Announcement

- Homework 6
  - Available in Blackboard -> Homework
  - Due: May. 2, 11:59pm

Information Extraction

SLP Ch 17; INLP Ch 17

Information Extraction: Overview

WASHINGTON/SELMA, Ala. (Reuters) - **Democratic** U.S. presidential front-runner **Bernie Sanders** raised **$46.5 million** in February, his campaign said on Sunday, and will launch new television ad buys in nine states with primaries later this month after this week's Super Tuesday contests. **Joe Biden**'s campaign reported raising $5 million the day of the South Carolina primary. His February haul was **$18 million**, spokesman Michael Gwin said. Meanwhile, rival **Elizabeth Warren**, who struggled to a fifth-place finish in South Carolina, raised more than **$29 million** in February, her campaign manager Roger Lau said in a memo to supporters on Sunday.

Candidate	Party	Feb Fundraising Total
Sanders	D	$46,500,000
Biden	D	$18,000,000
Warren	D	$29,000,000

Named Entity Recognition

Elizabeth Warren, the liberal firebrand who emerged as a top Democratic contender for the **White House** on the strength of an anti-corruption platform backed by a dizzying array of policy proposals, ended her campaign on Thursday. A former bankruptcy law professor who forged a national reputation as a scourge of **Wall Street** even before entering politics, **Warren** had banked on a strong showing on Super Tuesday after a string of disappointing finishes in the early states. But she trailed far behind front-runners **Bernie Sanders** and **Joe Biden**, placing third in her home state of **Massachusetts**, which she continues to represent in the **U.S.** Senate.

- Label certain kinds of proper nouns:
  - Personal names
  - Organizations
  - Geopolitical entities
  - Locations
  - Etc.

Nested Named Entity Recognition

Elizabeth Warren, the liberal firebrand who emerged as a top Democratic contender for the **White House** on the strength of an anti-corruption platform backed by a dizzying array of policy proposals, ended her campaign on Thursday...

… as a top Democratic contender for the White House on …

But she trailed far behind front-runners **Bernie Sanders** and **Joe Biden**, placing third in her home state of **Massachusetts**, which she continues to represent in the **U.S.** Senate.

… in the U.S. Senate.

Entity Linking

Elizabeth Warren, the liberal firebrand who emerged as a top Democratic contender for the White House on the strength of an anti-corruption platform backed by a dizzying array of policy proposals, ended her campaign on Thursday. A former bankruptcy law professor who forged a national reputation as a scourge of Wall Street even before entering politics, **Warren** had banked on a strong showing on Super Tuesday after a string of disappointing finishes in the early states. But she trailed far behind front-runners **Bernie Sanders** and **Joe Biden**, placing third in her home state of Massachusetts, which she continues to represent in the U.S. Senate.

https://www.wikidata.org/wiki/Q434706

Relation Extraction

member_of

WASHINGTON/SELMA, Ala. (Reuters) - **Democratic** U.S. presidential front-runner **Bernie Sanders** raised $46.5 million in February, his campaign said on Sunday, and will launch new television ad buys in nine states with primaries later this month after this week's Super Tuesday contests. **Joe Biden**'s campaign reported raising $5 million the day of the South Carolina primary. His February haul was $18 million, spokesman Michael Gwin said. Meanwhile, rival **Elizabeth Warren**, who struggled to a fifth-place finish in South Carolina, raised more than $29 million in February, her campaign manager Roger Lau said in a memo to supporters on Sunday.

Event Extraction

WASHINGTON/SELMA, Ala. (Reuters) - **Democratic U.S. presidential front-runner Bernie Sanders raised $46.5 million** in **February**, his campaign said on Sunday, and will launch new television ad buys in nine states with primaries later this month after this week's Super Tuesday contests. Joe Biden's campaign reported raising $5 million the day of the South Carolina primary. His February haul was $18 million, spokesman Michael Gwin said. Meanwhile, rival Elizabeth Warren, who struggled to a fifth-place finish in South Carolina, raised more than $29 million in February, her campaign manager Roger Lau said in a memo to supporters on Sunday.

Event	TRANSFER-MONEY
Trigger	raised
Recipient-Arg	… Bernie Sanders
Money-Arg	$46.5 million
Time-Arg	February

Similar to SRL

Methods: NER

- As sequence labeling
  - Use the BIO scheme to represent entity spans and types

…	as	a	top	Democratic	contender	for	the	White	House	on	…
…	O	O	O	O	O	O	O	B-ORG	I-ORG	O	…

Methods: NER

- As sequence labeling
  - Use the BIO scheme to represent entity spans and types
  - Problematic when handling nested entities
    - Ambiguous labeling

…	as	a	top	Democratic	contender	for	the	White	House	on	…
…	O	O	O	O	O	O	O	**B-ORG**	**I-ORG**	O	…
…	O	O	O	B-MISC	I-MISC	I-MISC	I-MISC	**I-MISC**	**I-MISC**	O	…

Methods: NER

- As span classification
  - For each span, predict its entity type (including NONE)

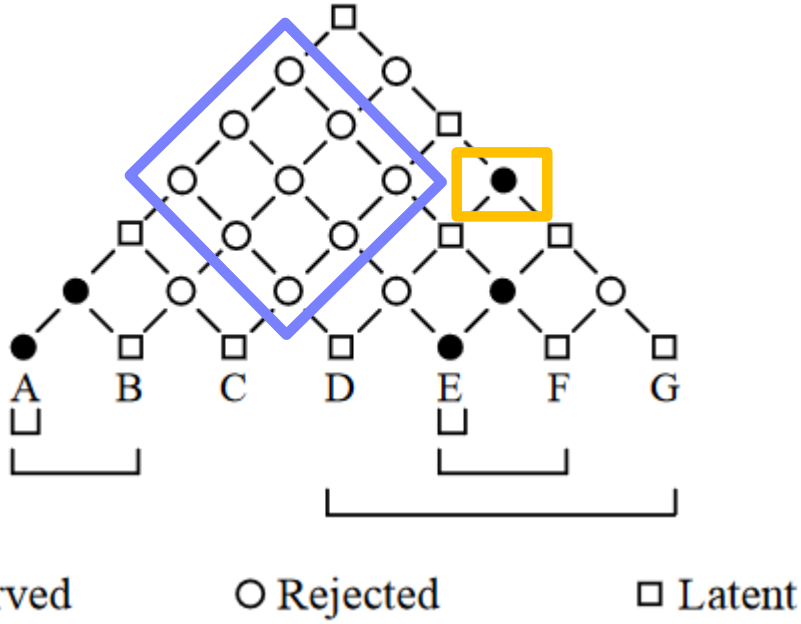
… 0 as 1 a 2 top 3 **Democratic** 4 **contender** 5 **for** 6 **the** 7 **White** 8 **House** 9 on 10 …

	…	6	7	8	9	10	…
…	…	…	…	…	…	…	…
3	…	NONE	NONE	NONE	**MISC**	NONE	…
4	…	NONE	NONE	NONE	NONE	NONE	…
5		NONE	NONE	NONE	NONE	NONE	…
6			NONE	NONE	NONE	NONE	…
7				NONE	**ORG**	NONE	…
8					NONE	NONE	…
…							…

Methods: NER

- As constituency parsing
  - Entities are constituents in a partially-observed constituency parse tree
  - Compared with span classification, here we restrict entities to be non-crossing spans.

These spans cannot be entities because they cross with the yellow entity.



Nested Named Entity Recognition with Partially-Observed TreeCRFs. Fu et al. 2020.

Methods: Relation Extraction

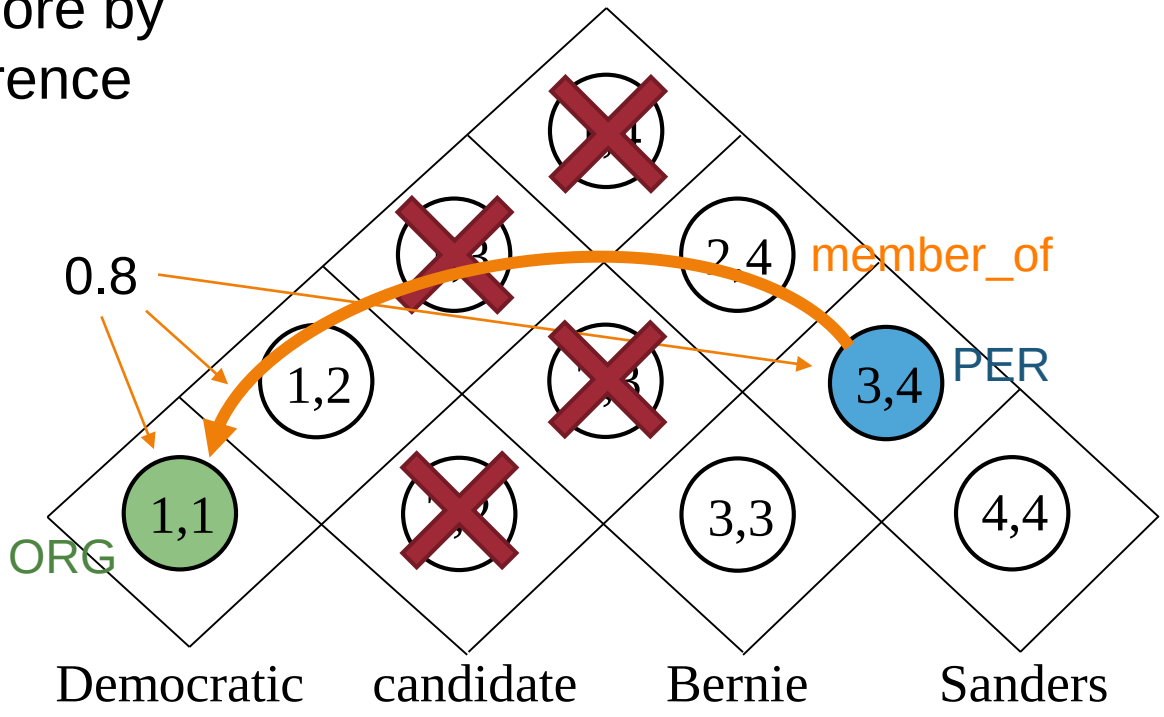
- Given entity spans, predicting relations between them is just like predicting dependency arcs between words
  - Input features
    - Labels of the two entities
    - Text spans of the two entities
    - Text between the two entities
    - Syntactic dependency path between the two entities
  - Output
    - Relation type (including NONE)

Joint Extraction

- Predicting entities and relations jointly, instead of the "entity → relation" pipeline
  - Avoid error propagation
  - Relations may place constraints on entity types
    - Ex: the LIVEIN relation should appear between a PERSON and a LOCATION entity.
- Event extraction is similar
  - Triggers and arguments are like entities
  - Roles are like relations
  - They can be predicted jointly

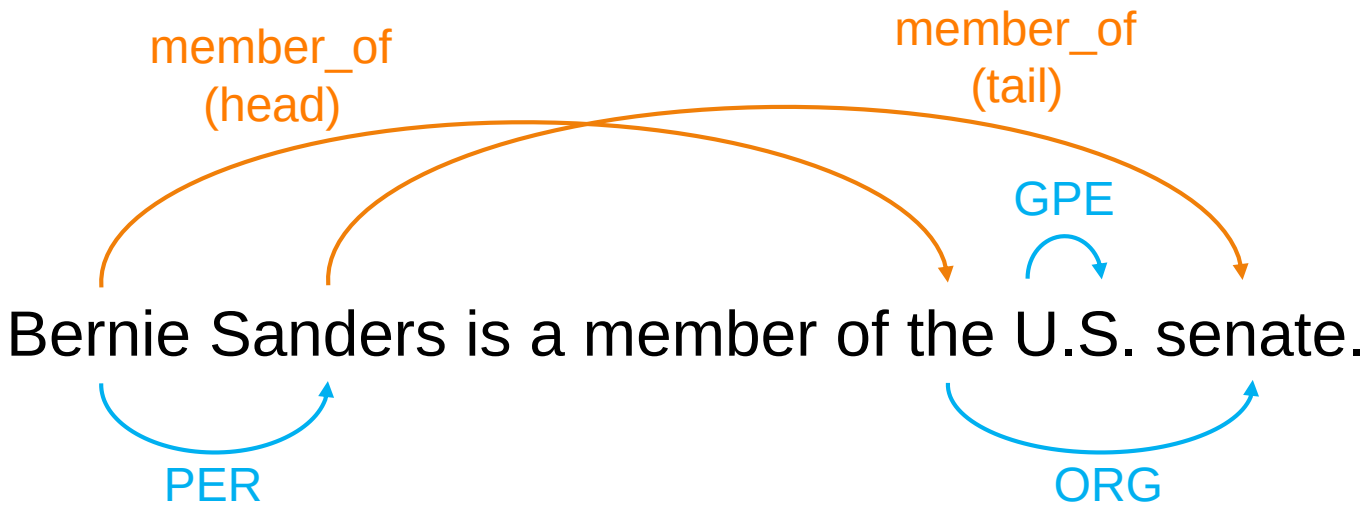
Joint Extraction

- Method 1
  - Score every labeled triple containing two entities and one relation (including null labels)
    - Entity pruning needed to reduce time complexity
  - Maximize total score by approximate inference



Joint Extraction

- Method 2
  - Each entity represented by one arc
  - Each relation represented by two arcs
  - Scoring and inference are similar to high-order dependency parsing



TPLinker: Single-stage Joint Extraction of Entities and Relations Through Token Pair Linking. Wang et al. 2020.

Decoding-based IE methods

- Seq2Seq

Steve became CEO of Apple in 1997.

```
(
   (person: Steve
      (work for: Apple)
   )
   (start-position: became
      (employee: Steve)
      (employer: Apple)
      (time: 1997)
   )
   (organization: Apple)
   (time: 1997)
)
```

Unified Structure Generation for Universal Information Extraction. Lu et al. 2022.

Decoding-based IE methods

- Conversation based (e.g., ChatGPT)
  - NER:
    - Q: The given sentence is “My Love Diary is a TV series released in Beijing in 1990”. Given the list of entity types: Person, Location, what entity types are included in this sentence?
    - A: Location
    - Q: Please identify the entities of type “Location” in the given sentence.
    - A: Beijing
  - The same method can be applied to relation and event extraction.

Zero-Shot Information Extraction via Chatting with ChatGPT. Wei et al. 2023.

Summary

Information Extraction

- Subtasks
  - Named entity recognition
  - Relation extraction
  - Event extraction
  - ...
- Methods
  - Sequence labeling
  - Span/arc classification
  - Joint extraction
  - Decoding based
  - ...