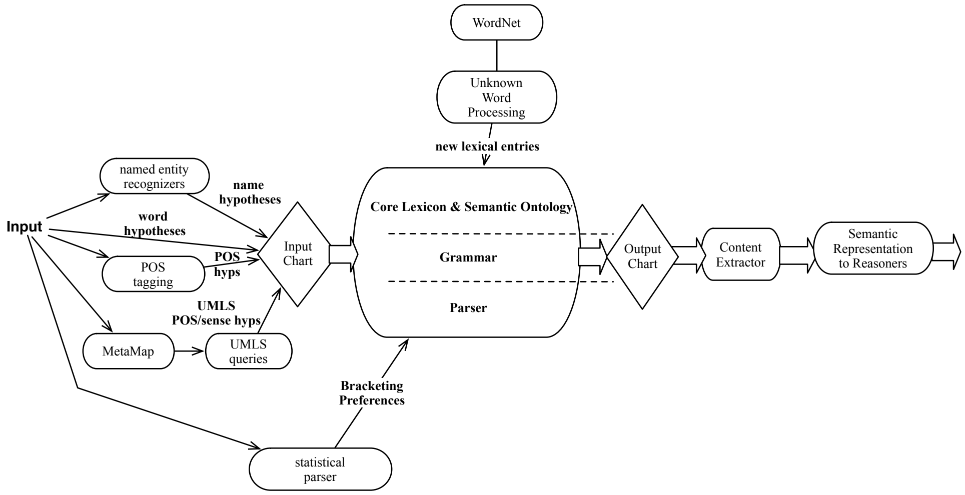

**Figure 1.** Natural Language Understanding Components

sentence, the parser identifies likely fragments, which, in many cases, are enough to construct a reasonable understanding of the text.

Most applications will not need the full LF representation for the input text. The Concept Extractor module extracts useful content from the LF based on application-specific rules tied into the ontology. This approach enables us to fully exploit the contextual information produced by our system's deep natural language understanding capabilities.

### Porting the System to the Clinical Domain

In October 2009, we began work on an NIH-funded project to support CER for cancer treatment, our role being to use NLP for extracting useful patient information from patients health records. Although TRIPS had been under development for over a decade, it had never been used for biomedical information extraction. However, the generic structure described above has proven useful for quick porting of the system to handle this new domain.

By far the most critical aspect in this process was the need for domain-specific lexical and ontological information. One widely used comprehensive resource that provides both is the National Library of Medicine's Unified Medical Language System (UMLS)[6]. UMLS was integrated into or system via MetaMap[7], a tool also developed by NLM, that can identify and rank UMLS concepts in text.

Specifically, we added MetaMap as a special kind of named entity recognizer feeding advice into the TRIPS parser's input chart (Figure 1). We run MetaMap twice on the input text to obtain UMLS information both for maximal constituents and for individual words in those constituents (for example, "lung cancer", as well as "lung" and "cancer"). We note here that the parser can entertain multiple hypotheses, even conflicting ones, about the same text span; it will evaluate all hypotheses and come up with the globally best answer.

The new words and phrases are added dynamically to the lexicon. However, there remains the issue of how the corresponding concepts fit in the TRIPS ontology in order to be represented in the resulting Logical Form. Our general approach for dealing with this problem uses an ontology specialization mechanism which we call *ontology grafting*, whereby new branches in the root ontology are created from third party ontological sources, and attached to appropriate leaf nodes in the TRIPS ontology.

The UMLS Semantic Network and certain vocabularies included in the UMLS Metathesaurus (SNOMED-CT[8] and the NCI Thesaurus[9] are of particular interest for our cancer-oriented CER system) define concept hierarchies along multiple axes. First, we established links between the 15 UMLS semantic groups[10] and corresponding concepts in our ontology. Second, we selected a list of nodes from the SNOMED-CT and NCI hierarchies (27 and 11 nodes, respectively) and formed ontological branches rooted in these nodes that we grafted onto our ontology. Finally, we implemented a blacklist/whitelist mechanism to remove from consideration UMLS concepts or (parts of) branches in the two hierarchies of interest. This mechanism was useful because for concepts that we consider "common", our own lexicon and ontology contain richer semantics and are therefore preferable. Whitelists are used to cancel the effect of blacklists; that is, smaller twigs from pruned branches may be grafted back to where the closest pruning cut was made. For SNOMED-CT, there are 24 concepts blacklisted and 16 concepts whitelisted. For the NCI Thesaurus, there are 27 concepts blacklisted and 48 concepts whitelisted.

Further details on lexical and ontological integration of UMLS data are included in a separate paper that has been submitted for publication[11].

**An example.** To illustrate the way our system works, consider the sentence "Oral intake was not adequate." Figure 2 shows the LF for this sentence, in a much simplified representation.

Although in general the word "adequate" might not be of interest, in this particular context it indicates an abnormal finding. The Content Extractor has a rule for such cases; the left hand side of the rule is shown in Figure 3, and can be interpreted as: "match any situation where a *bodily process* is said to *not* be *adequate.*" As a result of this rule's matching the LF in Figure 2, two extractions will be produced: one for a BODILY-PROCESS (spanning "oral intake") and one for a CONDITION (spanning "not adequate").

It is, of course, impossible to capture with one example the full range of capabilities afforded by our deep semantic representation. We note here that these semantic structures and the extraction algorithm allow hierarchical ontological matching, and modifiers (such as negations) and quantifiers can be extracted effortlessly.

### Adapting the System for the i2b2/VA NLP Challenge

The first task for the 2010 i2b2/VA NLP Challenge involved extracting three classes of concepts – medical problems, tests and treatments – from discharge summaries and progress notes. Since these are concepts of interest in our CER system, it appeared that we could easily adapt our system for the i2b2 task simply by ignoring all extractions that are not in one of these three classes.

In addition to the core NLP phase, we had to add: a) a pre-processing phase for preparing the i2b2 data for processing by our system; and b) a post-processing phase, for mapping our system's output into the i2b2 required format. In the following we will present conceptual and functional details about each of the three processing phases in our i2b2 system.

**Pre-processing.** The pre-processing phase involved three sub-processes:

*1. Text Cleanup*: Some changes were required to make sure the text is in a format and encoding that is not incompatible with downstream components' expectations. Some of the unfortunate slips in the development data included "</chief_complaint>", "&lt ;", "&amp ;", etc.; such tokens tend to cause problems when embedded in XML messages. It was particularly important to change "&apos;" back to apostrophe, since it is an important linguistic cue and wasn't handled properly by several NLP modules.

```
(f v1 (:* ont::HAVE-PROPERTY w::be)
        :theme v2 :property v4 :tense w::past)
(bare v2 (:* ont::BODILY-PROCESS w::intake) :mod v3)
(f v3 (:* ont::BODY-PART-VAL w::oral) :of v2)
(f v4 (:* ont::ADEQUATE w::adequate) :of v2 :mod v5)
(f v5 (:* ont::NEG w::not) :of v4)
```

**Figure 2.** Part of the LF for a sentence.

```
(f ?x1 (? type1 ont::HAVE-PROPERTY)
        :theme ?x2 :property ?x3)
(bare ?x2 (? type2 ont::BODILY-PROCESS))
(f ?x3 (? type3 ont::ADEQUATE) :mod ?x4)
(f ?x4 (? type4 ont::NEG))
```

**Figure 3.** Left hand side of an extraction rule.

*2. Content Filtering*: This component classifies each line in the report as header, footer and body, and inside the body, as section heading or content; all lines other than body-content and header-title are filtered out. This speeds up processing, and helps eliminate the false positives that may result from the erroneous interpretation of ambiguous words that appear in certain section headings.

We implemented a simple rule-based partition classifier, using a set of weighted heading identification pattern rules, supplemented with a continuity constraint. The list of header patterns was compiled from the i2b2 development data. Weights were assigned by hand.

On the development data, there was a very small number of content lines (ie, lines containing at least one annotated concept) filtered out because they were misclassified; on further inspection, it turned out that the majority were either annotated incorrectly (for example, services like "Medicine" and "Surgery" being annotated as treatments, "present illness" being classified as a problem in the "History of present illness" heading, etc.) or inconsistently (for example, the "physical examination" heading being classified as a test). Estimating how many non-content lines were misclassified as content lines is more difficult, since we don't have annotations at that level.

Although for the i2b2 test we discarded the section headings, we plan on eventually using them to provide context for interpreting the section bodies.

*3. Line Splitting*: This component was used to break extremely long lines (500-1000 characters) which exceed the tolerances of other components. It simply divides such long lines into smaller "clauses", broken at commas. This heuristic worked fine for the i2b2 data, where such long lines tended to be lists of tests. Because of its heuristic nature, we tried to minimize its use by using a very high line length threshold; it is likely that a lower threshold might have worked just as well, since true sentences rarely exceed 200 characters in length.

Eventually, this small, heuristic component will be replaced by a list processing tool that recognizes lists of similar things and parses them into items.

**NLP.** The only change to our CER clinical text NLP system was in the extraction rules used by the Content Extractor. Specifically, in our CER system we were interested in a wider range of concepts, including patient demographic information, temporal concepts, etc. These rules were simply removed from the i2b2 system. Referring back to the example illustrated in Figures 2 and 3, for i2b2 the extraction of BODILY-PROCESS concepts is suppressed.

Other rules were looking at wider contexts to pull information from. For example, we were extracting negative polarity for diseases and symptoms from sentences of the form "She denies any nausea or vomiting." Such rules were simplified for i2b2, where only the diseases and symptoms themselves had to be extracted.

Finally, a number of new extraction rules were created. This was, in part, a reflection of the fact that we had relatively few extraction rules in our system to begin with. Therefore, we expect some of the new rules will eventually be enhanced and added back to the underlying system.

**Post-processing.** For i2b2, the output of the NLP processing phase could be thought of as a list of LF fragments. The main goal of the post-processing phase is to map this output into the three i2b2 categories: problems, tests, and treatments. We therefore built mappings between extraction ontology types and the i2b2 categories; these mappings are fairly straightforward. For example, ontological types CONDITION and SYMPTOM reliably map to the "problem" i2b2 tag. Thus, in the example discussed in the previous section, "not adequate" is extracted as a CONDITION, and will be mapped to "problem."

In some cases the extraction's ontology type was too high up in the hierarchy, rendering it ambiguous. For example, the PROCEDURE type could include both diagnostic procedures (i.e., tests) and treatment procedures (i.e., treatments). For these cases we added more detailed mappings, based on the main concept's ontology type and lexicon entry.

Finally, a last step in processing involved the correction of systematic differences between constituent boundaries as extracted by our system and the specific requirements of the i2b2 task. This step only included less than a dozen rules to handle some of the frequent mismatches. For example, drug prescriptions typically include dosage, form, when and how to take the drug, etc.; in the LF, these will appear as modifiers, and will be extracted together with the name of the drug. For i2b2, only the name of the drug ought to be extracted. Similarly, prefixed negations in phrases such as "no wheezing" would be extracted by our system as modifiers; for i2b2 concept extraction, they ought to be left out. We decided that it was easier to make such adjustments in the post-processing phase instead of modifying the way extraction rules operated.

| Test | R value | P value | F value |
|---|---|---|---|
| problem | 48.15% | 57.66% | 52.48% |
| test | 29.33% | 42.92% | 34.84% |
| treatment | 56.24% | 43.64% | 49.15% |
| overall | 45.19% | 48.69% | 46.87% |

**Table 1.** Results for the exact span match.

| Test | R value | P value | F value |
|---|---|---|---|
| problem | 66.42% | 79.55% | 72.40% |
| test | 47.50% | 69.51% | 56.43% |
| treatment | 76.83% | 59.62% | 67.14% |
| overall | 69.36% | 69.10% | 69.23% |

**Table 2.** Results for the inexact span match.

Except for the boundary adjustments outlined above, there was no attempt to implement the rules described in the i2b2 Concept Annotation Guidelines about what constitutes a "complete" noun phrase or adjectival phrase. Rather, the span of each extracted phrase was simply the maximal contiguous span for the underlying LF fragment.

### Results and Discussion

Tables 1 and 2 summarize the performance results obtained by our system in the official i2b2 test for the exact span match and the inexact span match, respectively. We show values for recall (R), precision (P) and the F-measure (F).

Clearly, our system had significantly poorer performance on the exact match test than on the inexact match test, a consequence of the fact that we did not implement the specific guidelines for what ought to be extracted. The constituent structures present in the parser's output would allow us to more closely match the requirements, but linking the semantic LF representation back to the syntactic parse is not trivial, and we didn't have time for it. In some cases, we also believe that the longer spans that our system extracts are, in fact, more informative; for example, where we extract "simple atheroma in the aortic root", the i2b2 development data usually has only "simple atheroma" annotated as a problem – we feel that ignoring the modifier representing the site where the condition manifests leads to an incomplete description of what the problem really is.

Test extraction performance is notably lower than all others (especially recall). In part, this is explained by the fact that tests appear to be misclassified significantly more often than the other two categories. By far the largest number of tag confusions is represented by tests misclassified as treatments. It appears that this originates from having too high-level ontology mappings from UMLS, which don't distinguish between different types of substances. More precise ontology grafting should help.

Another source of errors that disproportionately affects test extractions is constituted by abbreviations (e.g., "Pt" may be either a test or a treatment, though it often simply means "patient"). The selectional restrictions in our present system are insufficient for disambiguating many abbreviations. We also found numerous abbreviations that are not mapped to the correct UMLS concept by MetaMap and we have no means of correcting this in later processing stages.

At the time of the official test run, we had a few updates to the post-processing phase that we hadn't had time to implement. When we finally managed to add them in, we noticed an overall improvement in precision of about 1.5% absolute on the exact match test and abut 2.5% absolute on the inexact match test. It is likely that a thorough error analysis will reveal a lot more opportunities there for improving precision (recall is much less affected by post-processing).

## Conclusion and Future Work

Given that our base system had been under development for just about 9 months when we performed the i2b2 test, we view the above performance results as very encouraging, and as a validation of our approach. Nevertheless, a system this young has much room for improvement. We already mentioned a number of areas where we see potential benefits.

In the short term, we are planning to conduct a more thorough error analysis, in order to find particularly glaring mistakes and promising avenues for development. We also plan to carry out more thorough experiments to evaluate the relative importance of various decisions we made with respect to UMLS integration.

Additionally, we plan to continue development on our clinical text information extraction system for our CER project. In particular, we plan to expand our extraction rules to include both assertions and relations for extracted concepts (something which unfortunately we were not able to complete in time for the i2b2/VA challenge). Finally, we are going to incorporate our work on temporal extractions for events[12] to allow us to create a timeline for clinical events within the extracted information.

## Acknowledgements

The project described was supported by Award Number RC2CA1488332 from the National Cancer Center and the H. Lee Moffitt Cancer Center and Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Center, the National Institutes of Health, or the H. Lee Moffitt Cancer Center.

## References

1. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008; 128-44.
2. Allen J, Byron D, Dzikovska M, Ferguson G, Galescu L, Stent A. An architecture for a generic dialogue shell. Natural Language Engineering. 2000;6:213-228.
3. Allen JF, Swift M, de Beaumont W. Deep semantic analysis of text. In: Proceedings of the 2008 Conference on Semantics in Text Processing (STEP '08). 2008; 343-354.
4. Blaylock N, Swain B, Allen J. Mining geospatial path data from natural language descriptions. In Proc. 1st ACM SIGSPATIAL GIS International Workshop on Querying and Mining Uncertain Spatio-Temporal Data. Seattle, WA. 2009.
5. Miller GA. WordNet: A lexical database for English. Communications of the ACM. 1995;38:39-41.
6. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucl Acids Res. 2004; 32:D267-D270.
7. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010; 17:229-236.
8. http://www.ihtsdo.org/snomed-ct/ (accessed Sept. 1, 2010).
9. http://ncit.nci.nih.gov/ (accessed Sept. 1, 2010).
10. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. Journal of Biomedical Informatics 2003;36:414-432.
11. Swift M, Blaylock N, Allen J, de Beaumont W, Galescu L, Jung H. Augmenting a deep natural language processing system with UMLS. Submitted.
12. Naushad UzZaman and James Allen. Extracting events and temporal expressions from text. Proc 4th IEEE International Conference on Semantic Computing (ICSC2010), Pittsburgh, USA, 2010.