HTP-NLP: A New NLP System for High Throughput Phenotyping

Daniel R. SCHLEGEL [a,1], Chris CROWNER [b], Frank LEHOULLIER [b] and
Peter L. ELKIN [b]
[a] *Department of Computer Science, SUNY Oswego, Oswego, NY, USA*
[b] *Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, USA*

Abstract. Secondary use of clinical data for research requires a method to quickly process the data so that researchers can quickly extract cohorts. We present two advances in the High Throughput Phenotyping NLP system which support the aim of truly high throughput processing of clinical data, inspired by a characterization of the linguistic properties of such data. Semantic indexing to store and generalize partially-processed results and the use of compositional expressions for ungrammatical text are discussed, along with a set of initial timing results for the system.

Keywords. high throughput phenotyping, clinical NLP, compositional expressions

1. Introduction

Secondary use of clinical data for research requires a method to quickly process large amounts of data for cohort extraction. This is the foundation of phenotyping as the word is used in informatics. More concretely, phenotyping has been defined as "the algorithmic recognition of any cohort within an EHR for a defined purpose, including case-control cohorts for genome-wide association studies, clinical trials, quality metrics, and clinical decision support" [5]. The phenotyping task must be repeated for new versions of semantic resources. A system which performs phenotyping of large volumes of data must be able to do so quickly, and so we often refer to "high throughput" phenotyping.

The general technique for high-throughput phenotyping is to pre-process records to extract information salient to cohort selection. The task is one of information extraction – at least a subset of the contents of records are "understood" and mapped back to a well-defined semantics. Spans of text with identical clinical meaning must be indexed together. Queries may be structured and access this semantic index directly, or unstructured and processed through the same information extraction process, then matched with the contents of the index. These queries may be manually created using Boolean logic on features in the index, or generated automatically by machine learning algorithms, as in [10]. Algorithms for phenotyping are being collected by resources such as PheKB [4].

The term "High Throughput Phenotyping" has been used in this context since at least 2013, when the SHARPn consortium used a template-based method for extracting features related to medications, procedures, symptoms, labs, disorders, and anatomical

[1] Corresponding Author: Daniel R. Schlegel E-mail: daniel.schlegel@oswego.edu.

sites [5]. The system discussed here attempts to be more general, extracting many of the features discussed by the SHARPn consortium, but using much more general templates. Proposed solutions to the phenotyping problem have been around since much before 2013 (e.g., [1]), though the problem itself was not as well defined.

We present two advances in the new High Throughput Phenotyping Natural Language Processing (HTP-NLP) system.[2] Both advances are meant to support high throughput, inspired by a characterization of the linguistic properties of clinical data (Section 2). The HTP-NLP system makes use of a *semantic index* to store partially processed text which may be generalized and re-used (Section 3), and a new implementation of *compositional expressions*[3] for high-speed noun-phrase relation extraction, even from ungrammatical text (Section 4). Initial results are presented in Section 5.

2. Characterizing Unstructured EHR Data

Patient records have properties that are not common to all natural language text, and which contrast with the assumptions of most NLP systems. Records are patient-centric; highly conventionalized; often contain pseudo-structured text which is difficult to analyze; and have text in which the semantic scope is likely to be very local.

Medical records are centered around the patient. In narrative text it is important to track who has a given property, such as in “Susan and Joe met at the hospital. Susan has diabetes and Joe has sleep apnea.” where the correct relationship is between “sleep apnea” and “Joe” as opposed to “sleep apnea” and “Susan”. In patient records, outside of sections concerning family history, when “sleep apnea” is seen, the assumption is that the condition refers to the patient.

The natural language that occurs in patient records is far less likely to be novel than what is in many other kinds of natural language data. Phrases such as “arrhythmia of the heart” are likely to be frequently repeated, whereas a single phrase in a newspaper may appear extremely rarely. Indeed, patient records use a kind of formalized language. Health professionals establish a convention for expressing things repeatedly, and not always in a grammatical way. A pseudo-sentence such as “EKG: normal.” might appear often in fields that are designated for natural-language text, simply because clinicians have established this as the best way to write that a patient’s EKG is normal. Another source of conventionalized expressions is the data entry tools used to produce patient health records. A given tool might include a drop-down box with a specific phrase like “cancer of the lung”, leading to its repeated appearance.

In many natural language texts the semantic scope of expressions extends over several sentences. The most obvious example of this is pronominal co-reference, where the meaning of a pronoun is taken from the meaning of the previous sentences. In patient medical records cross-sentential semantic relations occur rarely; expressions such as “The patient was diagnosed with skin cancer in 2012. After two years he recovered.”, where both sentences are necessary in order to interpret the word “recovered”, are rare. The most common type of wide scope encountered in patient medical records is negation, *e.g.,* “The patient does not have skin cancer.” The scope of negation is almost always at the level of the sentence rather than having a wider scope.

[2] This software is made available to other CTSA sites, contact elkinp@buffalo.edu for details.

[3] First discussed in preliminary form in [2].

3. Semantic Indexing

The overall design of the HTP-NLP linguistic processing mechanism has been influenced by the linguistic features discussed in Section 2 and the task at hand — quickly building study cohorts. The system makes use of a series of cascading indexes, capitalizing on the highly repetitive and formulaic nature of the natural language data, while preserving the flexibility to adapt to differing study needs.

NLP applications usually processes each document through a pipeline, often without storing intermediate information. These pipelines include steps such as tokenization, sentence breaking, syntactic analysis, and named entity recognition. For a given document, these tasks are performed from left-to-right, top-to-bottom, regardless of how many times the same sentence has been processed. A well designed NLP system which follows this strategy may cache some named entities, for example, to prevent re-processing, but this cache is often limited to the document, and does not persist across an entire corpus.

The HTP-NLP system stores each level of linguistic analysis in a key-value store. Each key represents the incoming level of linguistic analysis and the value represents the outgoing level. One incoming level of linguistic analysis might be the sentence and the outgoing level of linguistic analysis might be noun-phrase chunks. Because of this, each unique sentence and each unique noun-chunk needs to be analyzed only a single time across an entire corpus. Given the formulaic and redundant nature of EHR data, a great deal of redundant processing can be avoided. For example, the phrase “atrial fibrillation” is only coded a single time, even if it appears in the data several hundred thousand times.

The present system uses the following indexes: a discourse level index from individual patient records to sentence/sentence fragments; a syntax level index that from sentences to phrases, words, negation markers, evidentiality markers etc.; a semantic index from phrases or words with polarity indicators to semantically related alternative phrases and synonyms; and an analytic index from linguistic semantic content to ontological specifications such as SNOMED CT, ICD 10, or study-specified ontology. Cohort selection simply requires back-tracking through these indexes from code to document.

This design allows for efficient re-coding in the case of different ontology requirements or updates to existing ontologies. Re-analysis needs only be performed at the level of a single index and only a single time for each indexed key, rather than over each document. Thus, if a study needed to augment an existing index with terms from the Gene Ontology, only the Gene Ontology terms in the index must be re-processed.

4. Resilience to Non-Grammatical Text[4]

Many NLP systems, such as cTAKES [6], rely on some sort of linguistic parse of text for relation extraction. We believe that for at least some cases in the information extraction task, noun-phrase processing can be done using the order of terms as they appear in the text. Relying only on term order allows for processing of ungrammatical text such as “EKG: Normal” and the recognition that it has the same meaning as other expressions such as “normal EKG”. Many systems are unable to handle non-grammatical text without specific templates. Our system performs this processing based on a special kind of post-coordination known as compositional expressions (CEs).

[4] Portions of this section adapted from [7]

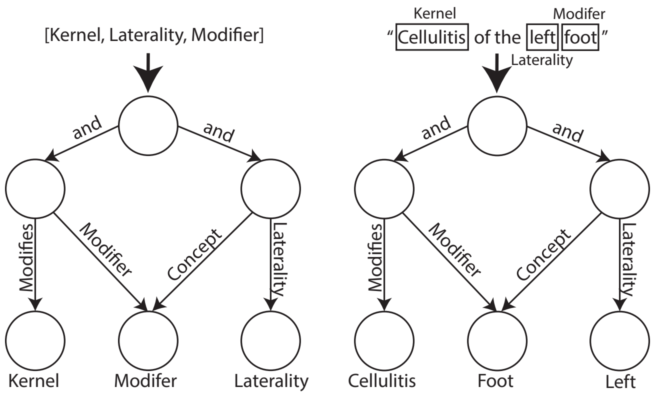

Figure 1. Left: a template graph for text which matches the pattern of a Modifier followed by Laterality followed by Kernel. Right: Instantiation of the template graph by the text “Cellulitis of the left foot”.

Post-coordinated concepts consist of multiple individual concepts related to each other using both existing portions of the terminology or ontology graph, as well as new instances of pre-defined relations. Post-coordination is a laborious manual task, in many cases requiring deep understanding of the text and the terminology. For many tasks, such as information extraction, where the goal is to recognize that two spans of text mean the same thing clinically, the deep understanding of post-coordination is often unnecessary.

Compositional expressions extend the idea of using portions of the existing ontology and terminology graph, adding logical and linguistic relations. The advantage of using CEs created from ontologies and terminologies is that multiple surface forms for the same concept are mapped to a single logical form (and hence, a graph structure [8]). For example, the following three forms, all representing hypertension which is uncontrolled, map to a logical form in which the SNOMED CT concept for “hypertension” is the first argument in a binary `hasModifier` relation with the SNOMED CT concept for “uncontrolled”: *Uncontrolled hypertension*; *HT, uncontrolled*; *Uncontrolled hypertensive disorder*. In addition, CEs add semantic data which is otherwise missing when text is coded using pre-coordinated terms alone. For example, when using SNOMED CT, 41% of clinical problems require CEs in order to be represented properly [3].

Rules for building CEs are based on the order of terms as they appear in the text. Some rules are based on upper level terms, *e.g.,* a `Procedure` adjacent to a `Body Structure` indicates the site of a procedure. Others use the *kind* of a term – either kernel, modifier, qualifier, or laterality. The kernel is the clinical concept under discussion. Modifiers change the meaning of of the kernel, such as “uncontrolled” in “uncontrolled hypertension.” Qualifiers specify status, such as “history.” Laterality has to do with sidedness, such as “right.” An example of this type of ordering would be a kernel, followed by laterality, followed by a modifier in “cellulitis of the left foot”.

Compositional expressions have been implemented to support high throughput. Term orderings are compiled into a discrimination tree, with the leaves mapped to small graph templates. (See left side of Figure 1). A template is instantiated when a sequence of terms matches a path in the discrimination tree, resulting in a graph such as in the right side of Figure 1. These are stored in a graph database, which has been shown to have fast retrieval times [7] for small graphs and subgraphs.

5. Initial Results

To compare processing speed with cTAKES, we processed 537,157 encounter notes for 97,964 patients from the UBMD Allscripts database on a single CPU. This took 48.3 minutes.[5] Of this: 29.3 minutes were spent on linguistic analysis and semantic indexing; 19 minutes were spent coding with SNOMED CT and synonym sets (*e.g.,* [9]), extracting 796 million codes. This is more than an order of magnitude faster than single-threaded cTAKES, which processed 14,021 notes per hour. This improvement is largely attributable to semantic indexing reducing the amount of data which was processed.

6. Conclusion

The HTP-NLP system represents a marked improvement over traditional pipeline-based models for information extraction such as cTAKES. The improvement comes from a system design based on a characterization of the unstructured EHR data which is being processed. The use of semantic indexing results in significant processing-time improvements, and CEs perform high-speed relation extraction, even for ungrammatical text.

Acknowledgements

This work has been supported by the Clinical and Translational Science Award (Number 1UI1TR001412-01), and the VA Big Data-Science Training Enhancement Program.

References

[1] Brown, S.H., et al.: eQuality: electronic quality assessment from narrative clinical reports. In: Mayo Clinic Proceedings. vol. 81, pp. 1472–1481. Elsevier (2006)

[2] Elkin, P.L., Brown, S.H., Chute, C.G.: Guideline for health informatics: Controlled health vocabularies-vocabulary structure and high-level indicators. Stud Health Technol Inform (1), 191–195 (2001)

[3] Elkin, P.L., et al.: Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. In: Mayo Clinic Proceedings. vol. 81, pp. 741–748 (2006)

[4] Kirby, J.C., et al.: PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc p. ocv202 (2016)

[5] Pathak, J., et al.: Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc 20(e2), e341–e348 (2013)

[6] Savova, G.K., et al.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17(5), 507–513 (2010)

[7] Schlegel, D.R., Bona, J.P., Elkin, P.L.: Comparing small graph retrieval performance for ontology concepts in medical texts. In: Wang, F., et al. (eds.) Biomedical Data Management and Graph Online Querying: VLDB 2015 Workshops, Big-O (Q) and DMAH, LNCS, vol. 9579. Springer (2016)

[8] Schlegel, D.R., Shapiro, S.C.: Visually interacting with a knowledge base using frames, logic, and propositional graphs. In: Croitoru, M., et al. (eds.) Graph Structures for Knowledge Representation and Reasoning, LNAI 7205, pp. 188–207. Springer-Verlag, Berlin (2012)

[9] Schlegel, D., Crowner, C., Elkin, P.: Automatically expanding the synonym set of snomed ct using wikipedia. Stud Health Technol Inform 216, 619 – 623 (2015)

[10] Yu, S., et al.: Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. J Am Med Inform Assoc 22(5), 993–1000 (2015)

[5] For a fair comparison with cTAKES, we ensured the same features were found in both. As cTAKES has no CE processor, we excluded it from our test. Processing with CEs roughly doubles coding time.