Information Extraction.

① Obtain structured text/ data / from unstructured data.

Objectives: IE system.

(i) Find & understand limited relevant parts of texts

(ii) Gather info from many different texts.

(iii) Produce a structured representation of the relevant info gathered.

Relations (database)

Knowledge base / knowledge graph.

End goals.

- Make the data "useful" for users
- Orgaize the info in a semantically precise manner. $\xrightarrow[\text{by}]{\text{used}}$ computer algorithms.

Input (unstructured / semi-structured data) $\longrightarrow$ structured data.

Information: (1) exact }
(2) factual }

Roughly : IE system

Who did What to whom and when?

The headquarters of BHP Billion Limited, and the global headquarters of the combined BHP Billion Group are located in Melbourne, Australia.

$$\underbrace{\boxed{\text{headquarters}}}_{\text{relation}}\left(\underbrace{\text{"BHP Billion Limited"}}_{\text{entity1}}, \underbrace{\text{"Melbourne, Australia}}_{\text{entity2.}}\right).$$

In 1998 Larry Page and Sergey Brin founded Google Inc.

FounderOf (Larry Page, Google Inc.)
FounderOf (Sergey Brin, Google Inc.)
FoundedIn (Google Inc., 1998).

Applications?

Biomedical domain.

① Large no. of scientific publications from the biomed domain

② Discover easily innovations (i) particular genes (ii) proteins & other entities.

③ Biomed entities ← ① many ambiguous names.
② lot of ~~synonymous~~ synonymous words.

Automatically identify the bio medical entities and link. them in knowledge bases.

↳ WikiData.

↓

subject	relation	object
p53	is_a	protein
p53	has_function	apoptosis
apoptosis	involved-in	cell death
⋮	⋮	⋮

News articles.

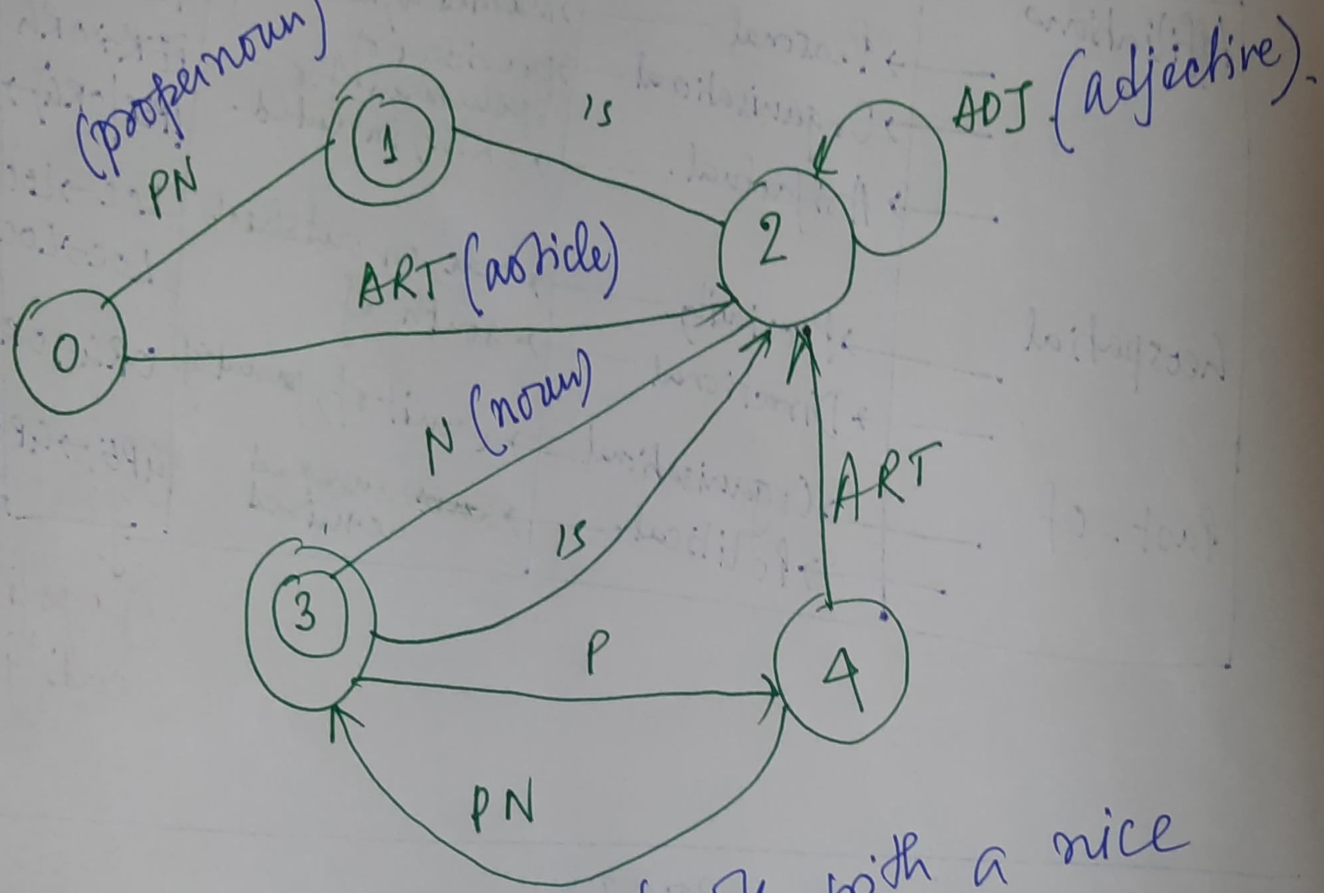
Relations.		Examples	Types.
Affiliations	→ Personal →	married to, mother of	PER → PER
	→ Organisational →	president of, spokeswoman for	PER → ORG
	→ Artifactual. →	owns, invented.	(PER|ORG) → ART
Geospatial	→ Proximity →	near, on outskirts	LOC → LOC
	→ Directional →	south of	LOC → LOC
Part-of	→ Organisational →	a unit of, parent of	ORG → ORG
	→ Political. →	annexed acquired	GPE → GPE

GPE → geopolitical entity.

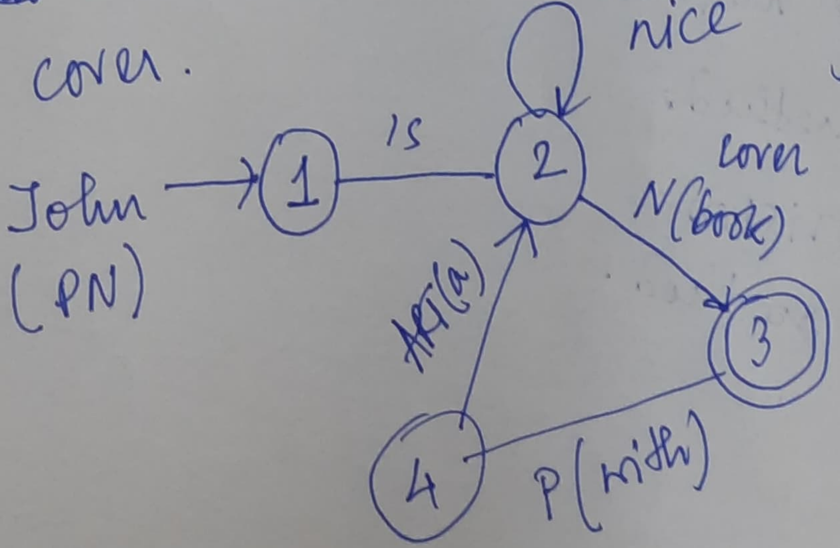
- Hand written rules (REGEX)
- Bootstrapping methods.
- Supervised methods
- Distant supervision
- Unsupervised method.

REGEX based IE.

Use finite automaton for noun groups:



John's interesting book with a nice cover.



many such REGEX to extract info.

→ Determine which person holds what position in what organization.

[person], [position] of [org]. ←

Vuk Draskovic, (leader of) the Serbian Renewal Movement }

[org] (named, appointed, etc.) [person] PREP [office].

(NATO) appointed Wesley Clark (as) (Commander in Chief)

org person preposition. office.

→ Determine where an organization is located.

[org] in [loc].

NATO headquarters in Brussels.

[org], [loc] (division, branch, headquarters, etc.).

KFOR, Kosovo headquarters.

Agar is a substance prepared from a mixture of red algae such as Gelidium for lab or industry use.