Natural Language Processing at Colorado

Martha Palmer, Jim Martin, Wayne Ward

SHARP Conference Call
University of Colorado
June 10, 2010
Outline

- Relationship between cTakes and CU-NLP
- Treebanking/PropBanking
- UMLS Relations
cTAKES: NLP Components
(Guergana’s 5/25 presentation)

- Core components
  - Sentence boundary detection (OpenNLP)
  - Tokenization (rule-based)
  - Morphologic normalization (NLM’s “norm”)
  - POS tagging (OpenNLP)
  - Shallow parsing (OpenNLP)
  - Named Entity Recognition
    - Diseases/disorders, signs/symptoms, procedures, anatomical sites, medications
    - Dictionary mapping (lookup algorithm)
    - Machine learning (MAWUI)
  - Negation and status identification (NegEx)
Colorado contribution to cTakes

- Semantic processing of the clinical text (in collaboration with Palmer, Martin and Ward)
  - Treebanking (deep parses)
  - Predicate-argument structure and semantic labeling (PropBanking)
  - UMLS relations (except temporal relations)
cTAKES: Components
(Guergana’s 5/25 presentation)

- Core components
  - Sentence boundary detection (OpenNLP)
  - Tokenization (rule-based)
  - Morphologic normalization (NLM’s “norm”)
  - POS tagging (OpenNLP)
  - Shallow parsing (OpenNLP)
  - Named Entity Recognition
    - Diseases/disorders, signs/symptoms, procedures, anatomical sites, medications
    - Dictionary mapping (lookup algorithm)
    - Machine learning (MAWUI)
  - Negation and status identification (NegEx)
  - Dictionary mapping (lookup algorithm)
  - UMLS Relations
  - Treebanking
  - PropBanking
Addressing challenges w/ current system

- Main sources of errors (from 5/25 presentation)
  - Abbreviations and WSD, e.g. “Dr.” with a mapping to “diabetic retinopathy”, “is” -> “immune suppression
  - Colorado has been doing the Sense tagging annotation and WSD system development for DARPA –GALE

  - Lexical variations and complex levels of synonymy
    “bladder showed very mild trabeculation” -> “trabeculated bladder”

These are exactly the types of variations that are normalized by propbanking
So what are treebanking and propbanking?

And how will they help with extracting event relations?
PropBanking enables Semantic Role Analysis

- Assigning semantic labels to sentence elements
- Elements are arguments of some predicate or participants in some event
  - **Who** did **What** to **Whom**, **How**, **When**, **Where**, **Why**

  - [ARGM-TMP In 1901] [ARG1 President William McKinley]
  - [TARGET was shot] [ARG0 by anarchist Leon Czolgosz]
  - [ARGM-LOC at the Pan-American Exposition]
Why?

- Richer annotation of documents
  - Labeling atomic events in texts along with event participants, locations, and other arguments
    - Facilitates IR/IE/Question answering
  - Labeling key arguments of specific kinds of events allows special purpose processing
    - Identifying temporal information facilitates causal and temporal reasoning with timelines

So what are Treebanks and PropBanks?
Analysts have been expecting a GM-Jaguar pact that would give the U.S. car maker an eventual 30% stake in the British company.
The same sentence, PropBanked

Analysts have been expecting a GM-Jaguar pact that would give *T*-1 the US car maker an eventual 30% stake in the British company.

expect(Analysts, GM-J pact) give(GM-J pact, US car maker, 30% stake)
Frames File Example: expect

Roles:
- Arg0: expecter (Agent)
- Arg1: thing expected (Theme)

Example: Transitive, active:

*Portfolio managers* expect *further declines in interest rates.*

Arg0: *Portfolio managers*
REL: *expect*
Arg1: *further declines in interest rates*
Frames File example: \textit{give}

Roles:

- Arg0: giver (Agent)
- Arg1: thing given (Theme)
- Arg2: entity given to (Recipient)

Example: double object

\textit{The executives gave the chefs a standing ovation.}

- Arg0: \textit{The executives}
- REL: \textit{gave}
- Arg2: \textit{the chefs}
- Arg1: \textit{a standing ovation}
Current Status of English PropBank

- Close to 2M words of annotated data
  - Newswire, Broadcast News, Talk Shows, WebText
- 5378 verb lemmas have frame files
  - 7212 verb senses
  - Links to WordNet, VerbNet, FrameNet
  - Over 99% coverage of verbs in new documents
- Approach has been ported to Korean, Chinese, Arabic, Hindi, French
Supervised ML Approaches to Language Processing

- Robust Machine Learning systems can *learn to produce* structured outputs...
- We now have tons of annotated data at our disposal
  - Trees, semantic roles, predicate-argument structures, word senses, entities, coreference, causal and temporal relations, ...
Training Set for Semantic Roles

- **Propbank I** (Palmer et al, 2005)
  - A million word corpus of Wall Street Journal Text
    - Syntactic parse trees for each sentence
    - Each argument to each non-copula verb hand-labeled
      - 132,000 labeled arguments in the training set
      - 7,000 arguments in a test set (Section 23)
    - Arguments are labeled with one of 23 semantic roles (Core Arg0-5, + Oblique ArgMs)
      - Arg0 roughly means prototypical agent
      - Arg1 roughly means prototypical patient
For each predicate in each training sentence, extract a set of features that characterize each semantic argument with respect to its predicate.

Hand the whole set of training examples over to a supervised ML system
- Support Vector Machines

Result is a set of classifiers that can be applied to unseen sentences
More Specifically

- Given an input sentence and a predicate, we can use those classifiers to
  1. Identify all (and only) the non-overlapping segments in the input that correspond to arguments to that predicate
  2. Assign all the identified segments a label, or tag, from a fixed set of labels that capture semantic roles
  3. Performance: between 80% and 90%
Constituent Based Algorithm

- Obtain syntax tree for the sentence
- For each predicate
  - For each node in the syntax tree
    - Generate feature vector relative to the predicate
    - Classify feature vector using One versus All Support Vector Machine (SVM) classifiers
    - Convert raw SVM scores to probabilities [Platt, 2000]
  - Create a lattice of arguments
  - Using an argument language model, perform Viterbi search on the lattice to obtain best sequence
A relevant example

The patient returns to the outpatient clinic today for follow-up

The patient will complete his thiotepa dose today, and he will return tomorrow for the last dose of his thiotepa.

His donor completed stem-cell collection yesterday.
The patient returns to the outpatient clinic today for follow-up. The patient will complete his thiotepa dose today.

, and he will return tomorrow for the last dose of his thiotepa.

His donor completed stem-cell collection yesterday.
The patient returns to the outpatient clinic today for follow-up. The patient will complete his thiotepa dose today, and he will return tomorrow for the last dose of his thiotepa.

His donor completed stem-cell collection yesterday.
The patient returns to the outpatient clinic today for follow-up. The patient will complete his thiotepa dose today, and he will return tomorrow for the last dose of his thiotepa. His donor completed stem-cell collection yesterday.
The patient returns to the outpatient clinic today for follow-up. He will complete his thiotepa dose today, and he will return tomorrow for the last dose of his thiotepa. His donor completed stem-cell collection yesterday.
The patient returns to the outpatient clinic today for follow-up. The patient will complete his thiotepa dose today, and he will return tomorrow for the last dose of his thiotepa.

His donor completed stem-cell collection yesterday.
Normalizing across syntactic variations

- “bladder showed very mild trabeculation” -> “trabeculated bladder”

- bladder showed very mild trabeculation
  - show-trabeculation
  - Light verb + nominalization construction, eg., make an appearance
  - Treated as a multi-word expression
A relevant example

bladder showed very mild trabeculation

trabeculated bladder
Instantiating Templates – UMLS relations

- CAP Protocol ColoRectal Cancer
  - Procedure
  - Tumor site
  - Size
  - Histology (=the final diagnosis)
  - Grade
  - Tumor extension
  - Margins
  - Lymph nodes
GROSS DESCRIPTION
A. Sigmoid, Colon biopsy: (1 piece 1.1 x 1.0 x 0.6 cm)
B. Rectum, Colon biopsy: (12 pieces 0.1 - 0.4 cm in diameter)

FINAL DIAGNOSIS
A1) Colon, sigmoid, endoscopic biopsy: Tubular adenoma, low grade dysplasia
B1) Colon, rectum, endoscopic biopsy: Invasive grade 3 (of 4) adenocarcinoma arising in a tubulovillous adenoma.
  High grade dysplasia. Tumor extends to biopsy margins.

annotator: creation date: Wed Jun 02 11:44:41
GROSS DESCRIPTION

A. Sigmoid, Colon biopsy: (1 piece 1.1 x 1.0 x 0.6 cm)
B. Rectum, Colon biopsy: (12 pieces 0.1 - 0.4 cm. in diameter)

FINAL DIAGNOSIS

A1) Colon, sigmoid, endoscopic biopsy: Tubular adenoma, low grade dysplasia

B1) Colon, rectum, endoscopic biopsy: Invasive grade 3 (of 4) adenocarcinoma arising in a tubulovillous adenoma, high grade dysplasia. Tumor extends to biopsy margins.
GROSS DESCRIPTION

A. Sigmoid, Colon biopsy  (1 piece 1.1 x 1.0 x 0.8 cm)
B. Rectum, Colon biopsy  (12 pieces 0.1 - 0.4 cm. in diameter)

FINAL DIAGNOSIS

A1) Colon, sigmoid, endoscopic biopsy, Tubular adenoma, low grade dysplasia.

B1) Colon, rectum, endoscopic biopsy, Invasive grade 3 of 4 adenocarcinoma arising in a tubulovillous adenoma, high grade dysplasia. Tumor extends to deep margins.
From the pathology note/clinical note

RESULTS IN →

Treatment prescribed by the doctor ....
Summary

- Domain specific annotation of rich semantic representations will enable machine learning of automatic taggers that will identify events and relations between them.
- Domain adaptation will also use error analysis, additional selective annotation and feature tuning to improve performance.