NLP Tools

L715/B659

Dept. of Linguistics, Indiana University
Fall 2014
Tools

We’ll provide a lot of examples from the Stanford NLP tools, but there are many others available

▶ http://nlp.stanford.edu/software/

Many tools come bundled as part of the CoreNLP package, which we’ll work with:

▶ http://nlp.stanford.edu/software/corenlp.shtml
CoreNLP tools

The tools include:

▶ Tokenizer (tokenize)
▶ Sentence splitter (ssplit)
▶ Part-of-speech (POS) tagger (pos)
▶ Named entity recognizer (NER) (ner)
▶ Parser (parse)
▶ Coreference resolution system (dcoref)
▶ Sentiment analysis system (sentiment)
▶ Bootstrapped pattern learning tools
./corenlp.sh
java -mx3g -cp "./" edu.stanford.nlp.pipeline.StanfordCoreNLP
Searching for resource: StanfordCoreNLP.properties
Searching for resource: edu/stanford/nlp/pipeline/StanfordCoreNLP.properties
Adding annotator tokenize
Adding annotator ssplit
Adding annotator pos
Reading POS tagger model ...
Adding annotator lemma
Adding annotator ner
Loading classifier ...
Initializing JollyDayHoliday for sutime with ...
Reading TokensRegex rules ...
Ignoring inactive rule: null
Ignoring inactive rule: temporal-composite-8:ranges
Reading TokensRegex rules from ...
Adding annotator parse
Loading parser from serialized file edu/stanford/nlp/models/lexparser/englishPCFG.ser.gz
Adding annotator dcoref
Interactive shell

Entering interactive shell. Type q RETURN or EOF to quit.
NLP> Analyze this sentence!
Sentence #1 (4 tokens):
Analyze this sentence!
[Text=Analyze CharacterOffsetBegin=0 CharacterOffsetEnd=7
  PartOfSpeech=VB Lemma=analyze NamedEntityTag=O]
[Text=this CharacterOffsetBegin=8 CharacterOffsetEnd=12
  PartOfSpeech=DT Lemma=this NamedEntityTag=O]
[Text=sentence CharacterOffsetBegin=13 CharacterOffsetEnd=21
  PartOfSpeech=NN Lemma=sentence NamedEntityTag=O]
[Text=! CharacterOffsetBegin=21 CharacterOffsetEnd=22
  PartOfSpeech=. Lemma=! NamedEntityTag=O]
(ROOT
  (S
    (VP (VB Analyze)
      (NP (DT this) (NN sentence)))
    (.
      !)))

root(ROOT-0, Analyze-1)
det(sentence-3, this-2)
dobj(Analyze-1, sentence-3)
Specifying annotators

1. Java properties file
   - Content of config.properties:
     annotators = tokenize, ssplit, pos, lemma, ner, parse, dcoref
   - Command: java -cp "*" -Xmx2g edu.stanford.nlp.pipeline.StanfordCoreNLP -props config.properties -file input.txt

2. Command line: java -cp "*" -Xmx2g edu.stanford.nlp.pipeline.StanfordCoreNLP -annotators tokenize,ssplit,pos,lemma,ner,parse,dcoref -file input.txt
   - These are the default annotators
Working from another directory

To call the files, make sure the classpath (cp) is set properly, e.g.,

▶ Command line: java -cp "/path/to/stanfordcorenlp/*" -Xmx2g edu.stanford.nlp.pipeline.StanfordCoreNLP -file input.txt
  ▶ Note the asterisk
  ▶ Make sure you have input.txt present
Output format
One token

<token id="1">
  <word>Stanford</word>
  <lemma>Stanford</lemma>
  <CharacterOffsetBegin>0</CharacterOffsetBegin>
  <CharacterOffsetEnd>8</CharacterOffsetEnd>
  <POS>NNP</POS>
  <NER>ORGANIZATION</NER>
</token>
Output format
Constituency parse

<parse>(ROOT (S (NP (NNP Stanford) (NNP University)))
  (VP (VBZ is) (VP (VBN located) (PP (IN in) (NP (NNP California)))))) (. .))) </parse>
<dependencies type="basic-dependencies">
  <dep type="nn">
    <governor idx="2">University</governor>
    <dependent idx="1">Stanford</dependent>
  </dep>
  <dep type="nsubjpass">
    <governor idx="4">located</governor>
    <dependent idx="2">University</dependent>
  </dep>
  ...
</dependencies>
<coreference>
  <coreference>
    <mention representative="true">
      <sentence>2</sentence>
      <start>3</start>
      <end>6</end>
      <head>5</head>
    </mention>
    <mention>
      <sentence>2</sentence>
      <start>1</start>
      <end>2</end>
      <head>1</head>
    </mention>
  </coreference>
</coreference>

See: http://nlp.stanford.edu/software/corenlp_xml_description.shtml
Outputting as text

java -cp "*" -Xmx2g edu.stanford.nlp.pipeline.StanfordCoreNLP -props config.properties -file input.txt -outputFormat text

[Text=Stanford CharacterOffsetBegin=0 CharacterOffsetEnd=8 PartOfSpeech=NNP Lemma=Stanford NamedEntityTag=ORGANIZATION] ...

(ROOT
  (S
    (NP (NNP Stanford) (NNP University))
    (VP (VBZ is)
      (ADJP (JJ located)
        (PP (IN in)
          (NP (NNP California))))))
  (. .)))

root(ROOT-0, located-4)
nn(University-2, Stanford-1)
Different annotator options

Each of the annotators has various options associated with it, e.g.,

▶ **clean.xmltags**: Discard xml tag tokens that match this regular expression. For example, .* will discard all xml tags

▶ **sutime.markTimeRanges**: Tells sutime to mark phrases such as ”From January to March” instead of marking ”January” and ”March” separately

▶ **regexner.mapping**: The name of a file, classpath, or URI that contains NER rules, i.e., the mapping from regular expressions to NE classes. ...

▶ **parse.flags**: flags to use when loading the parser model. ...
External code

There are some extensions to the CoreNLP package, including many wrappers in various languages

- Java, .NET, Python, Ruby, Perl, Scala, Clojure, JavaScript
Understanding the categories

For POS definitions:
- http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

For constituent syntactic phrases:
- table 3 of: https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html

For dependency categories:
Practice problem

1. Download & unpack the ICE-Nigeria corpus: http://sourceforge.net/projects/ice-nigeria/
   ▶ Note: part of the International Corpus of English (ICE): http://ice-corpora.net/ICE/INDEX.HTM
2. Under ice-nig/txt/written/, identify a file to work with
3. Provide linguistic annotations from the Stanford tools
   ▶ Consider the task of genre classification again: what linguistic annotations would be helpful?
4. Begin to extract linguistically-motivated features from the file you’re working with (using your favorite programming language)