An Overview of Biomedical Entity Recognition

Jeffry Scott

21 April 2010

1 Biomedical Entities

What is an Entity?

- NE: Named Entity, uniquely identified.
  - Person: Abraham Lincoln, Mae West
  - Place: Taj Mahal, Cairo, Washington D.C., The Pentagon
  - Thing: Gone With The Wind (a book), Gone With The Wind (a movie), Kleenex, The Beatles, Exxon Corporation, Her Majesty's Royal Navy.
  - Not a common noun.
- NER: Named Entity Recognition, techniques for named entity identification.

Problems in Entity Identification

- Feature Specification:
  - English: capitalization indicates proper names. Very language dependent. What of messages using all capital letters (telegrams, military message traffic)?
  - Other languages: ?
- Ambiguity
  - Will Smith: the actor? rapper? director? movie producer? UK comedien? football player?
  - May: the month? A girl named May? The verb *may* at the beginning of a question: *May I sit down?*

Problems in Entity Identification (cont.)

- Variability in Spelling
  - Exxon Corporation vs. Exxon vs. Exxon Corp. vs. Exxon, Inc.
- Volume
  - Low volume/small corpora easier to analyze with a lexicon or a rule set
  - Higher volume/very large corpora need to be quickly addressed with some loss of precision.
- Unknown Names
  - New names are introduced constantly, mandating constant updates to a lexicon or rule set.

1.1 Methods

Methods for NE Identification

- **Lexicon**: precise, but slow, requires constant update.
- **Rule Based**: quicker response than a lexicon, but exceptions are not handled well. Also requires regular updates.
- **Statistical**: many different statistical methods exist. Efficacy varies depending on training data, feature set and other factors. All require training data to *learn* desired patterns.
- **Hybrid Systems**: combining two or more methods.
- **Voting**: use several different statistical techniques where each method gets a vote. Choose the result with the most votes.

Computational Linguistic Statistical Methods

- Find statistical patterns when rule based patterns cannot be used.
- Heavily borrows from Pattern Recognition in EE and AI, Statistical Methods for Classification, Clustering in Statistics.
- Known as *'Machine Learning'*.
- Can handle higher volume than rule based methods.
- Allows for unknown data.
- Requires
  - A set of features defined to characterize the data
  - Training data to establish patterns for use in classification
  - A sufficient amount of training data
  - Training data that is representative of the target corpus

Common Statistical Methods

- Naive Bayes
- HMM (Hidden Markov Model)
- MaxEnt (Maximum Entropy)
- SVM (Support Vector Machine)
- CRF (Conditional Random Fields)

1.2 POS Tagging

Part-Of-Speech (POS) Analysis

- POS provides insight into what might be an entity.
- POS tagging an intermediate step
- Components:

- Tokenizer: find individual words, defined by white space or special characters.
- Lexicon: a list of words, corresponding POS tags
- Annotated corpus: used to define the lexicon, and provide training data
- Statistical method: method used for pattern recognition

POS Tagging Example

The patient was evaluated for repair of false femoral aneurysm.

- Simple Tagging
- *The*/D *patient*/N *was*/V *evaluated*/V *for*/P *repair*/N *of*/P *false*/ADJ *femoral*/ADJ *aneurysm*/N.
- Penn Treebank Style (S (NP (DT *The*) (NN *patient*) (VP (VP (VBD *was*) (VBG *evaluate*d)) (PP (IN *for*) (NN *repair*) (PP (IN *of*) (JJ *false*) (JJ *femoral)* (NN *aneurysm*)))))

Unknown Words in POS Tagging

Q: How are unknown words handled?

- The/D XXXX was/V evaluated/V for/P renal/ADJ failure/N.
  - The/D
  - The/D XXXX
  - The/D was/V
  - The/D was/V evaluated/V for/P renal/ADJ failure/N.
  - The/D XXXX/N was/V evaluated/V for/P renal/ADJ failure/N.

1.3 NP Chunking

NP Chunking

- Entities of interest beyond scope of Named Entities
- Usually defined as NPs (Noun Phrases)

- Different techniques required beyond Lexicon and Rule Sets
  - Compound nouns: *diabetes medication*
  - Adjectives: *distended abdomen*
  - NP + PP: *Queen of England*
  - Complex phrases: *the man who would be king*
- Segmentation primary method for NP Chunking

1.4 Segmentation (Sequence Labeling)

Segmentation (Sequence Labeling)

- Used to identify non-overlapping sequences of text.
- Requires a statistical method, features, training data.
- Uses IOB convention to identify tokens in segments
  - I: (token) inside segment
  - O: outside segment
  - B: beginning of segment
- POS tag a common feature for segmentation tasks

Segmentation Example

Sample sentence: *The patient was evaluated for repair of heart valve.*

Labeling NP (noun phrases):

$[The$	$patient]$	was	$evaluated$	for	$[repair]$	of	$[heart$	$valve]$	EOS
B	I	O	O	O	B	O	B	I	O

Labeling PP (prepositional phrases):

The	$patient$	was	$evaluated$	$[for$	$repair]$	$[of$	$heart$	$valve]$	EOS
O	O	O	O	B	I	B	I	I	O

* EOS == End of Sentence

BER Checklist

- Use NLP techniques for NER
  - POS Tagging
  - Segmentation
  - Word Sense Disambiguation
- Augment with biomedical lexicon
- Allow for new "unkown" entities

2 i2b2 Concepts

i2b2 Concepts

- Varied syntax beyond BER and NP
- Semantic categories:
  - *Medical Problems*
  - *Treatments*
  - *Tests*
- Exclusion of Concepts from Semantic Categories
- Relation and Assertion tasks build on the Concepts task

2.1 Concept Syntax

Concept Syntax Highly Varied

More than just NP chunks:

NP (noun phrase):	*high grade fever*
Compound noun:	*diabetes medication*
AP (adjective phrase):	*actively ischemic*
NP + PP*:	*placement of stent*
	subtotal occlusion of the RCA

* With restrictions on the type of PP (prepositional phrase) that may be used.

Q: what is the definition of a partial noun phrase?

2.2 Concept Semantic Categories

2.2.1 Medical Problems

Category: Medical Problems

- Disease name, syndrome, sign, symptom
- Mental or behavioral status
- Virus or bacterium
- Injury
- Abnormality

Concern: this could be a very long list of entities.

2.2.2 Treatments

Category: Treatments

- Medications: brand names, generic names, collective names
- Biological substances
- Drugs, treatment delivery devices
- Treatment procedures, related devices and hardware

2.2.3 Tests

Category: Tests

- Test procedures
- Panels and tests on body fluids
- Physiologic measures and vital signs
- Physical examination

2.3 Category Exclusion

Exclude from Categories

Medical Problems

- Normal states of health
- Physiologic measurements, vital signs
- Verbs describing outcome

Treatments

- Verbs indicating application of treatment

Tests

- Verbs indicating application of treatment
- Test values and measurements
- Mentions of tests stated as problems

Partial noun phrases are exluded from all Concept Categories.

Methods for Exclusion

Requires explicit steps/techniques to exclude concepts from a category.

1. Explicit rules for exclusion
2. Statistical training data for concepts to exclude
3. Both

2.4 Open Issues, Recommendations

Open Issues

1. New concepts properly identified? Steps to make concept identification robust, even for new concepts?
2. New concept mapping to Concept Category? How will that be done?
3. Low data volume: the i2b2 test data set will be small, and may not be sufficient for training.
4. What is a *partial noun phrase*?

Partial Noun Phrase Exclusion

Definition is by example, and seems incomplete.

Medical Problem: He was a [*moderately obese*] man in acute respiratory distress.

- *moderately obese* is marked as a partial noun phrase.
- wouldn't *a moderately obese man* be appropriate?
- or *man in acute respiratory distress*?

Treatment: [The needle jejunostomy tube] was utilized on the first post[operative] day.

- operative (in postoperative) is marked as a partial noun phrase.
- unclear as to what this example really shows.

Recommendations

1. Find Comprehensive Lexicon: UMLS or something like it, addressing wide range of biomedical entities.
2. Mix of POS Tagging and NP Chunking: low data volume gives more opportunities to focus on precision and recall where high throughput not needed.
3. Low data volume: supplement with additional corpora to test and train UW system.
4. GENIA corpus: use the GENIA corpus for training and test data. i2b2 annotation may be required.
5. Additional clinical data: acquire more annotated clinical data. Some annotation by the i2b2 team may be needed.

Recommendations (cont.)

1. Use Metamap: develop scheme for mapping Metamap (UMLS) concepts to i2b2 Semantic Categories. Use Metamap to vet concepts and assist in i2b2 concept classification. Given low data volume this seems reasonable.

3 Questions

Questions?