Lecture 12b: NLP (Introduction, Part 2)

Projects

Project #2
- Code freeze is now April 17
- Bug fixes, etc., allowed
- Goal: analysis
- Test on data I release.
- Instead of improving code, analyze what went wrong & what went well
- Failure/friction is key to picking research topics
- Papers due April 24

Announcements

Pipeline of NLP Tools

Scraping (not covered here)
Sentence splitting
Tokenization
(Stemming)
Part-of-speech tagging
Shallow parsing
Named entity recognition
Syntactic parsing
(Semantic Role Labeling)

Sentence Splitting

Usually easy (./?/! followed by spaces followed by capital letter), but...

IL-33 is known to induce the production of Th2-associated cytokines (e.g. IL-5 and IL-13).

Two solutions:
- Add more rules to handle exceptions
- Machine learning

Tokenization

Usually easy: group characters into tokens
- We don’t want to analyze documents at the level of characters
- But again, there are exceptions:
  - Commas
  - 2,6-diaminohexanoic acid
  - tricyclo(3.3.1.13,7)decanone
  - Four kinds of hyphens
    - "Syntactic."
    - Calcium-dependent
    - Hsp-60
    - Knocked-out gene: lush--flies
    - Negation: fever
    - Electric charge: Cl-

Full text → Index terms

K. Cohen NAACL-2007
Word frequency characteristics

Zipf’s Law: rank \* frequency = constant
(Most frequent word twice as common as second most frequent, three times as common as the third most frequent, etc.)

Organization of Inverted Files

Lexeme, Lexicon & Lemma

Lexeme: Smallest unit of language which has a meaning (roughly dictionary entry), e.g. run
- Takes various inflected word forms, e.g. runs, running, ran
- conduct (verb) is a different lexeme from conduct (noun)
Lexicon: A finite set of lexemes (roughly dictionary)
Lemma: The canonical or basic form that represents the lexeme, e.g. run

Lemmatization is not Trivial

May depend on the context
- He found the ball \(\rightarrow\) find
- He will found the Institute \(\rightarrow\) found
Depends on the part of speech
- He conducted the orchestra \(\rightarrow\) conduct (verb)

Is lemmatization useful?

For IR, some improvement
- especially for smaller documents
Helps on average, but not a lot
- Word sense disambiguation on query terms: business may be stemmed to busy, saw (the tool) to see
- Most studies for stemming for IR done for English
- may help more for other languages
- The possibility of letting people interactively influence the stemming has not been studied much

Part-of-speech tagging

The peri-kappa B site mediates human immunodeficiency

Assign a part-of-speech tag to each token in a sentence.
Part-of-speech tags

The Penn Treebank tagset: Part of Speech Tags (36)

https://www.clips.uantwerpen.be/pages/mbsp-tags

Part-of-speech tags

The Penn Treebank tagset: Chunk Tags (8)

https://www.clips.uantwerpen.be/pages/mbsp-tags

Part-of-speech tags

The Penn Treebank tagset: roles

https://www.clips.uantwerpen.be/pages/mbsp-tags

Writing rules for part-of-speech tagging

I have to go to school.  
I had a go at skiing.

If the previous word is “to”, then it’s a verb. 
If the previous word is “a”, then it’s a noun. 
If the next word is ...

: 

Writing rules manually is impossible

Learning from examples

The involvement of ion channels in B and T lymphocyte activation is supported by many reports of changes in ion fluxes and membrane 

Unseen text 

We demonstrate that... 

We demonstrate

Machine Learning Algorithm 

Unseen text
POS tagging with Hidden Markov Models

\[
P(t_1...t_n \mid w_1...w_n) = \frac{P(w_1...w_n \mid t_1...t_n)P(t_1...t_n)}{P(w_1...w_n)}
\]

\[
\propto P(w_1...w_n \mid t_1...t_n)P(t_1...t_n)
\]

\[
\approx \prod_{i=1}^{n} P(w_i \mid t_i)P(t_i \mid t_{i-1})
\]

Linguist Push Back

Where is the linguistic knowledge?
What does it teach us about language?

The Markov assumption is wrong!
- In this case, it assumes that \(P(\text{POS tag})\) is a function of only
  - The previous POS tag &
  - The current word
- Linguists have examples where this is wrong
  - POS tag depends on earlier words, not just earlier POS tags
  - POS tag depends on POS tag from several tags ago
  - I admit, I don’t know what these examples are
- But HMM’s work really well anyway
- What does that tell us?

Machine learning using diverse features

We want to use diverse types of information when predicting the tag.

He        opened         it

The word is \(\text{opened}\)\n
The suffix is \(\text{ed}\)\n
The previous word is \(\text{He}\)\n
Maximum entropy learning and Conditional Random Fields

Maximum entropy learning
- Log-linear modeling + MLE
- Parameter estimation
  - likelihood of each sample
  - Model expectation of each feature

Conditional Random Fields
- Log-linear modeling on the whole sentence
- Features are defined on states and edges
- Dynamic programming

POS tagging algorithms

Performance on the Wall Street Journal corpus

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Training Cost</th>
<th>Speed</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependency Net (2003)</td>
<td>Low</td>
<td>Low</td>
<td>97.2</td>
</tr>
<tr>
<td>Conditional Random Fields</td>
<td>High</td>
<td>High</td>
<td>97.1</td>
</tr>
<tr>
<td>Support vector machines (2003)</td>
<td>High</td>
<td>High</td>
<td>97.1</td>
</tr>
<tr>
<td>Bidirectional MEMM (2005)</td>
<td>Low</td>
<td>Low</td>
<td>97.1</td>
</tr>
<tr>
<td>Brill’s tagger (1995)</td>
<td>Low</td>
<td>High</td>
<td>98.6</td>
</tr>
<tr>
<td>HMM (2000)</td>
<td>Very low</td>
<td>High</td>
<td>96.7</td>
</tr>
</tbody>
</table>
POS taggers

Brill's tagger
  • http://www.cs.jhu.edu/~brill/

TnT tagger
  • http://www.coli.uni-saarland.de/~thorsten/tnt/

Stanford tagger
  • http://nlp.stanford.edu/software/tagger.shtml

SVMTool
  • http://www.lsi.upc.es/~nlp/SVMTool/

GENIA tagger
  • http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

Tagging errors made by a WSJ-trained POS tagger

… and membrane potential after mitogen binding.

... two factors, which bind to the same kappa B enhancers...

... by analysing the amino acid sequence.

... to contain T cell determinants than ...

Stimulation of interferon beta gene transcription monitored by

General Taggers work poorly on Special Domains

Performance of the Brill tagger evaluated on randomly selected 1000 MEDLINE sentences: 86.8% (Smith et al., 2004)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>84.4%</td>
<td></td>
</tr>
<tr>
<td>NNP = NN, NNPS = NNS</td>
<td>90.0%</td>
<td></td>
</tr>
<tr>
<td>LJ = NN</td>
<td>91.3%</td>
<td></td>
</tr>
<tr>
<td>JJ = NN</td>
<td>94.9%</td>
<td></td>
</tr>
</tbody>
</table>

Accuracies of a WSJ-trained POS tagger evaluated on the GENIA corpus (Tsuruoka et al., 2005)

Chunking (shallow parsing)

He reckons the current account deficit will narrow to only # 1.8 billion in September.

A chunker (shallow parser) segments a sentence into non-recursive phrases.

Extracting noun phrases from MEDLINE (Bennett, 1999)

Rule-based noun phrase extraction
  • Tokenization
  • Part-Of-Speech tagging
  • Pattern matching

Noun phrase extraction accuracies evaluated on 40 abstracts

<table>
<thead>
<tr>
<th></th>
<th>FastNPE</th>
<th>NPTool</th>
<th>Chopper</th>
<th>AZ Phraser</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>50%</td>
<td>85%</td>
<td>97%</td>
<td>92%</td>
</tr>
<tr>
<td>Precision</td>
<td>80%</td>
<td>80%</td>
<td>90%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Chunking with Machine learning

Chunking performance on Penn Treebank

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winnow (with basic features) (Zhang, 2002)</td>
<td>90.60</td>
<td>93.54</td>
<td>93.57</td>
</tr>
<tr>
<td>Perceptron (Carreras, 2003)</td>
<td>93.29</td>
<td>94.19</td>
<td>93.74</td>
</tr>
<tr>
<td>SVM + voting (Kudo, 2003)</td>
<td>93.92</td>
<td>93.89</td>
<td>93.91</td>
</tr>
<tr>
<td>SVM (Kudo, 2000)</td>
<td>93.51</td>
<td>93.45</td>
<td>93.48</td>
</tr>
<tr>
<td>Bidirectional MEMM (Tsuruoka, 2005)</td>
<td>93.70</td>
<td>93.70</td>
<td>93.70</td>
</tr>
</tbody>
</table>