DATA SCIENCE FOR DIGITAL HUMANITIES

INFORMATION EXTRACTION

PROF. DR. GORAN GLAVAŠ

Information Extraction

- **Information extraction (IE)** is the automatic identification of selected types of entities, relations, or events in free text

- Traditionally, IE tasks tasks are the following:
  - Named entity recognition and classification (NERC)
  - Coreference resolution
  - Relation extraction
  - Event extraction

- The following tasks loosely belong to IE:
  - Keywords/keyphrase extraction
  - Terminology extraction
  - Collocation extraction

Named entity recognition

Named Entity Recognition

Eastern Ukraine is gripped by an armed separatist uprising, with pro-Russian protesters occupying government buildings in more than a dozen towns and cities, despite an ongoing ”anti-terror” operation launched by the Ukrainian military. Vyacheslav Ponomaryov is the self-proclaimed pro-Russian mayor of Sloviansk, Donetsk region, the stronghold of the separatist movement in eastern Ukraine. He was involved in the seizure of a group of military observers from the Organization for Security and Co-operation in Europe (OSCE). One of the best-known leaders of the uprising, Igor Strelkov directs armed pro-Russian activists in eastern Ukraine, especially in Sloviansk. The word is he works for the GRU (Russian military intelligence agency), and his real name is Igor Girkin. He was born in 1970 and registered in Moscow.

- PERson, LOCation, ORGanization, TIME

- **Q:** What type of NLP task would NER be (from the machine learning perspective)?
- **A**: Sequence labeling

Rule-Based Named Entity Recognition

- Large number of extraction patterns / rules
- Each pattern detects some type of named entities

```
[capitalized-word]+[Corp.] ⇒ Organization
[' Mr.' ][capitalized-word]+ ⇒ Person
[in|at|on][capitalized-word]+ ⇒ Location
```

- Unfortunately, most rules have exceptions...

"She lost hope she would ever meet *Mr. Right One*." (Person?)
"God only knows what goes on in *Putin*'s mind." (Location?)

Building a Named-Entity Tagger

- We can add additional rules to handle exceptions
- E.g., gazetteers: word lists for each of the NER categories

- Some potential gazetteer rules:

```
[cap-word-names-gazetteer]+[cap-word-surnames-gazetteer]+
```

Personal names: Aaliyah, Aaron, Abbey, ..., Zygmunt, Zyta
Surnames: Abbott, Abney, Abraham, ..., Zysett, Zyskowsky
Organizations: Abbott Laboratories, Abercrombie & Fitch, Association for Computational Linguistics, . . . , WorldCom, World Help Foundation
Locations: Alabama, Arkansas, ..., Zimbabwe

- Problem: Gazetteers are always incomplete
- Generally, too many rules, difficult to maintain, etc.
- For some NE types rules are OK (**Q:** which ones?),
- But generally, it's better to go for **machine learning approaches**

Supervised Named Entity Recognition

- We need: a corpus manually annotated with named entities
- Annotations done according to annotation standard
- The most renowned annotation standard: **MUC-7**

- **MUC-7 named entity types**
  - Entity names (ENAMEX) – Person, Organization, Location
  - Temporal expressions (TIMEX) – Date, Time
  - Quantities (NUMEX) – Monetary value, Percentage

- Annotation of named entities is not particularly demanding
- No need to hire experts (e.g., linguists)
- Virtually any native speaker can annotate (after training)

Supervised Named Entity Recognition

- NER is a prototypical sequence labelling task
  - But named entities are generally multi-token expressions

- **Q:** What labels do we assign to individual tokens?
- We need to make a distinction between the first token of a named entity and all other tokens
- **Q:** Why?

Barcelona's/ORG draw/O with/O Atletico/ORG Madrid/ORG at/O Camp/LOC Nou/LOC was/O not/O expected/O, says/O British/ORG Broadcast/ORG Channel's/ORG La/ORG Liga/ORG football expert Andy/PER West/PER.

- „*British Broadcast Channel's La Liga*" – one or two organizations?

Supervised Named Entity Recognition

- NER is a prototypical sequence labelling task
  - But named entities are generally multi-token expressions

- B-I-O annotation scheme
  - **B** – Begins a named entity (i.e., first NE token)
  - **I** – Inside a named entity (i.e., second and subsequent NE tokens)
  - **O** – Outside of a named entity (i.e., token is not part of any NE)

Barcelona's/B-ORG draw/O with/O Atletico/B-ORG Madrid/I-ORG at/O Camp/B-LOC Nou/I-LOC was/O not/O expected/O, says/O British/B-ORG Broadcast/I-ORG Channel's/I-ORG La/B-ORG Liga/B-ORG football expert Andy/B-PER West/I-PER.

- „*British Broadcast Channel's La Liga*" – two organizations!

Supervised Named Entity Recognition

Supervised approaches to NER:

1. **Token-level classification**
   - Naive Bayes, SVM, Logistic regression, Feed-forward NN
   - Cannot use labels from both token sides as features

2. **Sequence labelling**
   - Hidden Markov Models (HMM), Conditional Random Fields (CRF)
     - Require manual feature design
   - **Deep neural networks (recurrent NNs or Transformers)**
     - Word embeddings as input, no feature design
     - State-of-the-art results

Common features (for feature-based learning algorithms):

- **Linguistic features**: word, lemma, POS-tag, sentence start, capitalization
- **Gazetteer features**: is gazetteer entry, starts gazetteer entry, inside of a gazetteer entry (for all gazetteers)

Named Entity Recognition – Document Level

- Sequence models predict BIO labels at the sentence level
- Thus, it's possible to have different labels for the same named entity at the document level

> Eastern Ukraine is gripped by an armed separatist uprising. Vyacheslav Ponomaryov is the self-proclaimed pro-Russian mayor of Sloviansk, Donetsk region, the stronghold of the separatist movement in eastern Ukraine. He was involved in the seizure of a group of military observers from the Organization for Security and Co-operation in Europe (OSCE). One of the best-known leaders of the uprising, Igor Strelkov directs armed pro-Russian activists in eastern Ukraine, especially in Sloviansk.

- Enforcing document-level consistency improves NER performance
- Approaches:
    - Simple rule-based approach (count-based)
    - Second sequence labelling model

Named Entity Recognition Evaluation

- Comparing system predicted Named Entities (NEs) with gold-annotated Nes

1. **Lenient (aka MUC) evaluation**
   - System NE and gold NE need to be of the same type and overlap in token spans in order to count as a match (i.e., true positive)
2. **Strict (aka Exact) evaluation**
   - System NE and gold NE need to be of the same type and exactly the same token span order to count as a match (i.e., true positive)

Gold: „The Faculty of Business Informatics and Mathematics issued a diploma...”
Sys1: „The Faculty of Business Informatics and Mathematics issued a diploma...”
Sys2: „The Faculty of Business Informatics and Mathematics issued a diploma...”

- State-of-the-art NER performance (coarse-grained entity types) is around **95%** F-score for English, slightly less for other languages

Coreference resolution and entity linking

Coreference Resolution

- Linking entity mentions that refer to the same entity in the real world
- Mentions refering to same real-world entities ⇒ coreferent mentions
- Set of coreferent mentions ⇒ coreference chain

> I no longer see the possibility of continuation of collaboration with General Flynn, said President in his most recent address. That is a political decision I had to make, considering the actions he was involved in. I thank the General for his contributions to the Government, said the Leader of the Free World.

- **Q:** Why do we need this for IE?

Approaches to Coreference Resolution

- Rule-based methods
  - Linguistically-motivated rules
  - Rules based on domain knowledge

- **Machine learning models**
  - Supervised machine learning (classification)
  - Unsupervised machine learning (clustering)
  - Hybrid: classification + clustering

Rule-Based Methods

- Hobbs' algorithm (Hobbs, 1986)
  - Within-sentence pronoun resolution algorithm
- Constraints on pronouns on the constituency syntactic parse of the sentence
  - Heuristics based on centering theory (Grosz et al., 1995)

- Centering theory – inference load on the reader is lower when mentions of the same entities occupy the same grammatical roles

Centering example

1. Johnny really goofs around sometimes. **(sub)**
2. He *was excited* about trying out his new sailboat. **(sub)**
3. He *wanted* Tony to join him on a sailing expedition. **(sub, obj)**
4. He *called* him at 6 AM. **(sub, obj)**
5. He was sick and furious at *being woken up* so early. **(obj)**

Supervised Coreference Resolution

- The first step of coreference resolution is **mention extraction**
- Deciding what is and what isn't a mention of some entity
- Not trivial: pronouns, names (i.e., named entities), nominals, nested NPs

- There are three groups of coreference resolution models:
  1. **Mention-pair models**
     - Classification model that produces pairwise coreference decisions
  2. **Entity-mention models**
     - Clusters are built directly by adding mentions to existing clusters
       - or by starting new clusters
     - Addition decisions are made by the classifier
  3. **Ranking models**
     - Determine which candidate antecedent is most probable with respect to the current mention

Mention-Pair Coreference Resolution

- A classifier that, given a description of two mentions, m_i and m_j, determines whether they are coreferent or not
  - Coreference as a pairwise classification task

- One training instance for each pair of mentions from text annotated with coreference information?

[Mary] said [John] liked [her] because [she]...

Positive instances: Mary – her, Mary – she, her – she
Negative instances: Mary – John, John – her, John – she

- Problem #1: Creating all possible mention pairs creates a huge and heavily skewed dataset (in favor of negative class)
- Solution #1: Heuristics for limiting the number of pairs (distance limit, gender matches, ...)

Mention-Pair Coreference Resolution

- Coreference is a transitive relation
- Problem #2: pairwise predictions (due to being imperfect) may violate transitivity

[Mary] likes [him] but [she]...

Pairwise classification decisions:

1. Mary – him: positive (error)
2. him – she: negative
3. Mary – she: positive; but should be negative by transitivity from 1. and 2.

- Solution #2: inducing mention chains instead of all pairwise classifications

Entity-Mention Coreference Resolution

- A classifier that determines whether (or how likely) a mention belongs to a preceding coreference cluster
    - More expressive than the mention-pair model
- A training instance is a pair of a mention and a preceding cluster

- Can employ cluster-level features defined over any subset of mentions in a preceding cluster
- All-most-none features:
    - Is a mention gender-compatible with all mentions in a preceding cluster?
    - Is a mention gender-compatible with most of the mentions in it?
    - Is a mention gender-compatible with none of them?

Entity Linking

- A task in its essence similar to coreference resolution
- Associating mentions of an entity in text to an entry representing that entity in a knowledge base

Iranian POW negotiator holds talks with Iraqi ministers

The head of **Iran**'s **prisoner of war** commission met with two **Iraqi** Cabinet ministers Saturday in a bid to glean information about thousands of Iranian POWs allegedly in Iraq, the official Iraqi News Agency reported.

Iraqi Foreign Minister **Mohammed Saeed al-Sahhaf** told Abdullah al-Najafi that the two states needed to ``speed up the closure of what remains from the POW and Missing-In-Action file,'' INA said.

The issue of POWs and missing persons remains a stumbling block to normalizing relations between the two neighbors.

Iraq has long maintained that it has released all Iranian prisoners captured in the **1980**-88 **Iran-Iraq War**. The countries accuse each other of hiding POWs and preventing visits by the **International Committee of the Red Cross** to prisoner camps.

The ICRC representative in **Baghdad**, Manuel Bessler, told The **Associated Press** that his organization has had difficulty visiting POWs on both sides on a regular basis.

In April, Iran released 5,584 fied as civil law detainees in the largest exchange since **1990**.

More than 1 million people w

Baghdad

Baghdad is the capital of Iraq and of Baghdad Governorate. With a metropolitan area estimated at a population of 7,000,000, it is the largest city in Iraq. It is the second-largest city in the Arab world (after Cairo) and the second-largest city in southwest Asia (after Tehran).

open in wikipedia

Entity Linking

- "James Cook" in **Wikipedia**
  - 4 different organizations (an University, an Institute, . . .)
  - 11 different people (British explorer, NFL player, . . .)
- In a way, a task dual to coreference resolution
  - Coref. resolution captures different mentions of the same entity
  - Entity linking often needs to disambiguate between different entities with the same surface form in text
- Entity linking is essentially an information retrieval (ranking) task
  - An entity mention (with the originating document) is a query
  - Knowledge base articles constitute the document collection
- Usually the context (multiple entity mentions) is jointly resolved
- „James Cook" mentioned with „NFL" is most likely the American football player

Relation extraction

Relation Extraction

- **Relation Extraction** refers to a recognition of an assertion of a particular relationship between two or more entities in text

> *„Located in Pittsburgh, Carnegie-Mellon University is one of the leading U.S. technical universities alongside Stanford, Berkeley, and MIT"*
>
> R1: **Located-in** (Carnegie-Mellon University, Pittsburgh)
> R2: **Peers** (Carnegie-Mellon University, Stanford, Berkeley, MIT)

- Applications require an understanding of semantic relations between entities
- Question answering, knowledge discovery, logical inference, ...

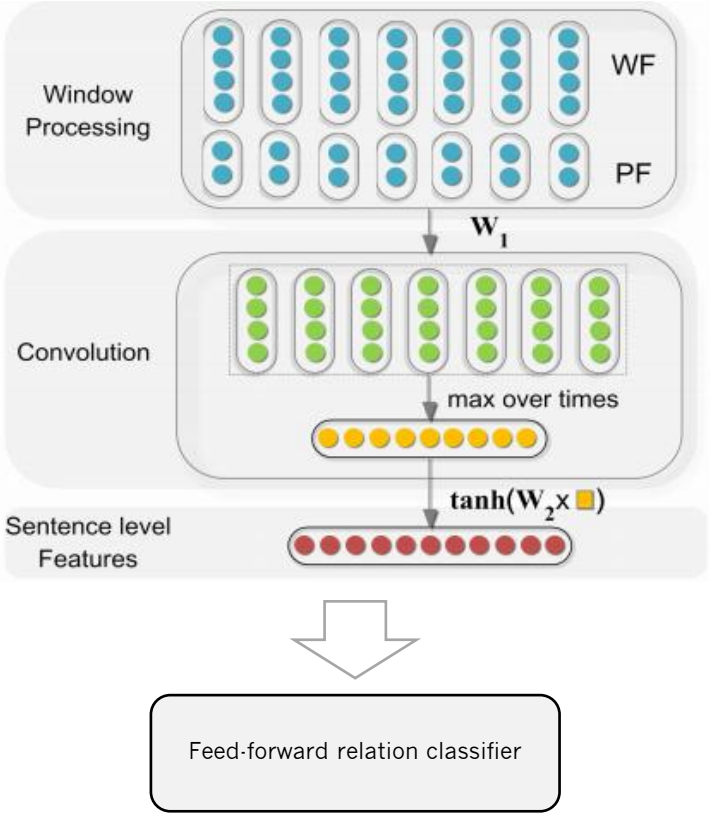
- Relation extraction is crucial in mining biomedical texts
  - Entities are chemical compounds (e.g., proteins, aminoacids, genes)
  - Relations are interactions between proteins
    - *„Gene X with mutation Y leads to malignancy Z"* ➔ *malig-mut(X, Y, Z)*

Relation Extraction Approaches

- **Supervised relation extraction**
  - Relation extraction as a classification task
  - Approaches: feature-based, kernel-based, deep learning-based
- **Semi-supervised approaches**
  - Bootstrapping
  - DIPRE, Snowball, TextRunner
- **Higher-order relation extraction**
  - More than two entities involved

A. Supervised Relation Extraction – Example

- As usual, DL methods remove the need for manual feature design
  - CNN-based approach to RE (Zeng et al., 2014)



- Each position in the sequence is described with
  - Word vectors (embeddings)
  - Positional vectors
    - Encoding relative distances to candidate phrases
- Example:
  - „*Some [people] have been moving into [downtown]*"
  - Positional features:
    - Some: [-1, -6],
    - people: [0, -5],
    - ...
    - into: [4, -1],
    - Downtown: [5, 0]

Open Relation Extraction

- **Open information extraction: set of relations not predefined**

- **RelFinder** (exploration of a „knowledge graph” extracted from text)
  http://www.visualdataweb.org/relfinder/relfinder.php