Natural Language Processing for the Semantic Web

Isabelle Augenstein

Department of Computer Science, University of Sheffield, UK
i.augenstein@sheffield.ac.uk

Linked Data for NLP Tutorial, ESWC Summer School 2014

August 24, 2014
Why Natural Language Processing?

Isabelle Augenstein
Why Natural Language Processing?

Isabelle Augenstein

This major exhibition will explore a golden age in China’s history.

Between AD 1400 and 1450, China was a global superpower run by one family – the Ming dynasty – who established Beijing as the capital and built the Forbidden City. During this period, Ming China was thoroughly connected with the outside world. Chinese artists absorbed many fascinating influences, and created some of the most beautiful objects and paintings ever made.

The exhibition will feature a range of these spectacular objects – including exquisite porcelain, gold, jewellery, furniture, paintings, sculptures and textiles – from museums across China and the rest of the world. Many of them have only very recently been discovered and have never been seen outside China.
Why Natural Language Processing?

Isabelle Augenstein

semi-structured information

unstructured information

How to link this information to a knowledge base automatically?
Why Natural Language Processing?

Isabelle Augenstein

semi-structured information

unstructured information

How to link this information to a knowledge base automatically?

Information Extraction!
Between AD 1400 and 1450, China was a global superpower run by one family – the Ming dynasty – who established Beijing as the capital and built the Forbidden City.
Between AD 1400 and 1450, China was a global superpower run by one family – the Ming dynasty – who established Beijing as the capital and built the Forbidden City.

Named Entity Recognition

Isabelle Augenstein
Between AD 1400 and 1450, China was a global superpower run by one family – the Ming dynasty – who established Beijing as the capital and built the Forbidden City.

Named Entity Recognition (NER)

Named Entity Classification (NEC):

China: /location/country
Ming dynasty: /royalty/royal_line
Beijing: /location/city
Forbidden City: /location/city
Between AD 1400 and 1450, China was a global superpower run by one family – the Ming dynasty – who established Beijing as the capital and built the Forbidden City.

Named Entity Recognition

**Named Entity Classification:**
- China: /location/country
- Ming dynasty: /royalty/royal_line
- Beijing: /location/city
- Forbidden City: /location/city

**Named Entity Linking:**
- China: /m/0d05w3
- Ming dynasty: /m/0bw_m
- Beijing: /m/01914
- Forbidden City: /m/0j0b2
Named Entities: Proper nouns, which refer to real-life entities

Named Entity Recognition: Detecting boundaries of named entities (NEs)

Named Entity Classification: Assigning classes to NEs, such as PERSON, LOCATION, ORGANISATION, or fine-grained classes such as ROYAL LINE

Named Entity Linking / Disambiguation: Linking NEs to concrete entries in knowledge base, example:
China -> LOCATION: Republic of China, country in East Asia
  -> LOCATION: China proper, core region of China during Qing dynasty
  -> LOCATION: China, Texas
  -> PERSON: China, Brazilian footballer born in 1964
  -> MUSIC: China, a 1979 album by Vangelis
  -> …
Between AD 1400 and 1450, China was a global superpower run by one family – the Ming dynasty – who established Beijing as the capital and built the Forbidden City.
Between AD 1400 and 1450, China was a global superpower run by one family — the Ming dynasty — who established Beijing as the capital and built the Forbidden City.

Named Entity Recognition

Relation Extraction

Temporal Extraction

Isabelle Augenstein
Between AD 1400 and 1450, China was a global superpower run by one family – the Ming dynasty – who established Beijing as the capital and built the Forbidden City.

Named Entity Recognition

Relation Extraction

Temporal Extraction

Event Extraction

Event: /royalty/royal_line: Ming dynasty

/royalty/royal_line/ruled_from: 1400-XX-XX
/royalty/royal_line/ruled_to: 1450-XX-XX
/royalty/royal_line/kingdom_s_ruled: China
Relations: Two or more entities which relate to one another in real life

Relation Extraction: Detecting relations between entities and assigning relation types to them, such as CAPITAL-OF

Temporal Extraction: Recognising and normalising time expressions: times (e.g. “3 in the afternoon”), dates (“tomorrow”), durations (“since yesterday”), and sets (e.g. “twice a month”)

Events: Real-life events that happened at some point in space and time, e.g. kingdom, assassination, exhibition

Event Extraction: Extracting events consisting of the name and type of event, time and location
• Information extraction (IE) methods such as named entity recognition (NER), named entity classification (NEC), named entity linking, relation extraction (RE), temporal extraction, and event extraction can help to add markup to Web pages

• Information extraction approaches can serve two purposes:
  • Annotating every single mention of an entity, relation or event, e.g. to add markup to Web pages
  • Aggregating those mentions to populate knowledge bases, e.g. based on confidence values and majority voting
    China LOCATION 0.9
    China LOCATION 0.8
    China PERSON 0.4
    → China LOCATION
• Focus of the rest of the tutorial and hands-on session: Named entity recognition and classification (NERC)

• Possible methodologies
  • Rule-based approaches: write manual extraction rules
  • Machine learning based approaches
    • Supervised learning: manually annotate text, train machine learning model
    • Unsupervised learning: extract language patterns, cluster similar ones
    • Semi-supervised learning: start with a small number of language patterns, iteratively learn more (bootstrapping)
  • Gazetteer-based method: use existing list of named entities
  • Combination of the above
Developing a NERC involves programming based around APIs..
Developing a NERC involves programming based around APIs. which can be frustrating at times.

Isabelle Augenstein
and (at least basic) knowledge about linguistics
Background: Linguistics

Acoustic signal

- \( /l/ \)

Phonology, phonetics

- \(/l/ \rightarrow \text{sh}\)

Lexicon

- \(\text{wait} + \text{ed} \rightarrow \text{waited}, \text{cat} \rightarrow \text{cat}\)

Morphology

- \(\text{wait} \rightarrow V, \text{cat} \rightarrow N\)

Syntax

- \(\text{The (D) farmer (N) hit (V) the (D) donkey (N).}\)

Sentences

- \(\text{Every farmer who owns a donkey beats it.}\)

Semantics, Discourse

- \(\forall x \forall y (\text{farmer}(x) \land \text{donkey}(y) \land \text{own}(x, y) \rightarrow \text{beat}(x, y))\)

Isabelle Augenstein
Background: NLP Tasks

Acoustic signal

Phonemes

Morphemes

Words

Sentences

Meaning

Phonology, phonetics

Lexicon

Morphology

Syntax

Semantics, Discourse

Speech recognition

Lemmatisation or stemming, part of speech (POS) tagging

Chunking, parsing

Semantic and discourse analysis, anaphora resolution

Isabelle Augenstein
Background: Linguistics

- Acoustic signal -> /l/
  - Phonology, phonetics
    - recognize speech – wreck a nice beach
      - /[rek.əɡ.naɪz]/ [spi.tʃ] – /[rek]/ /ə/ /[nɑɪs]/ /[biːtʃ]/
  - Lexicon
  - Morphemes
  - Morphology
  - Words
  - Syntax
  - Sentences
  - Meaning
    - Semantics, Discourse

Isabelle Augenstein
Background: Linguistics

Acoustic signal -> /l/

Phonology, phonetics

Phonemes

Lexicon

Morphemes

Morphology

Words

Syntax

Sentences

Semantics, Discourse

Isabelle Augenstein
Background: Linguistics

Acoustic signal -> /l/

Phonology, phonetics
Phonemes
Lexicon
Morphemes
Morphology
Words
Syntax
Sentences
Meaning

recognize speech – wreck a nice beach
/'rek.əɡ.naɪz/ /spiːtʃ/ – /rek/ /ə/ /nɑɪs/ /biːtʃ/

She’d /ʃiːd/ -> she would, she had

Isabelle Augenstein
Background: Linguistics

Acoustic signal -> /l/

Phonology, phonetics
Phonemes

Lexicon
Morphemes

Words
Morphology

Sentences
Syntax

Meaning
Semantics, Discourse

Isabelle Augenstein
Background: Linguistics

**Acoustic signal**

- **Phonemes**
  - **Phonology, phonetics**
    - *recognize speech – wreck a nice beach*
  - **Lexicon**
    - *She’d /ʃiːd/ -> she would, she had*
  - **Morphology**
    - *Time flies(V/N) like(V/P) an arrow*

**Words**

**Syntax**

**Sentences**

**Meaning**

Semantics, Discourse

Isabelle Augenstein
Background: Linguistics

Acoustic signal -> /ɪ/

Phonology, phonetics
Phonemes
Lexicon
Morphemes
Morphology
Words
Syntax
Sentences
Meaning

Isabelle Augenstein

recognize speech – wreck a nice beach
/ˈrek.əɡ.naɪz/ /spiːtʃ/ – /rek/ /ə/ /naɪs/ /biːtʃ/

She’d /ʃiːd/ -> she would, she had

Time flies(V/N) like(V/P) an arrow

The woman saw the man with the binoculars.
Background: Linguistics

Acoustic signal -> /l/

- Phonology, phonetics
  - Phonemes
    - Lexicon
      - Morphemes
        - Morphology
          - Words
            - Syntax
              - Sentences
                - Meaning

- Recognize speech – wreck a nice beach
  - Phonology, phonetics
    - Phonemes
      - Lexicon
        - Morphemes
          - Morphology
            - Words
              - Syntax
                - Sentences
                  - Meaning

- The woman saw the man with the binoculars.

Isabelle Augenstein
Background: Linguistics

Acoustic signal

Phonemes

Phonology, phonetics

Lexicon

Morphemes

Morphology

Words

Syntax

Sentences

Meaning

Semantics, Discourse

Isabelle Augenstein

recognize speech – wreck a nice beach

She’d -> she would, she had

Time flies(V/N) like(V/P) an arrow

The woman saw the man with the binoculars. -> Who had the binoculars?

Somewhere in Britain, some woman has a child every thirty seconds.
Background: Linguistics

- Acoustic signal -> /l/
  - Phonology, phonetics
    - Phonemes
      - Lexicon
        - Morphemes
          - Morphology
            - Words
              - Syntax
                - Sentences
                  - Semantics, Discourse
                    - Meaning
                      - Recognize speech – wreck a nice beach
                        - Phonetics
                          - /ˈrek.əɡ.naɪz/ /spiːtʃ/ – /rek/ /ə/ /naiʃ/ /biːtʃ/
                          - Lexicon
                            - She’d /ʃiːd/ -> she would, she had
                              - Morphology
                                - Time flies (V/N) like (V/P) an arrow
                                  - Words
                                    - Syntax
                                      - Sentences
                                        - Semantics, Discourse
                                          - Meaning
                                            - Recognize speech – wreck a nice beach
                                              - Phonetics
                                                - /ˈrek.əɡ.naɪz/ /spiːtʃ/ – /rek/ /ə/ /naiʃ/ /biːtʃ/
                                                  - Lexicon
                                                    - She’d /ʃiːd/ -> she would, she had
                                                      - Morphology
                                                        - Time flies (V/N) like (V/P) an arrow
                                                          - Words
                                                            - Syntax
                                                              - Sentences
                                                                - Semantics, Discourse
                                                                  - Meaning
                                                                    - Recognize speech – wreck a nice beach
                                                                      - Phonetics
                                                                        - /ˈrek.əɡ.naɪz/ /spiːtʃ/ – /rek/ /ə/ /naiʃ/ /biːtʃ/
Background: Linguistics

Acoustic signal -> /l/

Phonology, phonetics

Phonemes

Lexicon

Morphemes

Morphology

Words

Syntax

Sentences

Semantics, Discourse

Meaning

Isabelle Augenstein

recognize speech – wreck a nice beach
/'rek.əɡ.naɪz/ /spiːtʃ/ – /rek/ /eɪ/ /naɪs/ /biːtʃ/

She’d /ʃiːd/ -> she would, she had

Time flies(V/N) like(V/P) an arrow

The woman saw the man with the binoculars. -> Who had the binoculars?

Somewhere in Britain, some woman has a child every thirty seconds. -> Same woman or different women?
Background: Linguistics

Acoustic signal -> /l/

Phonemes

Morpheme

Words

Sentences

Meaning

Ambiguities on every level

vreck a nice beach

buld, she had

/P) an arrow

man with the

ad the binoculairs?

woman some woman

ty seconds.

different women?

Isabelle Augenstein
Language is ambiguous..

Can we still build named entity extractors that extract all entities from unseen text correctly?

Isabelle Augenstein
Language is ambiguous.

Can we still build named entity extractors that extract all entities from unseen text correctly?

Isabelle Augenstein
Language is ambiguous..

Can we still build named entity extractors that extract all entities from unseen text correctly?

However, we can try to extract most of them correctly using linguistic cues and background knowledge!

Isabelle Augenstein
What can help to recognise and/or classify named entities?

- Words:
  - Words in window before and after mention
  - Sequences
  - Bags of words

Between AD 1400 and 1450, China was a global superpower

w: China  w-1: ,  w-2: 1450  w+1: was  w+2: a
seq[-]: 1450,  seq[+]: was a
bow: China  bow[-]: ,  bow[-]: 1450  bow[+]: was  bow[+]: a

Isabelle Augenstein
What can help to recognise and/or classify named entities?

- **Morphology:**
  - Capitalisation: is upper case *(China)*, all upper case *(IBM)*, mixed case *(eBay)*
  - Symbols: contains $, £, €, roman symbols *(IV)*, ..
  - Contains period *(google.com)*, apostrophe *(Mandy’s)*, hyphen *(speed-o-meter)*, ampersand *(Fisher & Sons)*
  - Stem or Lemma *(cats -> cat)*, prefix *(disadvantages -> dis)*, suffix *(cats -> s)*, interfix *(speed-o-meter -> o)*
What can help to recognise and/or classify named entities?

- POS (part of speech) tags
  - Most named entities are nouns

![Bar chart showing counts of valid and invalid POS tags](Prokofyev (2014))
<table>
<thead>
<tr>
<th>Number</th>
<th>Tag</th>
<th>Description</th>
<th>Number</th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>CC</td>
<td>Coordinating conjunction</td>
<td>18.</td>
<td>PRP</td>
<td>Personal pronoun</td>
</tr>
<tr>
<td>2.</td>
<td>CD</td>
<td>Cardinal number</td>
<td>19.</td>
<td>PRP$</td>
<td>Possessive pronoun</td>
</tr>
<tr>
<td>3.</td>
<td>DT</td>
<td>Determiner</td>
<td>20.</td>
<td>RB</td>
<td>Adverb</td>
</tr>
<tr>
<td>4.</td>
<td>EX</td>
<td>Existential <em>there</em></td>
<td>21.</td>
<td>RBR</td>
<td>Adverb, comparative</td>
</tr>
<tr>
<td>5.</td>
<td>FW</td>
<td>Foreign word</td>
<td>22.</td>
<td>RBS</td>
<td>Adverb, superlative</td>
</tr>
<tr>
<td>6.</td>
<td>IN</td>
<td>Preposition or subordinating conjunction</td>
<td>23.</td>
<td>RP</td>
<td>Particle</td>
</tr>
<tr>
<td>7.</td>
<td>JJ</td>
<td>Adjective</td>
<td>24.</td>
<td>SYM</td>
<td>Symbol</td>
</tr>
<tr>
<td>8.</td>
<td>JJR</td>
<td>Adjective, comparative</td>
<td>25.</td>
<td>TO</td>
<td>to</td>
</tr>
<tr>
<td>9.</td>
<td>JJS</td>
<td>Adjective, superlative</td>
<td>26.</td>
<td>UH</td>
<td>Interjection</td>
</tr>
<tr>
<td>10.</td>
<td>LS</td>
<td>List item marker</td>
<td>27.</td>
<td>VB</td>
<td>Verb, base form</td>
</tr>
<tr>
<td>11.</td>
<td>MD</td>
<td>Modal</td>
<td>28.</td>
<td>VBD</td>
<td>Verb, past tense</td>
</tr>
<tr>
<td>12.</td>
<td>NN</td>
<td>Noun, singular or mass</td>
<td>29.</td>
<td>VBG</td>
<td>Verb, gerund or present participle</td>
</tr>
<tr>
<td>13.</td>
<td>NNS</td>
<td>Noun, plural</td>
<td>30.</td>
<td>VBN</td>
<td>Verb, past participle</td>
</tr>
<tr>
<td>14.</td>
<td>NNP</td>
<td>Proper noun, singular</td>
<td>31.</td>
<td>VBP</td>
<td>Verb, non-3rd person singular present</td>
</tr>
<tr>
<td>15.</td>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td>32.</td>
<td>VBZ</td>
<td>Verb, 3rd person singular present</td>
</tr>
<tr>
<td>16.</td>
<td>PDT</td>
<td>Predeterminer</td>
<td>33.</td>
<td>WDT</td>
<td>Wh-determiner</td>
</tr>
<tr>
<td>17.</td>
<td>POS</td>
<td>Possessive ending</td>
<td>34.</td>
<td>WP</td>
<td>Wh-pronoun</td>
</tr>
<tr>
<td>18.</td>
<td></td>
<td></td>
<td>35.</td>
<td>WP$</td>
<td>Possessive wh-pronoun</td>
</tr>
<tr>
<td>36.</td>
<td></td>
<td></td>
<td></td>
<td>WRB</td>
<td>Wh-adverb</td>
</tr>
</tbody>
</table>
### Morphology: Penn Treebank POS tags

<table>
<thead>
<tr>
<th>Number</th>
<th>Tag</th>
<th>Description</th>
<th>Numbers</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>CC</td>
<td>Coordinating conjunction</td>
<td>18.</td>
<td>PRP</td>
</tr>
<tr>
<td>2.</td>
<td>CD</td>
<td>Cardinal number</td>
<td>19.</td>
<td>PRP$</td>
</tr>
<tr>
<td>3.</td>
<td>DT</td>
<td>Determiner</td>
<td>20.</td>
<td>RB</td>
</tr>
<tr>
<td>4.</td>
<td>EX</td>
<td>Existential there</td>
<td>21.</td>
<td>RBR</td>
</tr>
<tr>
<td>5.</td>
<td>FW</td>
<td>Foreign word</td>
<td>22.</td>
<td>RBS</td>
</tr>
<tr>
<td>6.</td>
<td>IN</td>
<td>Preposition or subordinating conjunction</td>
<td>23.</td>
<td>RP</td>
</tr>
<tr>
<td>7.</td>
<td>JJ</td>
<td>Adjective</td>
<td>24.</td>
<td>SYM</td>
</tr>
<tr>
<td>8.</td>
<td>JJR</td>
<td>Adjective, comparative</td>
<td>25.</td>
<td>TO</td>
</tr>
<tr>
<td>9.</td>
<td>JJS</td>
<td>Adjective, superlative</td>
<td>26.</td>
<td>UH</td>
</tr>
<tr>
<td>10.</td>
<td>LS</td>
<td>List item marker</td>
<td>27.</td>
<td>VB</td>
</tr>
<tr>
<td>11.</td>
<td>MD</td>
<td>Modal</td>
<td>28.</td>
<td>VBD</td>
</tr>
<tr>
<td>12.</td>
<td>NN</td>
<td>Noun, singular or mass</td>
<td>29.</td>
<td>VBG</td>
</tr>
<tr>
<td>13.</td>
<td>NNS</td>
<td>Noun, plural</td>
<td>30.</td>
<td>VBN</td>
</tr>
<tr>
<td>14.</td>
<td>NNP</td>
<td>Proper noun, singular</td>
<td>31.</td>
<td>VBP</td>
</tr>
<tr>
<td>15.</td>
<td>NNPS</td>
<td>Proper noun, plural</td>
<td>32.</td>
<td>VBJ</td>
</tr>
<tr>
<td>16.</td>
<td>PDT</td>
<td>Predeterminer</td>
<td>33.</td>
<td>WDT</td>
</tr>
<tr>
<td>17.</td>
<td>POS</td>
<td>Possessive ending</td>
<td>34.</td>
<td>WP</td>
</tr>
<tr>
<td>18.</td>
<td>POS</td>
<td>Possessive ending</td>
<td>35.</td>
<td>WP$</td>
</tr>
<tr>
<td>19.</td>
<td>POS</td>
<td>Possessive ending</td>
<td>36.</td>
<td>WRB</td>
</tr>
</tbody>
</table>

- **Adjectives (all start with J)**
- **Nouns (all start with N)**
- **Verbs (all start with V)**

#### Notes:
- **PRP**: Personal pronoun
- **PRP$**: Possessive pronoun
- **RB**: Adverb
- **RBR**: Adverb, comparative
- **RBS**: Adverb, superlative
- **SYM**: Symbol
- **TO**: to
- **UH**: Interjection
- **VB**: Verb, base form
- **VBD**: Verb, past tense
- **VBG**: Verb, gerund or present participle
- **VBN**: Verb, past participle
- **VBP**: Verb, non-3rd person singular present
- **VBZ**: Verb, 3rd person singular present
- **WDT**: Wh-determiner
- **WP**: Wh-pronoun
- **WP$**: Possessive wh-pronoun
What can help to recognise and/or classify named entities?

- POS (part of speech) tags
  - Most named entities are nouns

\[\text{count of valid vs. invalid POS tags} \]

\[\text{Prokofyev (2014)}\]
What can help to recognise and/or classify named entities?

• Gazetteers
  • Retrieved from HTML lists or tables
    
    | Religion          | Heaven worship, Taoism, Confucianism, Buddhism, Chinese folk religion, Islam |
    |-------------------|--------------------------------------------------------------------------------|

• Using regular expressions patterns and search engines (e.g. “Religions such as * ”)
• Retrieved from knowledge bases
Extensive choice of machine learning algorithms for training NERCs
Extensive choice of machine learning algorithms for training NERCs

- SVM
- Ensemble Classifiers
- Naive Bayes
- Text Data
- K-means
- MiniBatch Kmeans
- MeanShift YIGMM
- Spectral Clustering GMM
- SVC

- K-means
- SVC
- Ridge Regression SVR (kernel='linear')
- LLE

- SVR (kernel='rbf')
- Ensemble Regressors

- NOT WORKING
- <10K samples
- Tough luck
- Producing structure
- Kernel approximation

- NERD: Training Models
NERC: Training Models

• Unfortunately, there isn’t enough time to explain machine learning algorithms in detail
• CRFs (conditional random fields) are one of the most widely used algorithms for NERC
  • Graphical models, view NERC as a sequence labelling task
  • Named entities consist of a beginning token (B), inside tokens (I), and outside tokens (O)
    China (B-LOC) built (O) the (O) Forbidden (B-LOC) City (I-LOC) (O)
• For now, we will focus on rule- and gazetteer-based NERC
• It is fairly easy to write manual extraction rules for NEs, can achieve a high performance when combined with gazetteers
  • This can be done with the GATE software (general architecture for text engineering) and Jape rules
  -> Hands-on session

Isabelle Augenstein
Natural Language Processing:
- GATE (general purpose architecture, includes other NLP and ML software as plugins)
- Stanford NLP (Java)
- OpenNLP (Java)
- NLTK (Python)

Machine Learning:
- scikit-learn (Python, rich documentation, highly recommended!)
- Mallet (Java)
- WEKA (Java)
- Alchemy (graphical models, Java)
- FACTORIE (graphical models, Scala)
- CRFSuite (efficient implementation of CRFs, Python)
Ready to use NERC software:
- ANNIE (rule-based, part of GATE)
- Wikifier (based on Wikipedia)
- FIGER (based on Wikipedia, fine-grained Freebase NE classes)

Almost ready to use NERC software:
- CRFSuite (already includes Python implementation for feature extraction, you just need to feed it with training data, which you can also download)

Ready to use RE software:
- ReVerb (Open IE, extracts patterns for any kind of relation)
- MultiR (Distant supervision, relation extractor trained on Freebase)

Web Content Extraction software:
- Boilerpipe (extract main text content from Web pages)
- Jsoup (traverse elements of Web pages individually, also allows to extract text)
Thank you for your attention!

Questions?

Isabelle Augenstein