Beyond NER: Towards Semantics in Clinical Text

Clare T. Grasso[1], Anupam Joshi[1], and Eliot Siegel[2]

[1] University of Maryland Baltimore County
[2] University of Maryland Medical School
Baltimore, Maryland, USA
cgrasso@umbc.edu

Abstract. While clinical text NLP systems have become very effective in recognizing named entities in clinical text and mapping them to standardized terminologies in the normalization process, there remains a gap in the ability of extractors to combine entities together into a complete semantic representation of medical concepts that contain multiple attributes each of which has its own set of allowed named entities or values. Furthermore, additional domain knowledge may be required to determine the semantics of particular tokens in the text that take on special meanings in relation to this concept. This research proposes an approach that provides ontological mappings of the surface forms of medical concepts that are of the UMLS semantic class *signs/symptoms*. The mappings are used to extract and encode the constituent set of named entities into interoperable semantic structures that can be linked to other structured and unstructured data for reuse in research and analysis.

Keywords: knowledge representation · semantic search · information extraction · unstructured data · clinical decision support · health informatics

1 Introduction

Medical natural language processing (NLP) systems have become very effective in recognizing and normalizing named entities in clinical text by mapping them to standardized biomedical terminologies such as those contained in the Unified Medical Language System (UMLS) [7]. For instance, the token *Metrorrhagia* can be mapped to the UMLS concept identifier C0025874 which can then be cross-referenced with biomedical literature to provide clinicians with supportive information to aid in determining treatment options. However, some medical concepts in clinical text, such as those that constitute the semantic class *signs and symptoms*, as defined by the UMLS Semantic Network [5], intrinsically consist of multiple components such as anatomical location, severity, onset, and duration as well as the date/time of the event. When assessing patient status, the clinician considers all these properties as a semantic whole.

Additionally, some attributes are idiosyncratic for a particular medical concept and require specific domain knowledge. For example, pain severity can be

noted in the chart using different scales such as the Wong-Baker Faces scale [9] (6 levels) or the Numeric Rating Scale [12] (11 levels). Severity may be also expressed with terms such as *little*, *annoying*, or *excruciating* which may or may not appear in any standard medical terminology. These expressions must be normalized to standardized terms such as *mild*, *moderate*, or *severe* if the data are to be practically reused in analysis.

This research proposes an ontology-based data access approach to retrieving semantic representations of composite medical concepts of the type *sign or symptom*. This lightweight, highly scalable approach allows for clinical data to be integrated with other structured data for use in clinical decision support or to be reused in biomedical research for cohort identification and analysis. The research described in this paper is being performed in conjunction with the University of Maryland Medical School and the Baltimore Veterans Administration (VA) Hospital Emergency and Radiology Departments. This research came as a result of an expressed need by the physicians at the VA to enhance care by integrating semantic search within a patient's chart into their electronic health system. These physicians were particularly interested in monitoring a patient's pain over time.

This research focused on extracting and encoding the medical concept of pain both as a proof of concept and because of the physicians' strong interest in it. Pain is an especially difficult target because of its prevalence in human experience. This results in a very wide variety of ways in which the concept of pain and its severity and quality may be expressed. In this ongoing work, this paper will detail the approach as it relates to pain and its severity.

2 Related Work

Several medical NLP systems have been developed which are effective in normalizing and annotating medical concepts in clinical and biomedical text such as cTAKES (clinical Text Analysis and Knowledge Extraction System) [13] and the National Library of Medicines MetaMap [4]. These systems are designed to process the entire document, one sentence at a time, and identify all named entity mentions of the type: *drugs, diseases and disorders, signs and symptoms, anatomical sites*, and *procedures*. Other entities such as numeric values, times, and dates are also labeled. In addition, part of speech tagging and shallow dependency parsing are also performed. However, post-processing must still be implemented on the annotated text in order to extract and encode the domain semantics for composite medical concepts that are made of up several attributes and their values. NLP algorithms are generally computationally complex, and may not be appropriate for use in semantic search in a near real-time environment [8].

Other systems such as Medical Language Extraction and Encoding System (MedLEE) [10] use a frame-based parser to extract and encode medical concepts into a semantic representation. Clinical notes are processed offline, stored,

and made available in real-time. However, the system is proprietary and the representations are not semantically interoperable with other systems.

More recently, the Strategic Health IT Advanced Research Projects Research Focus Area 4 (SHARP4) program aims to develop open-source tools for large-scale health record data sharing. One of the core initiatives is the development of the Clinical Element Model (CEM) [14] which is a detailed information model of composite clinical concepts and is associated with standard reference terminologies. This model is being converted to OWL-DL (Web Ontology Language Description Logic) as a representation formalism for interoperability, and to allow the use of inference and reasoning capabilities over extracted and encoded clinical data. The research described in this paper uses and builds on this model.

3 Methods

3.1 Corpus

The corpus consists of segments of ten deidentified patient charts that were downloaded from the VA's VistA [6]) electronic health system (EHS). Each patient has significant health problems involving pain such as cancer, kidney disease, and appendicitis. There are over forty different note types contained in the charts including triage, emergency department, surgery, radiology, laboratory, nursing assessments, and physical therapy. The EHR formats and outputs the chart as 80-character lines of ASCII text. All the notes for each patient are contained in a single file. There are a total of 93,375 lines of which the first 63,785 lines are being used in development; and the remaining 29,590 as the test set.

Table 1. Corpus Statistics

Patient Chart	Number Lines	Number Tokens	Unique Tokens	Number Notes	Note Types
Patient - Appendicitis	7,671	30,591	3,046	55	27
Patient - Syncope	3,095	13,927	2,116	15	12
Patient - Perirenal Abscess	28,424	118,332	4,893	229	48
Patient - Hypercalcemia	24,595	103,890	4,486	191	40
Subtotal	63,785	266,740		490	
Patient - Colon cancer	7,504	34,624	3,019	45	24
Patient - Pain	8,500	31,579	3,383	82	34
Patient - Anemia	5,762	24,858	2,593	32	17
Patient - Lung Cancer	7,824	34,255	3,071	52	25
Subtotal	29,590	125,316		211	
Total	93,375	392,056		701	

3.2 Approach

The general approach pursued in this research is that of ontology-based data access over unstructured data. Two ontologies and one analysis engine are used for each medical concept and for each attribute that is a constituent part of that concept. The first ontology assists in mapping the surface forms of the medical concepts in the text to their normalized representation. The second ontology defines semantic representation of the medical concept. These ontologies are implemented separately so that inference and reasoning over the knowledge base of facts is not encumbered by knowledge that is only needed for extraction. The semantic representation will use the CEM/OWL representation of the *sign and symptom* class as described above. The analysis engine uses the knowledge in the mapping ontologies to extract and encode them into their semantic representations. Separating the mapping and semantic representation from the analysis engine also allows alternative analysis engines to be used. These triple-sets form composable building blocks. For instance, onset and duration can be reused as attributes of many other medical concepts that are of the type *sign and symptom*.

The upper level mapping ontology contains all the knowledge needed to extract the main medical concept of interest. It also imports the subontologies for its constituent attributes such as body location and severity. The sequence of extractions is performed based on the subontologies. When order is important, (e.g., severity depends on the body location extractor), the order is specified in the top level mapping. When order is not important, attribute extraction may be parallelized to increase throughput.

For medical concepts and attributes in which a particular vocabulary has been assembled, such as for pain and severity, these terms are stored in the mapping ontology. Each term is a subclass of a **Term** class in which the base form is stored along with a regular expression that matches its surface lexical forms. Other domain knowledge needed for the extraction of the term may also be asserted in the class. Terms may also be subclasses of other terms. For example, in the severity mapping, the term *Minor* may be a subclass of the term *Mild* for normalization purposes.

For mappings that require external sources, the mapping ontology contains references to resources such as terminologies, lexicons, and other ontologies. For example, the body location mapping ontology contains references to Foundational Model of Anatomy (FMA) [1] and SNOMED-CT [3].

Once the extraction is complete, the pertinent information is encoded as an instance in the ontology that contains the semantic representation of the medical concept. This includes the reference to the document level information, normalized concepts identifiers, the original text found, and the line number and text span in which it was found in the note.

3.3 Pain Severity Analysis Engine

For many medical concepts of the class *signs and symptoms*, severity is expressed generally as *mild, moderate*, and *severe*. However, it may also be expressed with

a large number of synonyms that imply some level of pain such as *minimal* or *torturing*. Many of these terms are found in one or more medical terminology systems. Others are only found in general English thesauri. Severity terms were mined from these sources as well as the development data, pain questionnaires, online health sites, and research papers. The terms were then added to the mapping ontologies. With the help of physicians, these terms were related as subclasses to the normalized terms of *mild, mild to moderate, moderate, moderate to severe*, and *severe* which have been adopted by SHARP. As pain may also be expressed as a value on a numeric scale, these values were also normalized.

Listed below are various ways that pain severity was expressed in this corpus.

1. "SCALE:7", "Pain:4"
2. "Patient reports pain during shift: No"
3. "pain was tolerable", "minor abdominal soreness", "denies any pain"
4. "s: pain: 8.5/10 b hands"

Algorithm. The pain severity analysis engine modifies and extends the graph-based ConText [11] algorithm which relies on pattern matching trigger terms and scoping. Target rules provide the mapping for the main concepts which, in this case, is *pain* and all its alternative expressions such as *discomfort*. Modifier rules provide the mapping for contextual features of the targets, such as whether it is negated, historical, or hypothetical. For example, in the phrase "no pain", *pain* is the target that is modified by the negator, *no*. Each rule is given a unique name, a categorical type and a pattern used in matching. Modifier rules also are given a direction that indicate whether the modifier looks forward, backward, or both. For example: `denies, DEFINITE_NEGATED_EXISTENCE, \bdenies\b, FORWARD`. Target and modifier terms found in the text are added as a nodes in the graph. Each node contains the scope (text span) in which it is active.

The algorithm processes one textual unit, such as a sentence, at a time. The scope of target nodes is the entire textual unit. The scope of modifier nodes begins at the start span of the modifier term in the text and ends at either the beginning or end of the textual unit depending on the direction of its rule. Target nodes are connected to modifier nodes whose scope falls within their own.

Because this algorithm does not rely on syntactic parsing, it was very effective on the type of non-canonical text entered by the clinician at the point of care that is prevalent in this corpus such as: "64 yo AAM with h/o stage IV RLL adenocarcinoma T2N0M1 s/p treatment with carboplatin/taxol"

Algorithmic Modifications to Scoping. Because sentence segmentation can be very difficult with this type of text, scoping was changed to treat the newline characters as the basic unit of scope. Scoping modifications were also made for other types of punctuation such as semicolons and periods.

Modifications to the Rule Database. Fifteen target rules were added to the database to identify the different expressions of pain and its severity. Seventy modifier rules were also added . Twelve were of the type DEFINITE

_NEGATED_EXISTENCE to accommodate the idiosyncrasies of this corpus such as "Pain during shift: No". Forty-eight were of the new type PAIN_SEVERITY to recognize both numeric and lexical terms indicating severity ("9/10","not too bad"). Ten were of the type PROBABLE_EXISTENCE ("c/o","states").

4 Evaluation

4.1 Gold Standard

The reference standard was created by three physicians. The first physician provided oversight concerning what aspects of the data would be extracted which was used to develop the annotation guidelines. Two other physicians independently annotated each line of text in the test set for two variables, i.e, **Existence** and **Literals**. **Existence** was labeled with *Affirmed, Probable, Negated, or N/A*. Literals were labeled with the specific severity quantifier found in the text - numeric or lexical. If an expression crossed a line boundary, the annotation was made on the first of the two lines. The first physician annotated text in cases where the two annotators did not agree. This annotation was used to break the tie for the final gold standard.

A combination of the **Existence** and **Literals** annotations were used to create the final **Severity** variable against which the system was tested. If a line contained a **Literal** value, that value was used for **Severity**. Otherwise, if **Existence** was labeled *Affirmed* or *Probable*, **Severity** was labeled *Unknown*. If **Existence** was labeled *Negated*, **Severity** was labeled *0*.

For **Existence**, $A_o = 0.86(290/337), A_e = 0.468$, and Fleiss's $\kappa = 0.738$. For **Literals**, $A_o = 0.79(79/100), A_e = 0.184$, and Fleiss's $\kappa = 0.743$.

5 Results

Results were evaluated using the vertical metrics for measuring system performance on individual fields as outlined in [15] and consisted of phrase-level precision, recall and F-measure. (*line number, severity value*) pairs were compared against the gold standard. Only exact matches were considered correct. Specifically:

1. *true positive* – extracted value exactly matched the annotation.
2. *false positive* - extracted value did not match the annotation.
3. *false negative* – no value was extracted for text containing an annotation
4. *true negative* - no value was extracted for text not containing an annotation.

$$\textbf{precision} = \frac{\#Correctly\ returned\ values\ by\ system}{\#Values\ returned\ by\ the\ system} \tag{1}$$

$$\textbf{recall} = \frac{\#Correctly\ returned\ values\ by\ system}{\#Values\ in\ gold\ standard} \tag{2}$$

This resulted in a positive precision of 90.31% (233/258), a positive recall of 88.26% (264/289), and a positive F1-score of 89.27%.

For comparison, the original ConText algorithm reported very good results for negated findings in a corpus of six different note types (P:97%, R:97%) [11]. However, the corpus described in this paper included many more note types (over 40), and the non-canonical text that occurs extensively in it creates a much larger lexical and syntactic space. When the unmodified ConText algorithm downloaded from [2] was run on this corpus, the results were (P:77%, R:90%) for negated findings. After the modifications described in this paper were applied, results for negated findings on this corpus was significantly higher (P:92%, R:92%).

6 Discussion

In these results, recall suffers due to not recognizing lexical severity values ("mildly"), misspellings ("dnied"), and missing whitespace ("Chest Pain–denies"). However, almost one-third of false negatives (11/31) were the result of pain mentions that cross line boundaries ("no c/o [new line] pain"). For text that was characterized by non-canonical use of grammar, attempts to ignore line breaks and segment into sentences was not successful and resulted in lower accuracy. However, for the more canonical text found in specialist reports (radiology, surgery), this would be beneficial. Plans for future work include the use of machine learning techniques to train a classifier to recognize these different types of text.

Another source of error in recall was when a pain score was associated with body locations instead of to a target term, for example, "7.5 both hands." When work on the body locations extractor is complete and can identify these terms in the text, the severity extractor should be able to handle these cases as well.

An additional difficulty is that there may be more than one pain mention within the same line/sentence, for example, "new onset headaches, no c/o chest pain". Likewise, several body locations may share the same negation modifier, such as, "pt denies any abdominal pain, headaches," The algorithm will need to be modified to handle these as separate pain mentions.

An interesting issue came up regarding which terms imply pain. In the narrative note, terms such as pressure, discomfort, and sensation are treated differently from pain, for example, "feels pressure, no pain." In this case the algorithm classified it as a negated pain mention, but the physician annotator classified it as *Probable* pain with *Unknown* severity. In addition, the term *tender* was not regarded by the physician as an indicator of pain if it appeared in a physical exam section of a note and was the result of a palpation, for example, "Abd: tender, nondistended". However, the physician did classify it as pain if it appeared in a triage note, for example, "redness w/ warmth and tenderness." The semantics will need to be extended to recognize the broader context such as the type of note and sections within a note. This type of domain knowledge can be asserted in the mapping ontology used to encode the entity into its full semantic representation.

7 Conclusion

This ongoing research investigates an approach that allows clinicians to perform semantic search through a patient record to explore diagnostic hypotheses. It extracts and encodes the medical concept of pain into a semantic whole that can be stored in a knowledge base and used with reasoning and inference engines for clinical decision support and for reuse in analysis. In addition, it operates over all note types, including the very difficult non-canonical text that typifies narrative notes that are entered by clinicians at the point of care. Future work will undertake to extend it to other concepts of the type *signs and symptoms*.

References

[1] Fma - foundation model of anatomy. `http://sig.biostr.washington.edu/projects/fm/AboutFM.html`, accessed June 2014
[2] pycontextnlp. `https://github.com/chapmanbe/pyConTextNLP`, accessed June 2015
[3] Systematized nomenclature of medicine–clinical terms. `http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html`, accessed June 2015
[4] Umls metamap. `http://metamap.nlm.nih.gov/`, accessed June 2014
[5] Umls semantic network. `http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html`, accessed June 2014
[6] Worldvista. `http://http://www.worldvista.org/`, accessed June 2015
[7] Bodenreider, O.: The unified medical language system (umls): Integrating biomedical terminology. Nucleic Acids Research pp. D267–D270 (2006)
[8] Chard, K., Russell, M., Lussier, Y., Mendona, E., Silverstein, J.: A cloud-based approach to medical nlp. AMIA Annu Symp Procs pp. 207–216 (2011)
[9] Cote, C., Lerman, J., Todres, I.: A practice of anesthesia for infants and children. Elsevier Health Sciences p. 940 (2009)
[10] Friedman, C., Johnson, S., Forman, B., Starren, J.: Architectural requirements for a multipurpose natural language processor in the clinical environment. In: Proc Annu Symp Comput Appl Med Care. pp. 347–351 (1995)
[11] Harkema, H., Dowling, J., Thornblade, T., Chapman, W.: Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform 42(5), 839–851 (2009)
[12] McCaffery, M., Beebe, A.: Pain: Clinical Manual for Nursing Practice. VV Mosby Company, Baltimore, MD (1993)
[13] Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. J Am Med Inform Assoc 17, 507–513 (2010)
[14] Tao, C., Jiang, G., Oniki, T., Freimuth, R., Q. Zhu, D.S., Pathak, J., Huff, S., Chute, C.: A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. J Am Med Inform Assoc 20(3), 347–351 (2013)
[15] Uzuner, O., Solti, I., Cadag, E.: Extracting medication information from clinical text. J Am Med Inform Assoc 17(5), 514–518 (2010)