University at Buffalo *The State University of New York* | REACHING OTHERS

OSWEGO
STATE UNIVERSITY OF NEW YORK

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

Daniel R. Schlegel

Department of Computer Science, SUNY Oswego

Chris Crowner, Frank LeHoullier, and Peter L. Elkin

Department of Biomedical Informatics, University at Buffalo

Presented by Mark Jensen

Department of Biomedical Informatics, University at Buffalo

High Throughput Phenotyping

- Buzz-phrase, floating around the last few years
- Phenotyping: “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”[1]
- High-Throughput: Phenotyping on many records quickly.

[1] 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)

Our Contribution

- Started from examination of the nature of text in medical records
- Two advances:
  - Semantic Indexing
  - Resilience to Non-Grammatical Text

Unstructured EHR Data

- Patient-centric
  - Outside of certain sections, we can assume all text is related to the patient
- Highly conventialized
  - The same phrases are likely to repeat hundreds or thousands of times across a corpus
  - “arrhythmia of the heart”
  - “EKG: normal” (pseudo-sentences)
- Local semantic scope
  - Rarely is understanding >1 sentence required to derive meaning.

Semantic Indexing

- Since there is so much repeated text, repeated processing is wasteful.
- Store each level of analysis in key-value store

Level	Key-Value Store
Discourse	Medical Record -> Sentence / Fragment
Syntax	Sentence -> Word / Phrase (including Polarity / Evidentiality /… markers)
Semantic	Word / Phrase -> Semantically related phrases / synonyms
Analytic	Linguistic semantic content -> Ontological term

- Cohort selection: backtrack from terms -> documents

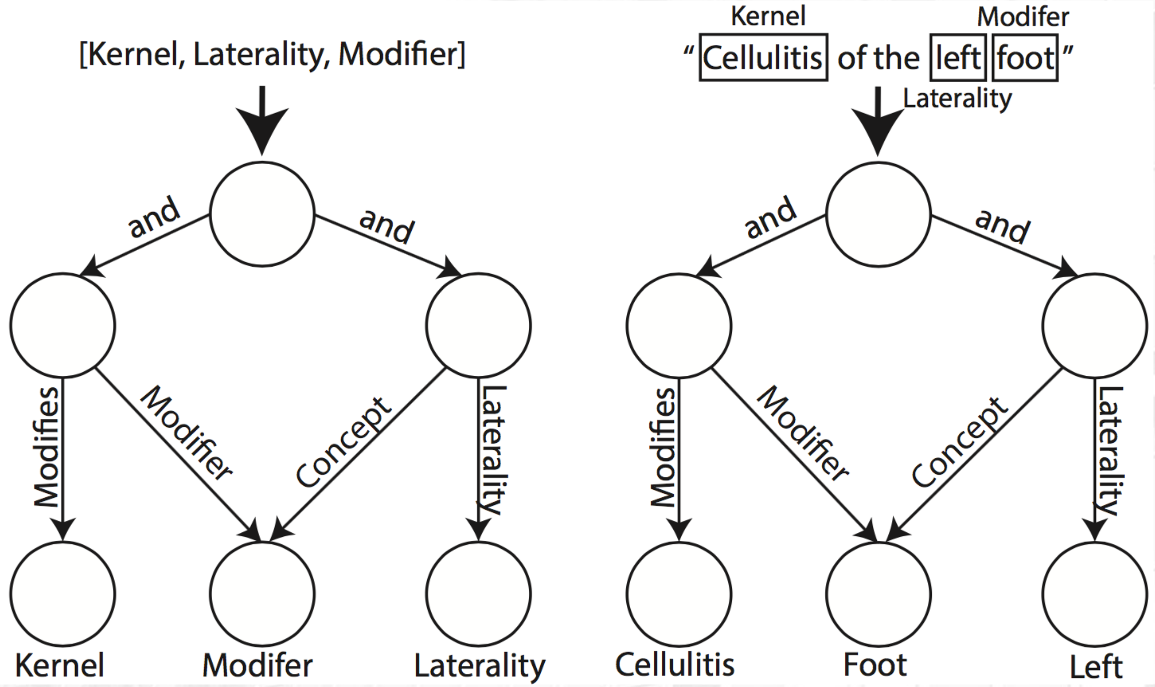
Resilience to Non-Grammatical Text

- Noun-phrase processing can sometimes be done by examining word order alone.
  - Allows for processing ungrammatical text
  - Different surface forms map to the same logical form
  - e.g., *EKG: Normal* vs. *Normal EKG*
- A kind of post-coordination: Compositional Expressions

Compositional Expressions

- Created from the ontological representation of parts of linguistic phrases or sentences.
- Ontology terms (hierarchies) are tagged as being a:
  - kernel – the clinical concept under discussion
  - modifier – changing the meaning of the kernel
  - qualifier – specifying status (e.g., ‘history’)
  - representing laterality – e.g., left or right

Compositional Expression Example



Template graphs for orders of components of CEs are built and matched against

Comparison Testing: Methodology

- Goal: Compare cTAKES vs. HTP-NLP
  - Just speed, no accuracy checks
- Dataset: UBMD Allscripts Database
  - 537,157 encounter notes for 97,964 patients
- Configuration: Single CPU, only components in both cTakes and HTP-NLP tests (no CE extraction)

Comparison Testing: Results

System	Time (minutes)
HTP-NLP	48.3 - 29.3 - linguistic analysis / semantic indexing - 19 - SNOMEC CT / synonym coding
cTAKES	2,299

HTP-NLP was **47.6** times faster than cTAKES in this experiment.

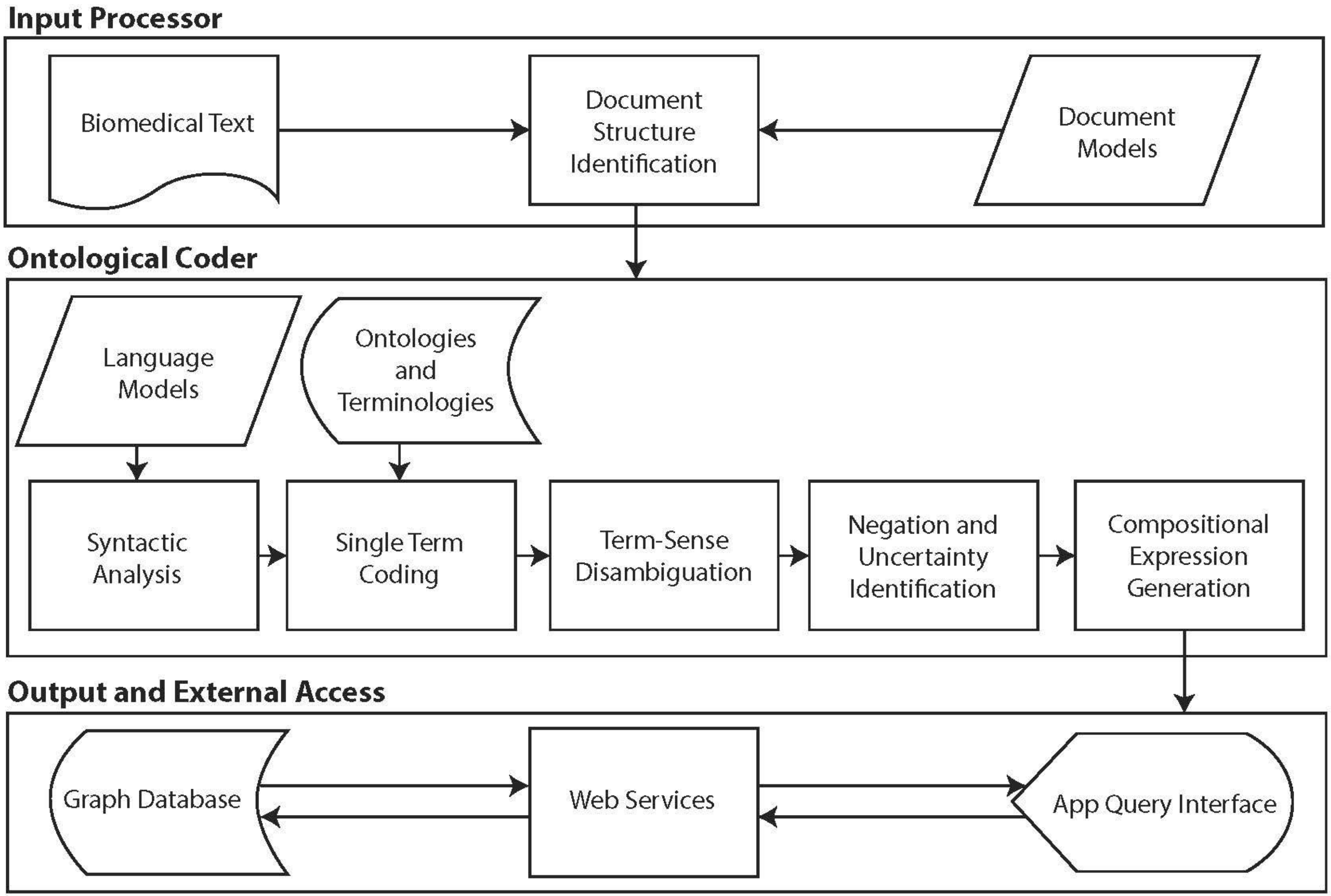
Adding CEs, which cTAKES does not have, roughly doubles processing time.

Conclusion

- The HTP-NLP system improves both portions of the High Throughput Phenotyping ideal
  - Semantic Indexing vastly increases throughput
  - Compositional expressions allow for dealing with non-grammatical text
- We achieved this by designing the system based on an understanding of the input data.

Thank you for your attention!

Contact: daniel.schlegel@oswego.edu


Input Processor
Biomedical Text
Document Structure Identification
Document Models
Ontological Coder
Language Models
Ontologies and Terminologies
Syntactic Analysis
Single Term Coding
Term-Sense Disambiguation
Negation and Uncertainty Identification
Compositional Expression Generation
Output and External Access
Graph Database
Web Services
App Query Interface