AI6122 Text Data Management & Analysis

Topic 9: Information Extraction

Topics to be covered

- Information Extraction
  - Named entity recognition
  - Relation extraction
- Named entity recognition
  - The task as sequence classification
  - Features
- Relation extraction
  - Relation extraction via supervised learning
  - Alternative approaches

Information Extraction

- Information Extraction (IE) refers to techniques for extracting limited kinds of semantic content from text.
  - Turns the unstructured information embedded in texts into structured data,
  - An important step in creating or populating knowledge bases

- Named Entity Recognition
  - Finding the proper names or named entities in a text. Examples are person names, company names, locations.
  - **Entity linking**: link an extracted NE to a real world entity. Victoria may refer to different real-world entities

- Relation Extraction
  - Finding and classifying semantic relations among the text entities.
  - These are often binary relations like child-of, employment, part-whole, and geospatial relations.

Named Entity Recognition

- Named Entity Recognition (NER)
  - The first step in most IE tasks
  - The task of NER is is to find each **mention** of a named entity in the text and **label its type**.
  - Typical NE types:
    - 3 class: Location, Person, Organization
    - 4 class: Location, Person, Organization, Misc
    - 7 class: Location, Person, Organization, Money, Percent, Date, Time
  - Many pre-trained models are available for above NE types

- Real applications
  - NEs can be products, medication, financial asset, protein names, or any other domain-specific names

Date and time

- **Temporal expression**: to figure out the time and date mentioned in text
  - Days of the week, Friday or Wednesday
  - Relative expressions: two days from now, next year, a while later
  - Time in different format: 2:20PM, 14:20, around 2 o'clock in the afternoon

- Temporal expression normalization
  - Recognized temporal expression normalized to specific calendar dates and times of day. Example: tomorrow → 29 March 2020
  -
- Many tools are based on regular expressions or rules
  - https://github.com/HeidelTime/heideltime
  - SUTime https://nlp.stanford.edu/projects/time.shtml

Named entity recognition

- A named named entity entity is, roughly speaking, anything that can be referred to with a proper name: a person, a location, an organization.
  - Example text below contains 13 mentions of named entities including 5 organizations, 4 locations, 2 times, 1 person, and 1 mention of money.

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

More examples

- Named entities are useful for many other language processing tasks
  - sentiment analysis: a consumer's sentiment toward a particular product (or named entity)
  - Entities are a useful first stage in question answering
  - For linking text to information in structured knowledge sources like Wikipedia

- Example: a list of generic named entity types with example entity instances

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Named Entity Recognition

- Two main tasks
  - Finding the spans of text that constitute proper names (also known as entity mention detection), and
  - Classifying the type of the entity.

- Challenges in the two tasks
  - Ambiguity of segmentation: what's an entity and what isn't, and where the boundaries are, e.g., "Bishan" vs "Bishan GRC"
  - Type ambiguity: what's an entity's type

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Vehicle
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

[PER Washington] was born into slavery on the farm of James Burroughs.
[ORG Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [LOC Washington] for what may well be his last state visit.
In June, [GPE Washington] passed a primary seatbelt law.
The [VEH Washington] had proved to be a leaky ship, every passage I made...

NER as Sequence Labelling

- The standard algorithm for named entity recognition is as a word-by-word sequence labeling task
  - The assigned tags capture both the boundary and the type.
  - A sequence classifier then trained to label the tokens in a text with tags that indicate the presence of particular kinds of named entities.

- Sequence classifiers
  - Hidden Markov Model (HMM)
  - Maximum Entropy Markov Models (MEMM), aka. conditional Markov model (CMM)
  - Conditional Random Field (CRF)
  - More recently bi-LSTM, transformer, and other deep learning models

  - MEMM, CRF models are able to incorporate domain-specific features to represent observations, e.g. a person name often has its first letter in uppercase.

Labelling Scheme: BIO, IO, BIOLU

[ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said.

Words	IOB Label	IO Label
American	B-ORG	I-ORG
Airlines	I-ORG	I-ORG
,	O	O
a	O	O
unit	O	O
of	O	O
AMR	B-ORG	I-ORG
Corp.	I-ORG	I-ORG
,	O	O
immediately	O	O
matched	O	O
the	O	O
move	O	O
,	O	O
spokesman	O	O
Tim	B-PER	I-PER
Wagner	I-PER	I-PER
said	O	O
.	O	O

- IOB
  - **B** for the first token of a NE, **I** for tokens inside NE's, **O** for tokens outside any NE.
- IO
  - **I** for tokens inside NE's, **O** for tokens outside any NE.
- BIOLU
  - **B** for the first token of a NE, **I** for tokens inside NE's, **L** for the last token of a NE, **O** for tokens outside any NE, **U** for unit length NE, i.e., an NE has only one word.

What features to represent an observation (a word)

- Typical features used for NER

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i is all upper case
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
presence of hyphen

Word shape features

- Word shape features are used to represent the abstract letter pattern of the word by mapping
  - lower-case letters to ‘x’, upper-case to ‘X’, numbers to ’d’, and
  - retaining punctuation.

- Examples
  - I.M.F maps to X.X.X DC10-30 maps to XXdd-dd
  - A second class of shorter word shape features is to remove consecutive character types, so DC10-30 would be mapped to Xd-d but I.M.F would still map to X.X.X.

- This feature by itself accounts for a considerable part of the success of feature-based NER systems for English news text.

- Shape features are also particularly important in recognizing names of proteins and genes in biological texts.

Gazetteer

- A gazetteer is a list of place names, often providing millions of entries for locations with detailed geographical and political information.
  - A related resource is name-lists, e.g., lists of first names and surnames derived from external resources
  - Similar lists of corporations, commercial products, and all manner of things biological and mineral are also available from a variety of sources.

- Gazetteer and name features are typically implemented as a binary feature for each name list.
  - Such lists can be difficult to create and maintain, and their usefulness varies considerably.
  - While gazetteers can be quite effective, lists of persons and organizations are not always helpful (Mikheev et al., 1999).

Feature effectiveness and Evaluation

- Feature effectiveness depends on the application, genre, media, and language.
  - Shape features, critical for English newswire texts, are of little use with automatic speech recognition transcripts, or other non-edited or informally edited sources, e.g., social media.
  - Languages like Chinese don’t use orthographic case.

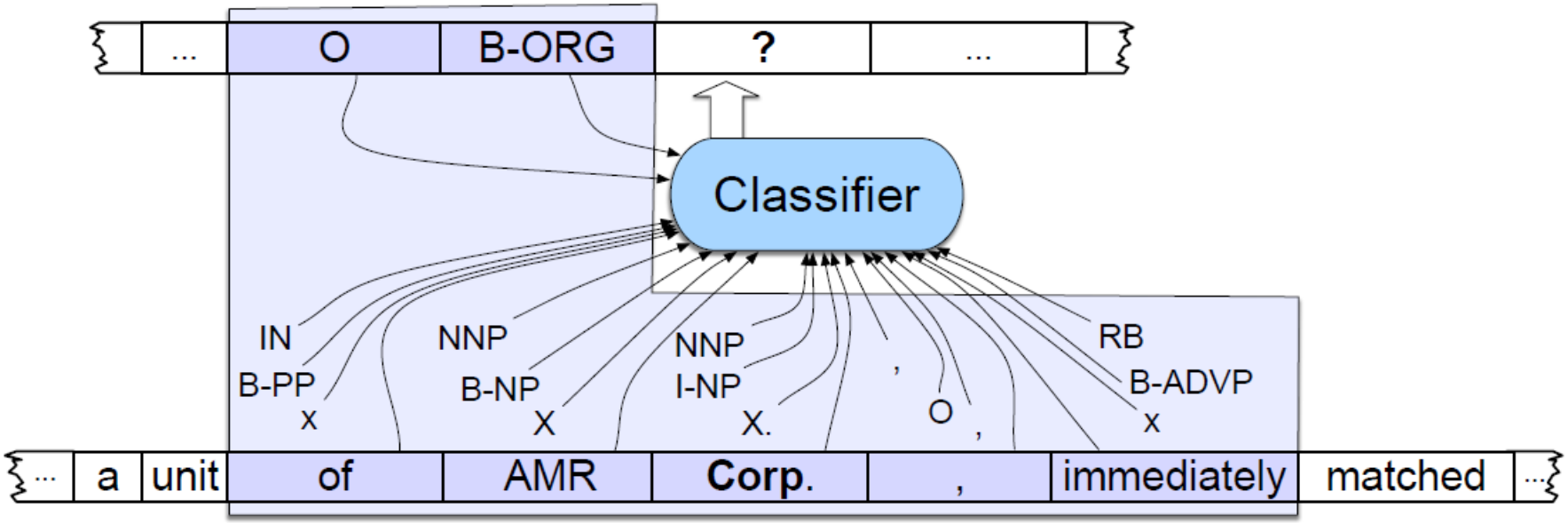
- NER (before deep learning) is heavily based on feature engineering

- Evaluation
  - Precision, Recall, F1

Example to predict the label of “Corp.”

Word	POS	Chunk	Short shape	Label
American	NNP	B-NP	Xx	B-ORG
Airlines	NNPS	I-NP	Xx	I-ORG
,	,	O	,	O
a	DT	B-NP	x	O
unit	NN	I-NP	x	O
of	IN	B-PP	x	O
AMR	NNP	B-NP	X	B-ORG
Corp.	NNP	I-NP	Xx.	I-ORG
,	,	O	,	O
immediately	RB	B-ADVP	x	O
matched	VBD	B-VP	x	O
the	DT	B-NP	x	O
move	NN	I-NP	x	O
,	,	O	,	O
spokesman	NN	B-NP	x	O
Tim	NNP	I-NP	Xx	B-PER
Wagner	NNP	I-NP	Xx	I-PER
said	VBD	B-VP	x	O
.	,	O	.	O

- Feature set in this example
  - POS tags
  - syntactic base-phrase chunk tags
  - (short) word shapes



NER in practical systems

- Commercial approaches to NER are often based on pragmatic combinations of lists and rules, with some smaller amount of supervised machine learning
  - First, use high-precision rules to tag unambiguous entity mentions.
  - Then, search for substring matches of the previously detected names.
  - Consult application-specific name lists to identify likely name entity mentions from the given domain.
  - Finally, apply probabilistic sequence labeling techniques that make use of the tags from previous stages as additional features.

- Some of the entity mentions in a text will be more clearly indicative of a given entity's class than others.
- Once an unambiguous entity mention is introduced into a text, it is likely that subsequent shortened versions will refer to the same entity

Relation Extraction

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

- Relations mentioned in this example
  - Tim Wagner is a spokesman for American Airlines
  - United is a unit of UAL Corp.
  - American is a unit of AMR.

Relation Extraction by Rules

- The earliest and still common algorithm for relation extraction is lexico-syntactic patterns, first developed by Hearst (1992a).

- Example pattern: $NP_0\ such\ as\ NP_1\ \{, NP_2, \ldots, (and|or)\ NP_i\}, i \geq 1$ implies that $\forall NP_i, i \geq 1, hyponym\ (NP_i, NP_0)$
  - "Agar is a substance prepared from a mixture of **red algae**, such as **Gelidium**, for laboratory or industrial use" implies that: Gelidium is a kind of (a hyponym of) red algae.

NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, and other important civic buildings
NP_H such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP_H as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP_H {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP_H {,} especially {NP}* {(or|and)} NP	European countries, especially France, England, and Spain

Relation Extraction via Supervised Learning

- A **fixed set of relations** and **entities** is chosen, a **training corpus** is hand-annotated with the relations and entities, and the annotated texts are then used to train classifiers to annotate an unseen test set.
  - To find pairs of named entities (usually in the same sentence).
  - A filtering classifier is trained to make a binary decision as to whether a given pair of named entities are related (by any relation).
    - Positive examples are extracted directly from all relations in the annotated corpus, and negative examples are generated from within-sentence entity pairs that are not annotated with a relation.
    - The use of the filtering classifier speeds up the final classification and also allows the use of distinct feature-sets appropriate for each task.
  - A classifier is trained to assign a label to the relations that were found by the filtering classifier.

Features for relation classification

- Example: **American Airlines**, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said.
  - Mention 1 (M1): American Airlines
  - Mention 2 (M2): Tim Wagner

- Word Features
  - The headwords of M1 and M2 and their concatenation
    Airlines Wagner Airlines-Wagner
  - Bag-of-words and bigrams in M1 and M2
    American, Airlines, Tim, Wagner, American Airlines, Tim Wagner
  - Words or bigrams in particular positions
    M2: -1 spokesman
    M2: +1 said
  - Bag of words or bigrams between M1 and M2:
    a, AMR, of, immediately, matched, move, spokesman, the, unit
  - Stemmed versions of the same

Features for relation classification

- Example: **American Airlines**, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said.
  - Mention 1 (M1): American Airlines
  - Mention 2 (M2): Tim Wagner

- The syntactic structure of a sentence can also signal relationships among its entities.
  - Syntax is often featured by using strings representing syntactic paths: the (dependency or constituency) path traversed through the tree in getting from one entity to the other
    - Base syntactic chunk sequence from M1 to M2
      NP NP PP VP NP NP
    - Constituent paths between M1 and M2
      $NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
    - Dependency-tree paths
      $Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Relation Extraction: further readings

- Semi-supervised Relation Extraction via Bootstrapping
  - Suppose we have a few high-precision seed patterns or perhaps a few seed tuples. Bootstrapping proceeds by taking the entities in the seed pair, and then finding sentences (on the web, or whatever dataset we are using) that contain both entities. From all such sentences, we extract and generalize the context around the entities to learn new patterns

- Distant Supervision for Relation Extraction
  - Instead of just a handful of seeds, distant supervision uses a large database to acquire a huge number of seed examples, creates lots of noisy pattern features from all these examples and then combines them in a supervised classifier.

- Unsupervised Relation Extraction
  - Unsupervised relation extraction is to extract relations from the web when we have no labeled training data, and not even any list of relations. This task is often called open information extraction or Open IE.
  - In Open IE, the relations are simply strings of words (usually beginning with a verb).

Summary

- Information Extraction
  - Named entity recognition
  - Relation extraction
- Named entity recognition
  - The task as sequence classification
  - Features
- Relation extraction
  - Relation extraction via supervised learning
  - Alternative approaches