Information Extraction with Linked Data

Isabelle Augenstein

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

Information Extraction with Linked Data Tutorial, ESWC Summer School 2015

2 September 2015
When Arctic Monkeys burst into the spotlight in 2006, breaking sales records, prompting bandwagon-jumping prospective Labour party leaders to make themselves sound heroically foolish and being hailed for their quintessential Britishness, you could have got long odds on the key figure in the second phase of their career being Californian stoner-rock pioneer Josh Homme. And yet it was under the production aegis of the Queens of the Stone Age frontman that they performed a stylistic handbrake turn with 2009's *Humbug*, ditching the indie influences of their first two albums for a markedly heavier, darker sound. It also marked the point where Alex Turner's lyrical
When Arctic Monkeys burst into the spotlight in 2006, breaking sales records, prompting bandwagon-jumping prospective Labour party leaders to make themselves sound heroically foolish and being hailed for their quintessential Britishness, you could have got long odds on the key figure in the second phase of their career being Californian stoner-rock pioneer Josh Homme. And yet it was under the production aegis of the Queens of the Stone Age frontman that they performed a stylistic handbrake turn with 2009's *Humbug*, ditching the indie influences of their first two albums for a markedly heavier, darker sound. It also marked the point where Alex Turner’s lyrical
Why Information Extraction?

semi-structured information

Arctic Monkeys
The Observer

Phil Mongredien
Sunday 8 September 2013 00.05 BST

W
hen Arctic Monkeys burst into the spotlight in 2006, breaking sales records, prompting bandwagon-jumping prospective Labour party leaders to make themselves sound heroically foolish and being hailed for their quintessential Britishness, you could have got long odds on the key figure in the second phase of their career being Californian stoner-rock pioneer Josh Homme. And yet it was under the production aegis of the Queens of the Stone Age frontman that they performed a stylistic handbrake turn with 2009's Humbug, ditching the indie influences of their first two albums for a markedly heavier, darker sound. It also marked the point where Alex Turner's lyrical

How to link this information to a knowledge base automatically?
Why Information Extraction?

semi-structured information

Arctic Monkeys
The Observer

Phil Mongredien
Sunday 8 September 2013 00.05 BST

W
hen Arctic Monkeys burst into the spotlight in 2006, breaking sales records, prompting bandwagon-jumping prospective Labour party leaders to make themselves sound heroically foolish and being hailed for their quintessential Britishness, you could have got long odds on the key figure in the second phase of their career being Californian stoner-rock pioneer Josh Homme. And yet it was under the production aegis of the Queens of the Stone Age frontman that they performed a stylistic handbrake turn with 2009’s *Humbug*, ditching the indie influences of their first two albums for a markedly heavier, darker sound. It also marked the point where Alex Turner’s lyrical

How to link this information to a knowledge base automatically?

Information Extraction!
The Arctic Monkeys almost exclusively played songs from their new album AM at Summerfest 2014 at Miller Lite Oasis in Milwaukee on 25 June 2014.
The Arctic Monkeys almost exclusively played songs from their new album AM at Summerfest 2014 at Miller Lite Oasis in Milwaukee on 25 June 2014.

Named Entity Recognition
The Arctic Monkeys almost exclusively played songs from their new album AM at Summerfest 2014 at Miller Lite Oasis in Milwaukee on 25 June 2014.

Named Entity Recognition

Named Entity Classification (NEC):

Arctic Monkeys: mo:MusicArtist
AM: mo:SignalGroup
Summerfest 2014: mo:Festival
Miller Lite Oases: geo:SpatialThing
Milwaukee: geo:SpatialThing
The **Arctic Monkeys** almost exclusively played songs from their new album **AM** at **Summerfest 2014** at **Miller Lite Oasis** in **Milwaukee** on 25 June 2014.

Named Entity Recognition

**Named Entity Classification (NEC):** **Named Entity Linking (NEL):**

**Arctic Monkeys**: mo:MusicArtist

**AM**: mo:SignalGroup

**Summerfest 2014**: mo:Festival

**Miller Lite Oases**: geo:SpatialThing

**Milwaukee**: geo:SpatialThing

**Arctic Monkeys**: mo:artist/ada7a83

**AM**: mo:release-group/a348ba2f-f8b3

**Summerfest 2014**: mo:event/3fc3

**Miller Lite Oases**: mo:place/3f26acf

**Milwaukee**: mo:area/4dc3fa97-cf9b
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, MISC or fine-grained classes such as SIGNAL GROUP

Named Entity Linking / Disambiguation: Linking NEs to concrete entries in knowledge base, example:
Milwaukee -> LOC: largest city in the U.S. state of Wisconsin
  -> LOC: Milwaukee, Oregon, named after the city in Wisconsin
  -> LOC: Milwaukee County, Wisconsin
  -> ORG: Milwaukee Tool Corp, a manufacturer of electric power tools
  -> MISC: early codename for what was to become the Macintosh II
  -> …
The Arctic Monkeys almost exclusively played songs from their new album AM at Summerfest 2014 at Miller Lite Oasis in Milwaukee on 25 June 2014.
The Arctic Monkeys almost exclusively played songs from their new album AM at Summerfest 2014 at Miller Lite Oasis in Milwaukee on 25 June 2014.
The Arctic Monkeys almost exclusively played songs from their new album AM at Summerfest 2014 at Miller Lite Oasis in Milwaukee on 25 June 2014.
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 LOCATED-IN

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. music festival, album release

Event Extraction: Extracting events consisting of the name and type of event, agent, 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

Milwaukee LOC 0.9
Milwaukee LOC 0.8
Milwaukee ORG 0.4
→ Milwaukee LOC
NERC: Methods

• 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

Isabelle Augenstein
Developing a NERC involves programming based around APIs.
Developing a NERC involves programming based around APIs.

A  N  D  Y  E  T  S  O  H  E  L  P  L  E  S  S  L  Y  A  L  O  N  E  
A  S  W  H  E  N  I  G  O  O  G  L  E  A  N  E  R  R  O  R  
A  N  D  T  H  R  E  E’  S  O  N  E  R  E  S  U  L  T  
A  T  H  R  E  A  D  B  Y  S  O  M  E  O  N  E  W  I  T  H  T  H  E  S  A  M  E  P  R  O  B  L  E  M  
A  N  D  N  O  A  N  S  W  E  R  
L  A  S  T  P  O  S  T  E  D  T  O  I  N  2  0  0  3  

W  H  O  W  E  R  E  Y  O  U,  D  E  N  V  E  R  C  O  D  E  R?  
W  H  A  T  D  I  D  Y  O  U  S  E  ?!

which can be frustrating at times
and (at least basic) knowledge about linguistics

TENSE? MOODY? IRREGULAR?

YOU MUST BE A VERB.

Isabelle Augenstein
Sentences

Lexicon

Tokens

Lexicon

Morphemes

Lexicon

Words

Morphology

Syntax

Sentences

Semantics,

Meaning

Discourse

The farmer hit the donkey.

The, farmer, hit, the, donkey, .

wait + ed -> waited, cat -> cat

wait -> V, cat -> N

The (D) farmer (N) hit (V) the (D) donkey (N).

Every farmer who owns a donkey beats it.

∀x ∀y (farmer(x) ∧ donkey(y) ∧ own(x, y) → beat(x, y))
# Background: NLP Tasks

<table>
<thead>
<tr>
<th>Level</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>Sentence splitting</td>
</tr>
<tr>
<td>Tokens</td>
<td>Tokenisation</td>
</tr>
<tr>
<td>Morphemes</td>
<td>Lemmatisation or stemming, part of speech (POS) tagging</td>
</tr>
<tr>
<td>Words</td>
<td>Chunking, parsing</td>
</tr>
<tr>
<td>Sentences</td>
<td>Semantic and discourse analysis, anaphora resolution</td>
</tr>
</tbody>
</table>

**Lexicon**

- Sentences
- Tokens
- Morphemes
- Words
- Sentences
- Meaning

**Morphology**

- Words
- Sentences
- Meaning

**Syntax**

- Tokens
- Morphemes
- Words
- Sentences
- Meaning

**Semantics, Discourse**

- Sentences
- Meaning

Isabelle Augenstein
Background: Linguistics

Isabelle Augenstein
Background: Linguistics

- Sentences
  - Lexicon
  - Tokens
    - Lexicon
    - Morphemes
      - Lexicon
      - Morphology
      - Words
        - Morphology
        - Syntax
        - Sentences
          - Syntax
          - Meaning
            - Semantics, Discourse

- New York-based
  - [New, York-based] or [New, York, -, based]
Sentences

**Lexicon**

**Tokens**

**Lexicon**

**Morphemes**

**Lexicon**

**Words**

**Morphology**

**Syntax**

**Sentences**

**Semantics, Discourse**

**New York-based**

**[New, York-based] or [New, York, -, based]**

**She’d**
Background: Linguistics

Sentences

Tokens

Morphemes

Words

Sentences

Meaning

Lexicon

Lexicon

Lexicon

Morphology

Syntax

Semantics, Discourse

New York-based
[New, York-based] or [New, York, -, based]

She’d -> she would, she had
Background: Linguistics

Sentences

Tokens

Morphemes

Words

Syntax

Sentences

Meaning

Lexicon

New York-based

[New, York-based] or [New, York, -, based]

Lexicon

She’d -> she would, she had

Morphology

Time flies like an arrow

Semantics, Discourse
Background: Linguistics

Sentences
  ↓
Tokens
  ↓
Lexicon
  ↓
New York-based
  ↓
[New, York-based] or [New, York, -, based]
  ↓
She’d -> she would, she had
  ↓
Morphemes
  ↓
Lexicon
  ↓
Time flies(V/N) like(V/P) an arrow
  ↓
Words
  ↓
Lexicon
  ↓
She’d -> she would, she had
  ↓
Morphology
  ↓
Syntax
  ↓
Sentences
  ↓
Meaning
  ↓
Semantics,
  ↓
Discourse
  ↓
Isabelle Augenstein
Sentences
  ↓
Tokens
  ↓
Lexicon
  ↓
Morphemes
  ↓
Lexicon
  ↓
Words
  ↓
Morphology
  ↓
Syntax
  ↓
Sentences
  ↓
Meaning
  ↓
Semantics, Discourse

New York-based
[New, York-based] or [New, York, -, based]

She’d -> she would, she had

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

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

Sentences

Tokens

Morphemes

Words

Syntax

Sentences

Meaning

Lexicon

New York-based
[New, York-based] or [New, York, -, based]

Lexicon

She’d -> she would, she had

Morphology

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

Words

Syntax

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

Semantics, Discourse
Sentences

Tokens

Morphemes

Words

Syntax

Sentences

Meaning

New York-based

[New, York-based] or [New, York, -, based]

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

Isabelle Augenstein

Sentences → Lexicon

Tokens → Lexicon

Morphemes → Morphology

Words → Syntax

Sentences → Semantics, Discourse

New York-based
[New, York-based] or [New, York, -, based]

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. -> Same woman or different women?
### Sentences

<table>
<thead>
<tr>
<th>Levels</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentences</strong></td>
<td>New York-based</td>
</tr>
<tr>
<td></td>
<td>[New, York-based] or [New, York, -, based]</td>
</tr>
<tr>
<td><strong>Tokens</strong></td>
<td>She’d -&gt; she would, she had</td>
</tr>
<tr>
<td><strong>Morphemes</strong></td>
<td>Time flies(V/N) like(V/P) an arrow</td>
</tr>
<tr>
<td><strong>Words</strong></td>
<td>The woman saw the man with the binoculars. -&gt; Who had the binoculars?</td>
</tr>
<tr>
<td><strong>Syntax</strong></td>
<td>Somewhere in Britain, some woman has a child every thirty seconds. -&gt; Same woman or different women?</td>
</tr>
<tr>
<td><strong>Semantics, Discourse</strong></td>
<td></td>
</tr>
</tbody>
</table>

*Isabelle Augenstein*
Sentences

Tokens

Morpheme

Words

Sentences

Meaning

Ambiguities on every level

Y U SO

[New, York, -, based]

She had

/V/P) an arrow

man with the

had the binoculars?

η, some woman

ty seconds.

different women?

Ambiguous?

Isabelle Augenstein
Language is ambiguous.

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

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

Summerfest 2014 took place at Miller Lite Oasis in **Milwaukee** on 25 June 2014.

- w: Milwaukee  w-1: in  w-2: Oasis  w+1: on  w+2: 25
- seq[-]: Oasis in  seq[+]: on 25
- bow: Milwaukee  bow[-]: in  bow[-]: Oasis  bow[+]: on  bow[+]: 25

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

Prokofyev (2014)
### Morphology: Penn Treebank POS tags

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

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

#### Nouns (all start with *N*)

#### Adjectives (all start with *J*)

#### Verbs (all start with *V*)
What can help to recognise and/or classify named entities?

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

Prokofyev (2014)
What can help to recognise and/or classify named entities?

- Gazetteers
  - Retrieved from HTML lists or tables [1]

  **Most top-ten albums**

  Source: [8]

  - The Rolling Stones (36)
  - Frank Sinatra (33)
  - Barbra Streisand (33) [9]
  - The Beatles (30)
  - Elvis Presley (27)

- Using regular expressions patterns and search engines (e.g. “Popular artists such as * ”)

- Retrieved from knowledge bases

NERC: Training Models

Extensive choice of machine learning algorithms for training NERCs
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**)
  
  took(**O**) place(**O**) at(**O**) Miller(**B-LOC**) Lite(**I-LOC**) Oasis(**I-LOC**) in(**O**)  

For now, we will **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**
NLP & ML Software

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, wolfe (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)
Application: Opinion Mining

- Extracting opinions or sentiments in text
- It’s about finding out what people think

service  speedy
stellar  quality
selection  diverse
rockin’ atmosphere
great  value
Application: Opinion Mining

- Opinion Mining is big business
- Someone just bought an album by a music artist
  - Writes a review about it
- Someone else wants to buy an album
  - Looks up reviews by fans and music critics
- Music artist and music producer
  - Get feedback from fans
  - Improve their product
  - Improve their marketing strategy
• “Miley Cyrus's attempts to shock would be more effective if she had songs to back up the posturing.”
  
  – The Guardian ★★★★★

• “Bangerz is an Amazing album with great lyrics and we can see the Miley Cyrus' musical evolution. Would love to buy it and I already did. ALBUM OF THE YEAR. Peace”
  
  – Rodolfoalmeida3 ★★★★★
Why is opinion mining and sentiment analysis challenging?

• Relatively easy to find sentiment words in sentences, difficult to identify which topic they are about
Why is opinion mining and sentiment analysis challenging?

- Relatively easy to find sentiment words in sentences, difficult to identify which topic they are about
  - “The album comes with a free bonus CD but I don't like the cover art much.”

Does this refer to the cover art of the bonus CD or the album?
Why is opinion mining and sentiment analysis challenging?

- Relatively easy to find sentiment words in sentences, difficult to identify which topic they are about
Why is opinion mining and sentiment analysis challenging?

- Relatively easy to find sentiment words in sentences, difficult to identify which topic they are about
- Whitney Houston was quite unpopular…

**Twitter Sentiment**

"Whitney Houston"  
Tweet: 273  
Like: 319  
+1: 20

Sentiment analysis for "Whitney Houston"
Why is opinion mining and sentiment analysis challenging?

- Relatively easy to find sentiment words in sentences, difficult to identify which topic they are about
- Whitney Houston was quite unpopular… or was she?

**Tweets about: "Whitney Houston"**

- *bazzyboy25*: Whitney houston…too soon? #CelebritiesThatLookLikeTheyStank
  - Posted 5 minutes ago
- *TeghanSimone*: Radio playing Whitney Houston.. I swear I’m about to cry… So sad
  - Posted 5 minutes ago
- *JB3LL*: hoes about to get whitney houston’d tonight! #TheWalkingDead
  - Posted 5 minutes ago
- *derickadams*: “@indreamville : Twitter I’m curious who do you think had more problems Michael Jackson or Whitney Houston???”
  - Posted 5 minutes ago

- Death confuses opinion mining tools
Why is opinion mining and sentiment analysis challenging?

- It’s not just about finding sentiment words, context is important too
  - “It's a great movie if you have the taste and sensibilities of a 5-year-old boy.”
  - “It's terrible Candidate X did so well in the debate last night.”
  - “I'd have liked the track a lot more if it had been a bit shorter.”

- If sentiment words are neutral, negative or positive depends on domain
  - “a long track” vs “a long walk” vs “a long battery life”
Why is opinion mining and sentiment analysis challenging?

• How much should every single opinion be worth?
  • experts vs non-experts
  • relationship trust
  • reputation trust
  • spammers
  • frequent vs infrequent posters
  • “experts” in one area may not be expert in another
  • how frequently do other people agree?
Subtopics

• **Opinion extraction**: extract the piece of text which represents the opinion
  • *Cyrus has made a 23-song, purposely strange psych-rock record. Make no mistake, some of this album is unlistenable. But Cyrus is also too skilled of an artist to not place some beauty inside this madness, and Miley Cyrus and Her Dead Petz swerves into thoughtful territory when it’s least expected.*

• **Sentiment classification/orientation**: extract the polarity of the opinion (e.g. positive, negative, neutral, or classify on a numerical scale)
  • negative: purposely strange, some is unlistenable
  • positive: skilled artist, beauty inside madness, thoughtful

• **Opinion summarisation**: summarise the overall opinion about something
  • Strange, some unlistenable: negative, skilled artist, beauty, thoughtful: positive, Overall 6/10
Subtopics

- **Feature-opinion association**: given a text with target features and opinions extracted, decide which opinions comment on which features.
  - “The tracks are *good* but not *so keen* on the cover art”
- **Target identification**: which thing is the opinion referring to?
- **Source identification**: who is holding the opinion?
Opinion Mining Resources

Bing Liu’s English Sentiment Lexicon
• 2006 pos words, 4783 neg words
• Useful properties: includes misspellings, morphological variants, slang
• Available from: http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

The MPQA Subjectivity Lexicon
• Polarities: positive, negative, both, neutral
• Subjectivity: strongsubj or weaksubj
• Download from: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
Opinion Mining Resources

WordNet Affect
- Extension of WordNet with affect words
- Useful properties: includes POS categories

Hands-on session: Applying standard opinion mining lexicons with GATE
- Spoiler: general purpose lexicons do not always perform well, for better performance, domain- or context-specific lexicons are necessary
Thank you for your attention!

(And thank you to Diana Maynard for allowing me to adapt and reuse her Opinion Mining slides!)

Questions?

Isabelle Augenstein