Information Extraction and Named Entity Recognition

Dr. Demetrios Glinos
University of Central Florida

CAP6640 – Computer Understanding of Natural Language

Today

- Information Extraction
- Named Entity Recognition
- Sequence Models for NER
- Evaluation of NER Systems

Information Extraction

- Information extraction (IE) systems
  - Find and extract limited kinds of semantic content from text
  - Can gather information from multiple document sources and types
  - Produce structured representations of the relevant information
    - as relations (in the database sense)
    - as a knowledge base for mining and/or other downstream processing
- IE systems turn the unstructured information that is embedded in free text into structured data

What Information is Extracted

- Basic factual information
  - who did what, to whom, and when
- Example:
  - gathering information about corporate mergers from financial news wires

    "East Rock Partners said it proposed to acquire A. P. Green Industries, Inc. for $40 a share."

➔ *merger_parties("East Rock Partners", "A . P. Green Industries, Inc.")*

  - learning drug-gene product interactions from medical research literature

Information Extraction Tasks

- Named entity recognition (NER)
  - finding and detecting all the proper names in the text
  - including reference resolution (e.g., "United Airlines" and "United")
- Relation detection and classification
  - finding and classifying important relations among named entities
- Event detection and classification
  - temporal expressions such as "yesterday" and "next Thursday"
  - mapping temporal expressions onto specific calendar dates and times
- Template filling
  - identifying fixed categories of interest
  - e.g., for merger: parties, price, underwriter

Today

- Information Extraction
- Named Entity Recognition
- Features for NER
- Sequence Models for NER
- Evaluation of NER Systems

Named Entity Recognition (NER)

- Named entity
  - anything that can be referred to with a proper name
- Named entity recognition
  - finding the spans of text that constitute proper names
  - and classifying the mentions according to their type

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles

source: J&M,, Fig. 22.1

Example: NER

Step 1: Finding the named entity mentions:

"Tim Wagner, spokesman for United Airlines, a unit of UAL Corp., said the price increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco."

Example: NER

Step 2: Classifying the named entity mentions:

"Tim Wagner, spokesman for United Airlines, a unit of UAL Corp., said the price increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco."

Legend:	Person
	Organization
	Date
	Location

Named Entity Uses

- How named entities may be used
  - can associate sentiment with named entities
  - many IE relations are associations between named entities
  - can use named entities for indexing the context and/or the documents in which they appear
  - can use named entities for linking to additional information about them

- Example
  - Web page links on named entities to biographical or topic pages, etc.

Ambiguity in Named Entity Recognition

- Same name can refer to two entities of the same type
  - e.g., "JFK" can refer to the former president or his son
- Same name can refer to two entities of different types
  - e.g., "JFK" can refer to the former president or to a major airport in New York City
- Some ambiguities can be completely coincidental
  - e.g., "IRA" can refer to an individual retirement account or to the International Reading Association, or to the Irish Republican Army
- Some ambiguities are the result of linguistic metonymy
  - e.g., "Washington" for the U.S. government
  - e.g., "suit" for "business executive"

Reference Resolution

- Different names for the same named entity
  - International Business Machines, Corp.
  - IBM, Corp.
  - IBM
- Anaphoric references must also be resolved
  - "Peter Townsend, spokesperson for IBM, said the company is financially sound. He later reiterated that position in a press release."
    - "the company" refers to IBM
    - "he" refers to Peter Townsend

Today

- Information Extraction
- Named Entity Recognition
- Features for NER
- Sequence Models for NER
- Evaluation of NER Systems

The Named Entity Recognition Task

- Assigning tags to tokens
  - identify type of named entity
  - identify boundaries of entity
- Common to use "IOB" style, with type indicator
  - "B" indicates the start of the named entity span
  - "I" indicates the token is within the span
  - "O" means token is not within any named entity span
- The **complete span** for a named entity runs from its "B" to the last consecutive "I"

Words	Label
American	B_{ORG}
Airlines	I_{ORG}
,	O
a	O
unit	O
of	O
AMR	B_{ORG}
Corp.	I_{ORG}
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B_{PERS}
Wagner	I_{PERS}
said	O
.	O

source: J&M,, Fig. 22.5

Entity class encoding schemes

	IO Encoding	IOB Encoding
Peter	PER	B-PER
showed	O	O
Ann	PER	B-PER
Juwei	PER	B-PER
Wang	PER	I-PER
's	O	O
new	O	O
painting	O	O

Features for Sequence Labeling

- Good features
  - plausible predictors of the class label
  - easily and reliably extractable from the source text

- Typical features

Feature	Explanation
Lexical items	The token to be labeled
Stemmed lexical items	Stemmed version of the target token
Shape	The orthographic pattern of the target word
Character affixes	Character-level affixes of the target and surrounding words
Part of speech	Part of speech of the word
Syntactic chunk labels	Base-phrase chunk label
Gazetteer or name list	Presence of the word in one or more named entity lists
Predictive token(s)	Presence of predictive words in surrounding text
Bag of words/Bag of N-grams	Words and/or *N*-grams occurring in the surrounding context

source: J&M,, Fig. 22.6

Features: Word Shapes

- Word shape features
  - Mappings to simplified representations that encode key attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.
- Shape templates
  - e.g., "xXXX" finds "mRNA"
  - e.g., "Xx-xxx" matches "Varicella-zoster"
- Other typical shape features

Shape	Example
Lower	cummings
Capitalized	Washington
All caps	IRA
Mixed case	eBay
Capitalized character with period	H.
Ends in digit	A9
Contains hyphen	H-P

source: J&M,, Fig. 22.7

Features: Predictive Words and Gazeteers

- Predictive words
  - words that serve as predictors for surrounding text
  - examples:
    - "Dr.", "MD", "Col.", etc. for persons
    - "Inc.", "Corp.", "LLP", etc. for organizations
- Gazeteers
  - detailed lists that contain millions of entries for many types of locations, along with geographical, geologic, and political information
  - similar lists of persons and organizations are not nearly as useful

Feature Encoding for NER

- Features are usually encoded as additional columns in IOB-encoded data

Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
a	DT	B_{NP}	lower	O
unit	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	B_{ORG}
Corp.	NNP	I_{NP}	cap_punc	I_{ORG}
,	PUNC	O	punc	O
immediately	RB	B_{ADVP}	lower	O
matched	VBD	B_{VP}	lower	O
the	DT	B_{NP}	lower	O
move	NN	I_{NP}	lower	O
,	PUNC	O	punc	O
spokesman	NN	B_{NP}	lower	O
Tim	NNP	I_{NP}	cap	B_{PER}
Wagner	NNP	I_{NP}	cap	I_{PER}
said	VBD	B_{VP}	lower	O
.	PUNC	O	punc	O

Today

- Information Extraction
- Named Entity Recognition
- Features for NER
- Sequence Models for NER
- Evaluation of NER Systems

Sequence Labeling

- We can view many NLP problems as not just labeling, but ***sequence*** labeling
  - Word segmentation
    - determining word boundaries in Chinese
  - POS tagging
    - tagging each token with a part of speech
  - Named entity recognition
    - tagging each token with its entity IOB tag
  - Text segmentation
    - splitting text into alternating questions and answers
    - identifying speakers in spoken text

ML Approach for NER

- Training

    1. Collect a representative set of training documents
    2. Label each token with its entity class or "O" ("outside")
    3. Design feature extractors appropriate to the text and classes
    4. Train a sequence classifier to predict the labels from the data

- Testing

    1. Given a set of test documents
    2. Run the trained sequence model to infer entity labels
    3. Identify entity spans and output the recognized entities

Example: POS Tagging

- Suppose we wish to apply a POS tag to the word "race" in this sentence:

  Secretariat/NNP is/VBZ expected/VBN to/TO race/??? tomorrow/???

- How we do this depends on
  - what we know
  - when we know it
  - what are the possibilities
- How we use the available information depends on the strategy that we employ
- These considerations apply to sequence labeling tasks generally, including NER

Strategy #1: Greedy

Secretariat/NNP is/VBZ expected/VBN to/TO race/??? tomorrow/???

- Greedy strategy
  - start at the left, and use our classifier at each position to assign a label
  - classifier can use
    - entire observed sequence
    - previously assigned labels
- Advantages
  - fast: no extra memory required
  - easy to implement
  - with rich features, including observations to the right, can work well
- Disadvantage
  - Cannot recover from labeling errors

Strategy #2: Beam

Secretariat/NNP is/VBZ expected/VBN to/TO race/??? tomorrow/???

- Beam strategy
  - at each position, keep the top k sequences so far
  - extend each sequence for each possible choice of current label
  - evaluate extensions and keep the k best ones
- Advantages
  - fast: beam sizes of 3-5 are almost as good as exact inference in many cases
  - easy to implement without need for dynamic programming
  - with rich features, including observations to the right, can work well
- Disadvantage
  - optimality not guaranteed: the best sequence can fall off the beam

Strategy #3: Viterbi Inferencing

Secretariat/NNP is/VBZ expected/VBN to/TO race/??? tomorrow/???

- Viterbi strategy
  - dynamic programming solution
  - requires a small window of state influence (e.g., past 2 labels)
  - can use surrounding observations (words), but only prior labels
- Advantages
  - global optimum solution is returned
- Disadvantage
  - difficult to implement long-distance dependencies (e.g., for particles), but no worse than beam strategy

Strategy #4: Conditional Random Fields

Secretariat/NNP is/VBZ expected/VBN to/TO race/??? tomorrow/???

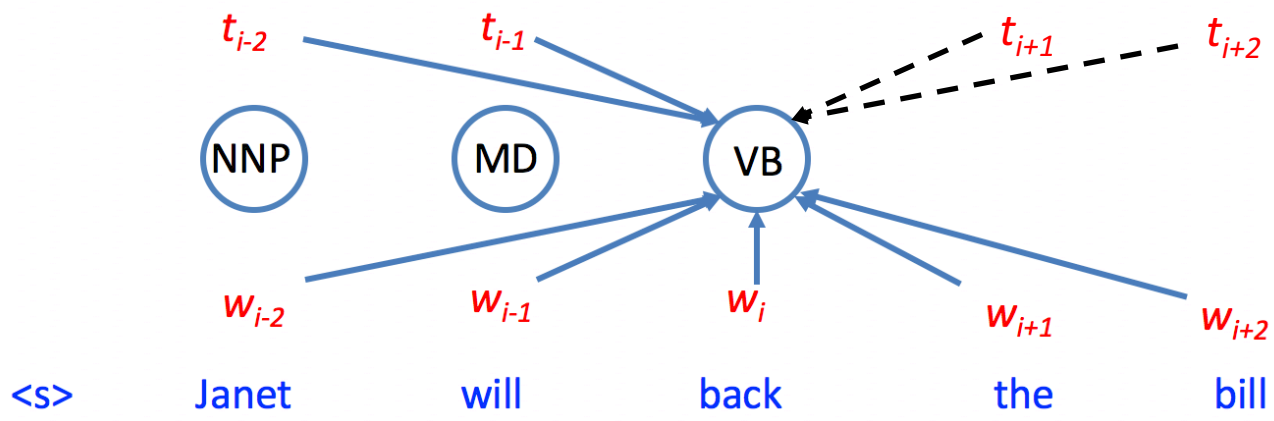
- Conditional Random Field (CRF) strategy
  - a complete-sequence conditional model instead of chaining local models

$$P(c|d,\lambda) = \frac{exp \sum_i \lambda_i f_i(c,d)}{\sum_{c'} exp \sum_i \lambda_i f_i(c',d)}$$

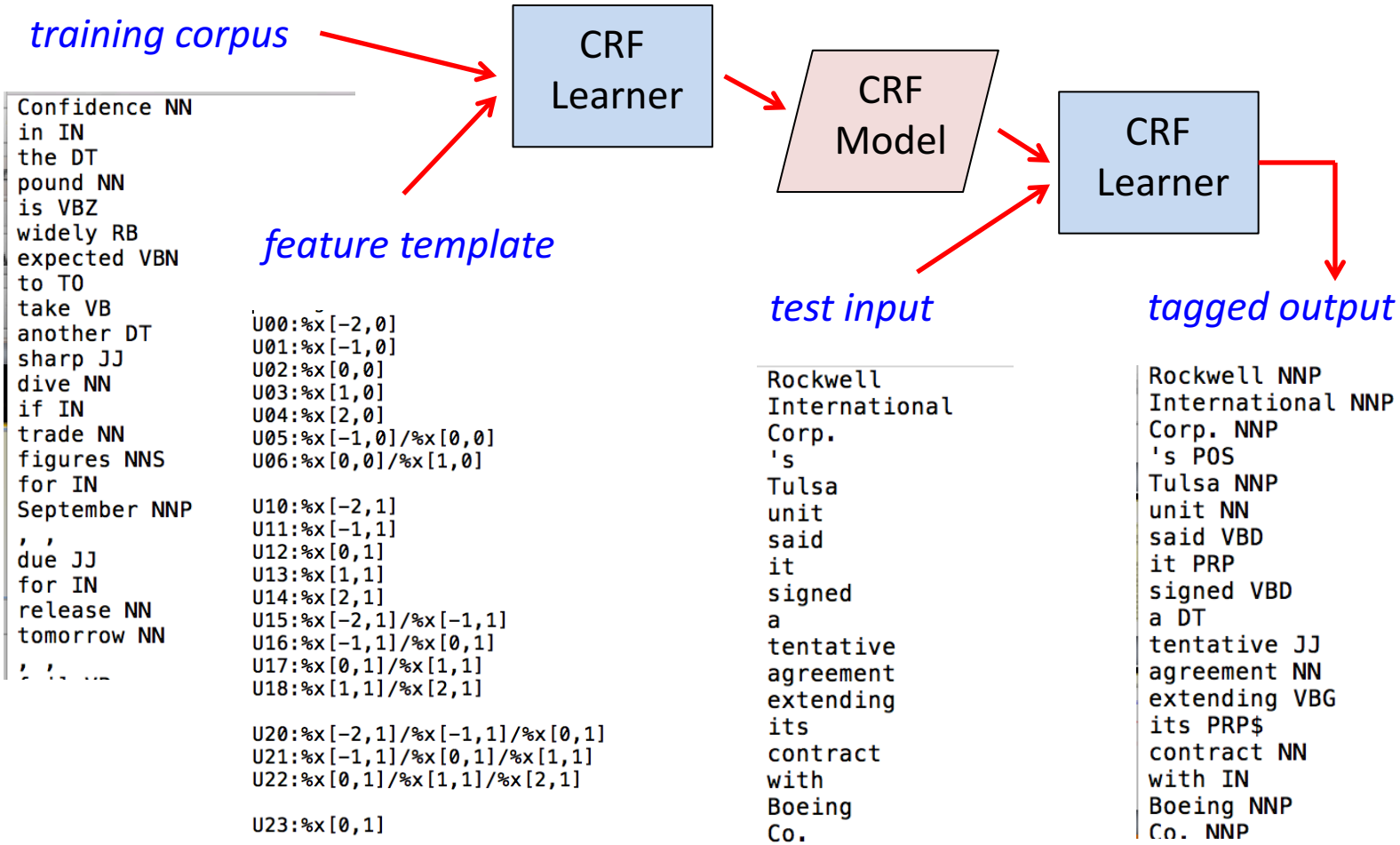
  - i.e., the space of c' is now the space of sequences of labels, not just the space of labels
- Advantage
  - if the features f_i remain local, the conditional sequence likelihood can be calculated exactly using dynamic programming
  - optimal for constrained topologies like linear chain
- Disadvantage
  - slow to train, but considered the state-of-the art in sequence labeling

Feature Templates

- CRF allows feature templates to include *subsequent* tags



POS Learning Using CRF++



CRF References

- Lafferty, J., McCallum, A., Pereira, F. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann. pp. 282–289.

- **CRF++: Yet Another CRF toolkit,** https://taku910.github.io/crfpp/

Today

- Information Extraction
- Named Entity Recognition
- Features for NER
- Sequence Models for NER
- Evaluation of NER Systems

The Named Entity Recognition Task

- Assigning tags to tokens
  - identify type of named entity
  - identify boundaries of entity
- Common to use "IOB" style, with type indicator
  - "B" indicates the start of the named entity span
  - "I" indicates the token is within the span
  - "O" means token is not within any named entity span
- The **complete span** for a named entity runs from its "B" to the last consecutive "I"

Words	Label
American	B_{ORG}
Airlines	I_{ORG}
,	O
a	O
unit	O
of	O
AMR	B_{ORG}
Corp.	I_{ORG}
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B_{PERS}
Wagner	I_{PERS}
said	O
.	O

source: J&M,, Fig. 22.5

Precision/Recall/F1 for IE/NER

- Precision and recall are straightforward at the document level
  - e.g., for IR or text categorization

- However, complications arise when used for named entity spans
  - boundary errors are common
  - example:

First Bank of Scotland announced earnings ...

NER returned

NER should have found

➔ This counts as both a *false positive* and a *false negative* (selecting nothing would have scored better

- Both traditional scoring and adjusted metrics (for partial credit) are used