CS689: Computational Linguistics for Indian Languages
Information Extraction

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs689/

2nd semester, 2023-24

Tue 10:30–11:45, Thu 12:00–13:15 at RM101/KD102

Information Extraction

- Extracting structured or semi-structured information from unstructured text is called information extraction (IE)
- Named entity recognition
- Relation extraction
- Event extraction
- Extracted information can be put in a structured form
- Knowledge base (KB)
  - Can include *grammar rules*
- Knowledge graph (KG) stores data as a graph
- Information extraction is useful for
  - *Question-answering (QA)*
  - *Interpretability*
  - *KB-augmented LLMs*

Named Entities (NE)

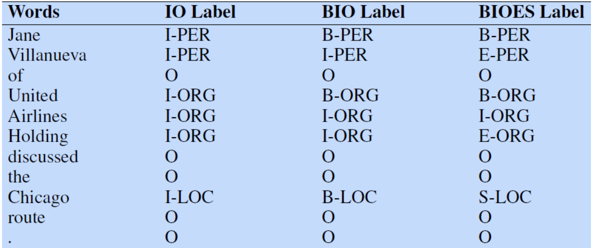
- Named entities (NE) are important information that are *distinct* from other similar information
- Mostly *proper nouns* in English
  - *Subhas Chandra Bose* established the *Azad Hind* government in *Singapore* on *Oct 21st, 1943* and declared war against the *Allied Forces* on *23rd [Oct, 1943]*.
- In Indian languages, may not be easily distinguishable
  - sarakāra kā jādū maṃtramugdha kara detā hai। (सरकार का जादू मंत्रमुग्ध कर देता है।)
  - yaha sarakāra kā jādū thā jisane garīboṃ ko bacāyā। (यह सरकार का जादू था जिसने गरीबों को बचाया।)
- Named entity detection is the task of *detecting* NEs
- Named entity recognition (NER) is the task of *detecting* NEs as well as assigning them correct *labels*

Named Entity Tags

- Named entity tagsets can vary depending on applications
  - Names: Person, Location, Organization, Geo-political Entity
  - Temporal: Date, Time
  - Numerical: Money, Number, Ordinal
  - Miscellaneous: Product
- For Indian languages, may require change
  - Hanumāna, Jatāyu, Ghaṭotkaca, etc. may be tagged as Person or another class "Non-Human"

NER Task

- NER task is to assign NER tags to words (single or span)
- Three schemes of tagging
  - IO: inside a span or outside (i.e., others)
  - BIO: indicates beginning of span as well
  - BIOES: indicates end of span also and single words



Words	**IO Label**	**BIO Label**	**BIOES Label**
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

- IO tagging may lose information
  - `rāma lakṣmaṇa bharata śatrughna ca gacchanti`
- For BIO, end tag is not needed, since there is others tag
- Number of classes for *BIO* tagging is $2n+1$ for n tags
- Can be cast as a *sequence-to-sequence* task

Relation Extraction

- Finding and classifying *semantic* relationships among entities
- Part-whole
  - `avayavī-avayava bhāva` (अवयवी-अवयव भाव)
- Location-of, Time-of
- Human relationships
  - Daughter-of, Mother-of, Sister-of, Wife-of, Ancestor-of, etc.
- Name-of, Alias-of
- Is-a (hypernymy)
- Is-kind-of (hyponymy)
- Owner-of
- Can be used to build KGs
- Domain-specific KGs require specialized entity and relationship types
  - Legal KG has statutes, is-precedent-of, etc.

Rule-based Relation Extraction

- Based on patterns or rules called Hearst patterns
- Genitive case (`sambandha vācaka` denoted by `ṣaṣṭhī vibhakti`)
  - `tasya bhrātā duḥśāsanaḥ`
- Example patterns
  - "...red algae, such as Gelidium, ..." → { Gelidium *is-kind-of* red algae *is-kind-of* algae }
  - Subhas Bose, Prime Minister of Azad Hind government → *person*, *position* of *organization*
- High precision but low recall

Machine Learning-based Extraction

- Supervised learning
- Training examples: sentences and corresponding relations
- Features
  - Word features: POS tags, head words in parse, bigrams
  - Entity features: NER tags
  - Parse tree features: phrases, paths
- Hard to collect large training data

Semi-supervised Learning

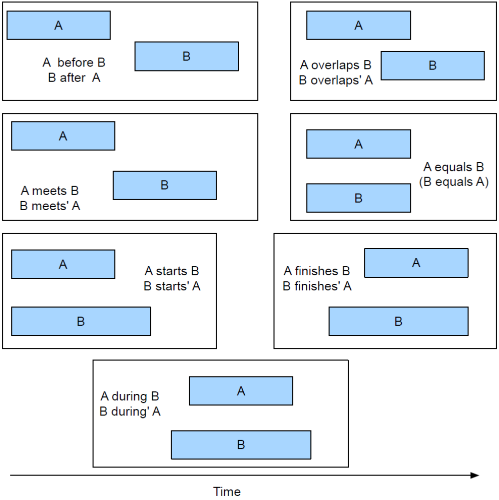
- A small amount of labeled data
- Seed patterns and seed tuples
- Bootstrapping
- Use seed tuples to identify sentences containing both entities
- Extract patterns from them
- Generate new seeds and patterns
- Can assign *confidence score* to patterns based on how many tuples follow
- Distant supervision
- Generates many patterns
- Uses features to classify them

Event Extraction

- Finding *events* in which entities participate
  - (Almost) every verb is an event
  - TAM (tense-aspect-modality) tags are important
- Types of events
  - Actions: go, kill, ...
  - States: sleep, ...
  - Reporting events: tell, discuss, explain, ...
  - Perception events: feel, think, ...
- Events are *temporal* in nature
- *Absolute* time
  - 21st October, 1943 or 29 Ashwin, 1865 Sakabda
- *Relative* time
  - two days from today
- *Duration*
  - this semester
- Can be cast as a sequence-to-sequence task with BIO tagging
- Can be rule-based or machine learning-based as well

Temporal Ordering

- Point events can be either *before* (*after*) or *equal*
- Temporal order between events with non-zero time-span
    - Allen relations


A
B
A before B
B after A
A
B
A overlaps B
B overlaps' A
A
B
A meets B
B meets' A
A
B
A equals B
(B equals A)
A
B
A starts B
B starts' A
A
B
A finishes B
B finishes' A
A
B
A during B
B during' A
Time