Information Extraction

Goals of Information Extraction

- An important research area for natural language processing and text mining is **the extraction and formatting of information from unstructured text**.
- Computers can be used to sift through a large amount of text and extract restricted forms of useful information, which can be **represented in a tabular format**.
- Information extraction can be regarded as a restricted form of full natural language understanding, where we know in advance what kind of **semantic information** we are looking for.
- The main task is then to extract parts of text to **fill in slots in a predefined template**.

Goals of Information Extraction

- A task defined as **executive position changes**:

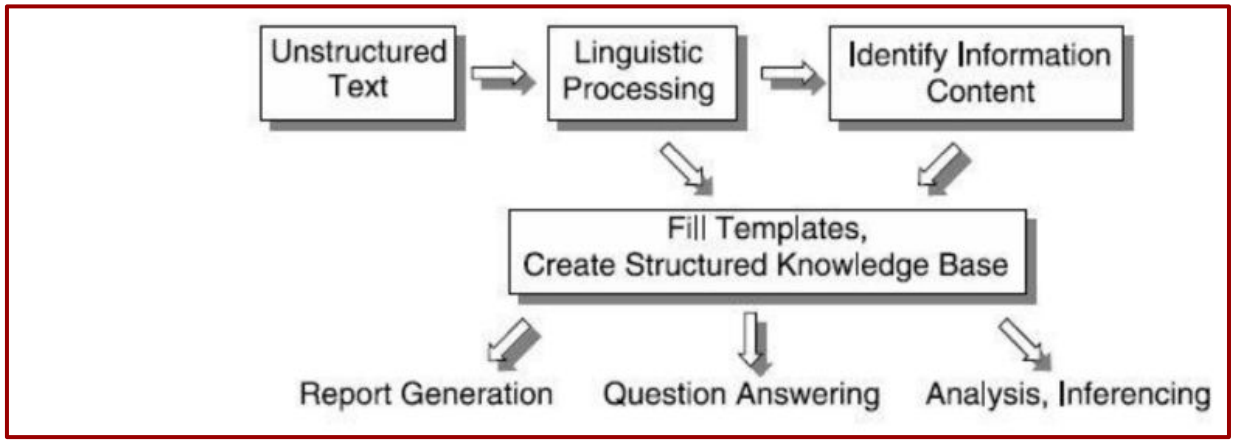
One of the many differences between *Robert L. James, chairman and chief executive officer of McCann-Erickson*, and *John J. Dooner, Jr.*, the agency's president and chief operating officer, is quite telling: Mr. James enjoys sailboating, while Mr. Dooner owns a powerboat.

Now, Mr. James is preparing to sail into the sunset, and Mr. Dooner is poised to rev up the engines to guide *Interpublic Group's McCann-Erickson* into the 21st century. Yesterday, *McCann* made official what had been widely anticipated: *Mr. James, 57* years old, is stepping down as chief executive officer on *July 1* and will retire as chairman at the *end of the year*. He will be succeeded by *Mr. Dooner, 45* ...

Predefined Domain	Extracted Information
Organization	*McCann-Erickson*
Position	*Chief executive officer*
Date	*July 1*
Outgoing person name	*Robert L. James*
Outgoing person age	*57*
Incoming person name	*John J. Dooner, Jr.*
Incoming person age	*45*

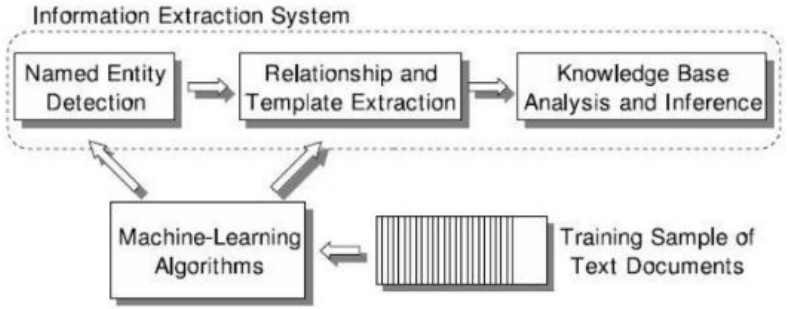
Goals of Information Extraction

- A general information extraction system is illustrated in the blow figure.
- The task of information extraction naturally decomposes into a sequence of processing steps, typically including **tokenization, sentence segmentation, part-of-speech assignment, named entity identification, phrasal parsing, sentential parsing, semantic interpretation, template filling,** and **merging**.



Goals of Information Extraction

- The most accurate information extraction systems often involve human effort: handcrafted language processing modules.
  - People's names have prefixes such as Mr., Mrs., Miss., Dr., Jr.,
  - People's names are recognized by phrases such as "according to.." or "...said"
- The application of machine-learning techniques to information extraction is motivated by the time-consuming process needed to handcraft these systems.
- The general architecture of a machine-learning-based information extraction system is given as below:



Goals of Information Extraction

- There are typically two main modules involved in such a system.
- The purpose of the first module is to annotate the text document and find portions of the text that interest us (**Name Entity extraction**)
  - For example, we want to identify the string *Robert L. James* as a **person** and the string *McCann-Erickson* as an **organization**. human effort: handcrafted language processing modules.
- Once such entity mentions are extracted, another module is invoked to extract high-level information based on the entity mentions (**Relationship extraction**).
  - In the example of Fig. 6.1, we want to identify that the person *Robert L.James* **belongs to** the organization *McCann-Erickson*, and his age is **57**.
- The information is then filled into slots of a predefined template.

Example of Information Extraction

- As a task: Filling slots in a database (template) from corpus

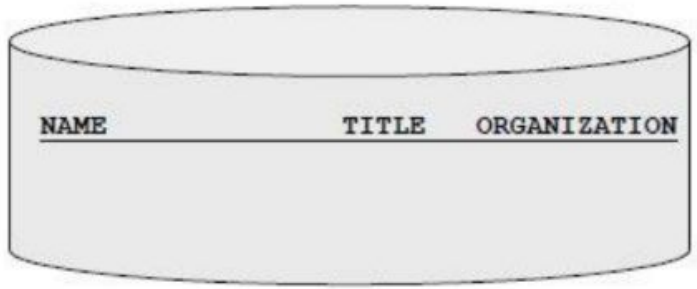
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



Example of Information Extraction

- As a task: Filling slots in a database (template) from corpus

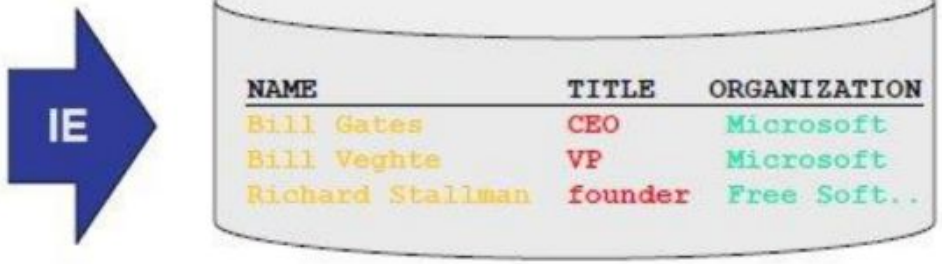
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

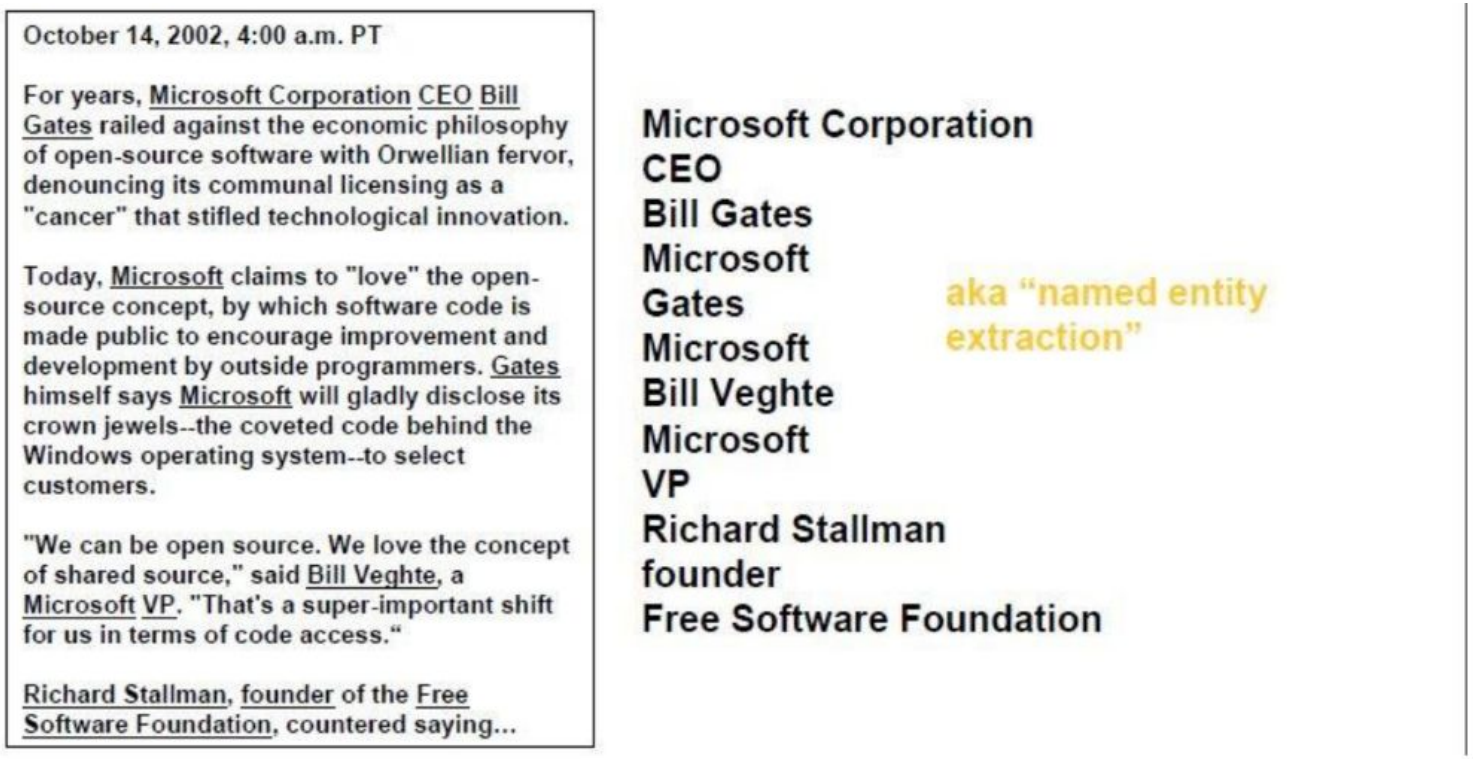
Richard Stallman, founder of the Free Software Foundation, countered saying...



Example of Information Extraction

- As a family of techniques:

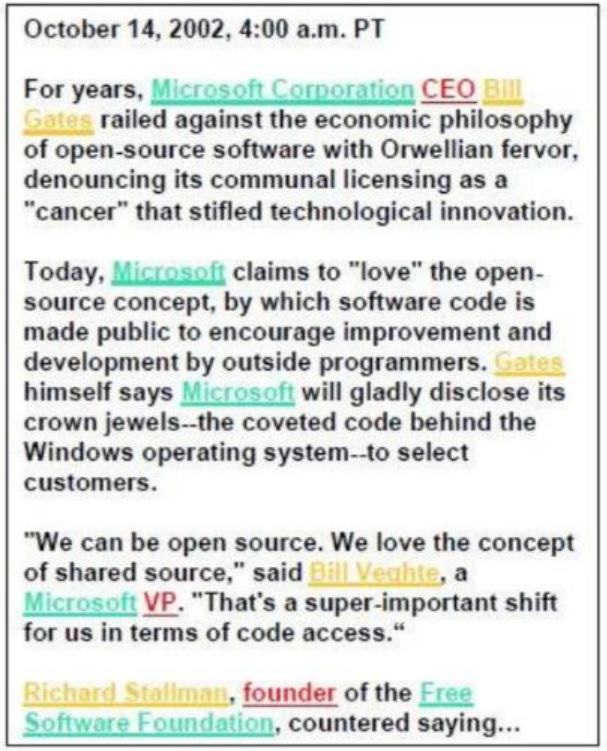
Information Extraction = Segmentation + Classification + Association + Clustering



Example of Information Extraction

- As a family of techniques:

Information Extraction = Segmentation + Classification + Association + Clustering

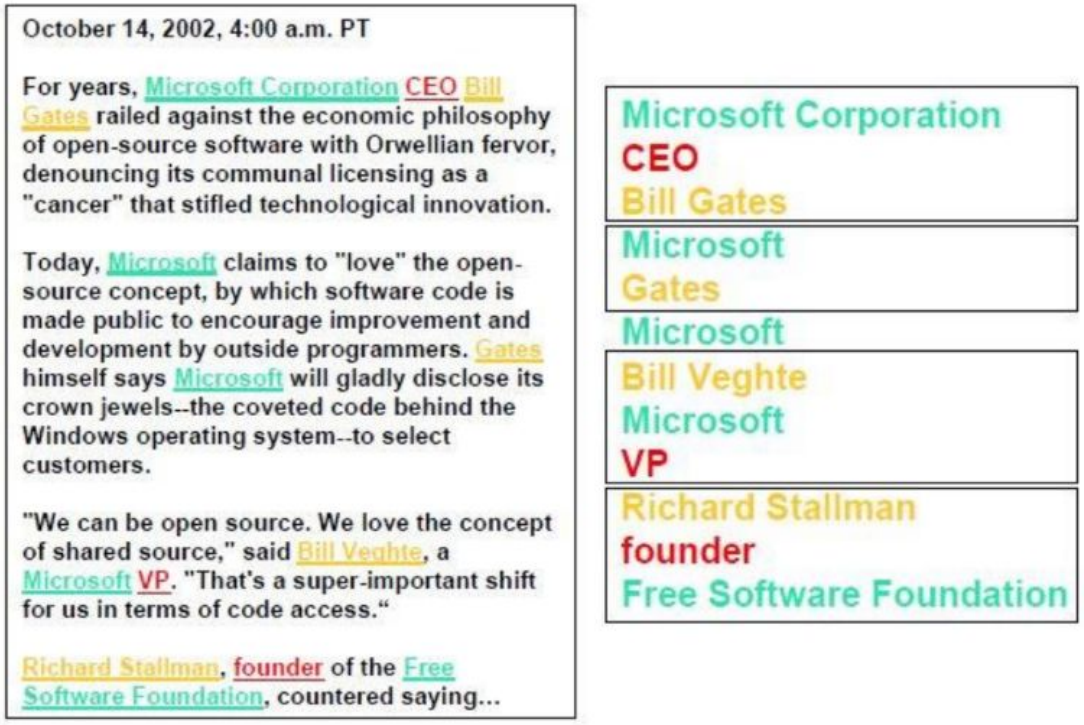


Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

Example of Information Extraction

- As a family of techniques:

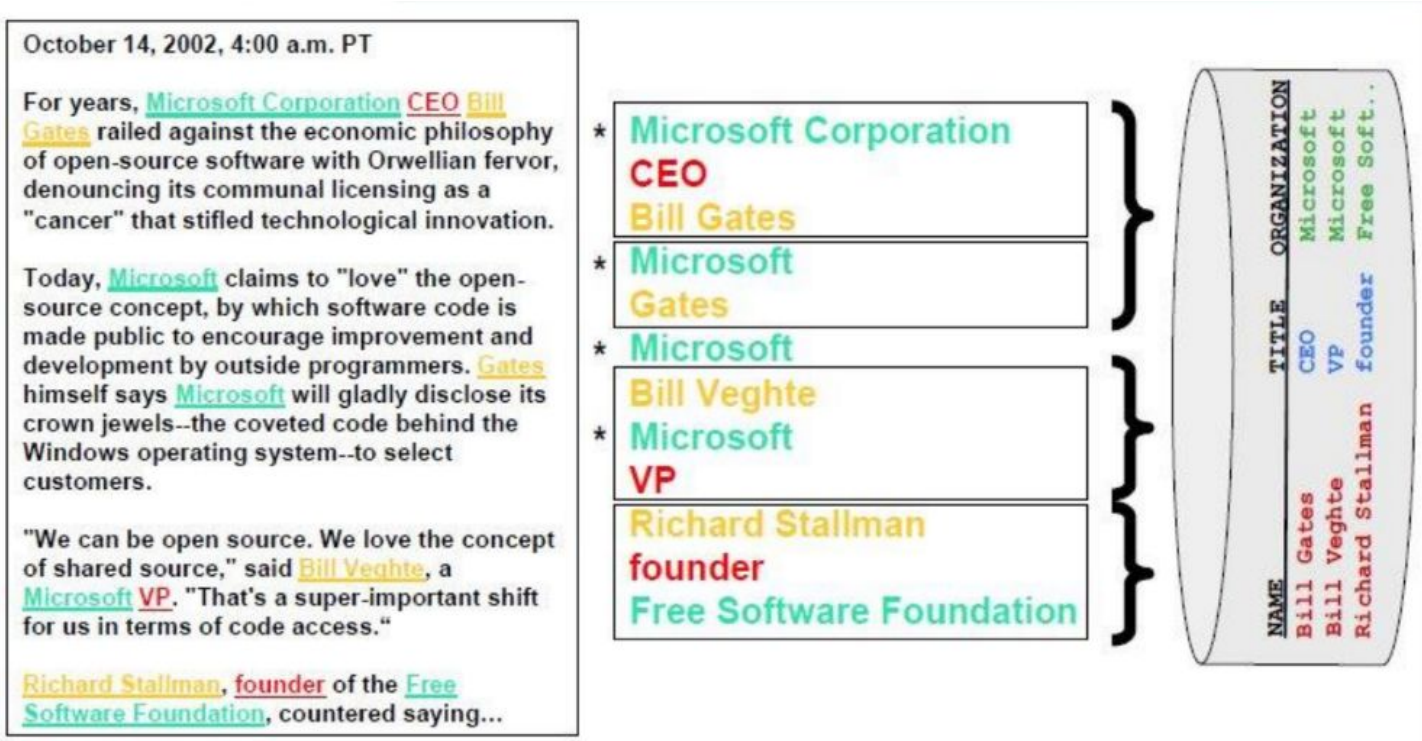
Information Extraction = Segmentation + Classification + Association + Clustering



Example of Information Extraction

- As a family of techniques:

Information Extraction = Segmentation + Classification + Association + Clustering



Named Entity Recognition

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text for example:

The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person
Date
Organization

Goals of Information Extraction

- The uses:
  - Named entities can be indexed, linked off, etc.
  - A lot of IE relations are associations between named entities
  - For question answering, answers are often named entities.
  - Sentiment can be attributed to companies or products.
- Concretely:
  - Many web pages tag various entities, with links to bio or topic pages, etc.
    - Retuers' OpenCalais, AlchemyAPI, Yahoo'sTerm Extraction,...
  - Microsoft: smart recognizers for document content
    - E.g., recognize a name, can take actions, such as add to contacts and open contacts.

The NER Task

- Task: Predict entities in a text

Foreign	ORG
Ministry	ORG
spokesman	O
Shen	PER
Guofang	PER
told	O
Reuters	ORG
:	:

Standard evaluation is per entity, *not* per token

Precision/Recall/F1 for IE/NER

- Recall and precision are straightforward for tasks like text categorization.
- The measure is a bit different for IE/NER when there are *boundary errors* (which are common):
  - First Bank of Chicago announced earnings…..
- This counts both a false positive and a false negative
- Select **nothing** would have been better.
- Some other metrics (e.g., MUC scorer) give partial credit (according to complex rules)

Sequence Problems

- In document, each sentence or phrase contains a sequence of words.
- We can label each item in a sequence for **name entity recognition**.
  - A sequence classifier or sequence labeler is a model whose job is to assign some label or class to each unit.

PERS	O	O	O	ORG	ORG
Murdoch	discusses	future	of	News	Corp.

Named entity recognition

VBG	NN	IN	DT	NN	IN	NN
Chasing	opportunity	in	an	age	of	upheaval

POS tagging

The ML sequence model approach to NER

- Training
  - Collect
  - Label each token for its entity class or other (O)
  - Design feature extractors appropriate to the text and classes
  - Train a sequence classifier to predict the labels from the data
- Testing
  - Receive a set of testing documents
  - Run sequence model inference to label each token
  - Appropriately output the recognized entities

Encoding Classes for Sequence Labeling

	IO encoding	IOB encoding (short for Inside, Outside, Beginning)
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	I-PER
Huang	PER	I-PER
's	O	O
new	O	O
painting	O	O

Which encoding method need more training data?

Features for sequence labelling

- Words
  - Current word
  - Previous/next word (context)
- Other kinds of inferred linguistic classification
  - Part-of-speech tags
- Label context
  - Previous label

Input (current word: London): *Thousands of demonstrators have marched through London to protest the war in Iraq and demand the withdrawal of' British troops from that country.*
Label: ....., ('through', 'O'), ('London', 'B-geo'), ('to', 'O'), ...

Features for the word **London**

```
{'bias': 1.0,
 'word.lower()': 'london',
 'word[-3:]': 'don',
 'word[-2:]': 'on',
 'word.isupper()': False,
 'word.istitle()': True,
 'word.isdigit()': False,
 'postag': 'NNP',
 'postag[:2]': 'NN',
 '-1:word.lower()': 'through',
 '-1:word.istitle()': False,
 '-1:word.isupper()': False,
 '-1:postag': 'IN',
 '-1:postag[:2]': 'IN',
 '+1:word.lower()': 'to',
 '+1:word.istitle()': False,
 '+1:word.isupper()': False,
 '+1:postag': 'TO',
 '+1:postag[:2]': 'TO'},
```

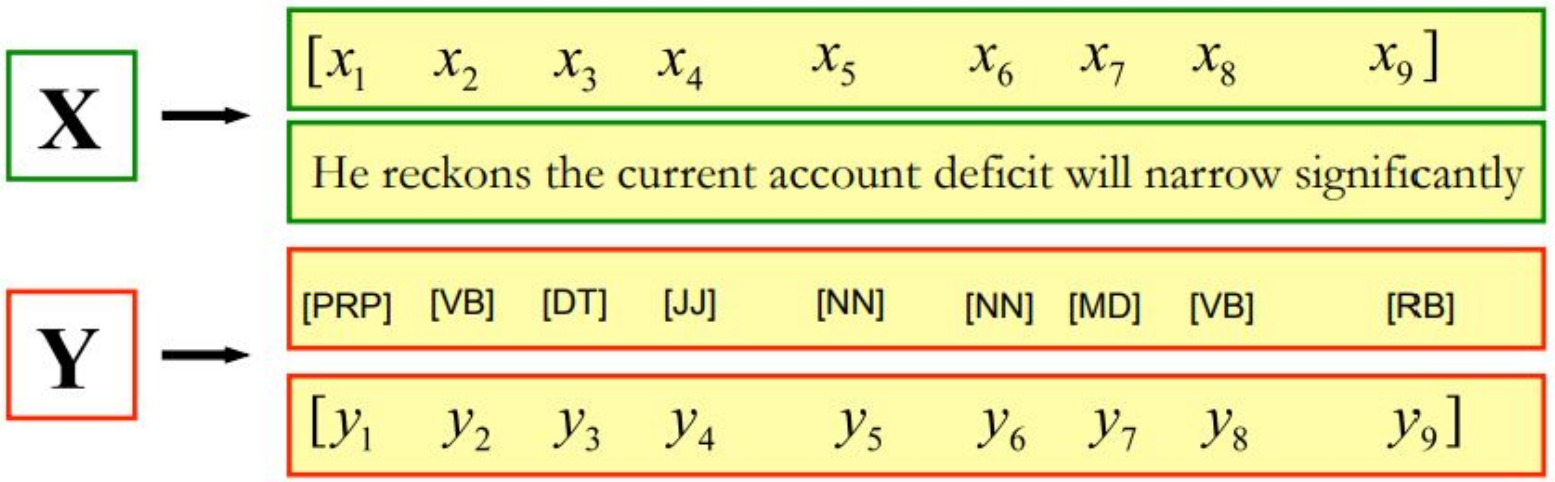
Sequence Labelling

- A widely used algorithm for sequence labelling
- Finds the most probable label sequence **y** given an observation sequence **x**

$$y = argmax_y p(y|x)$$

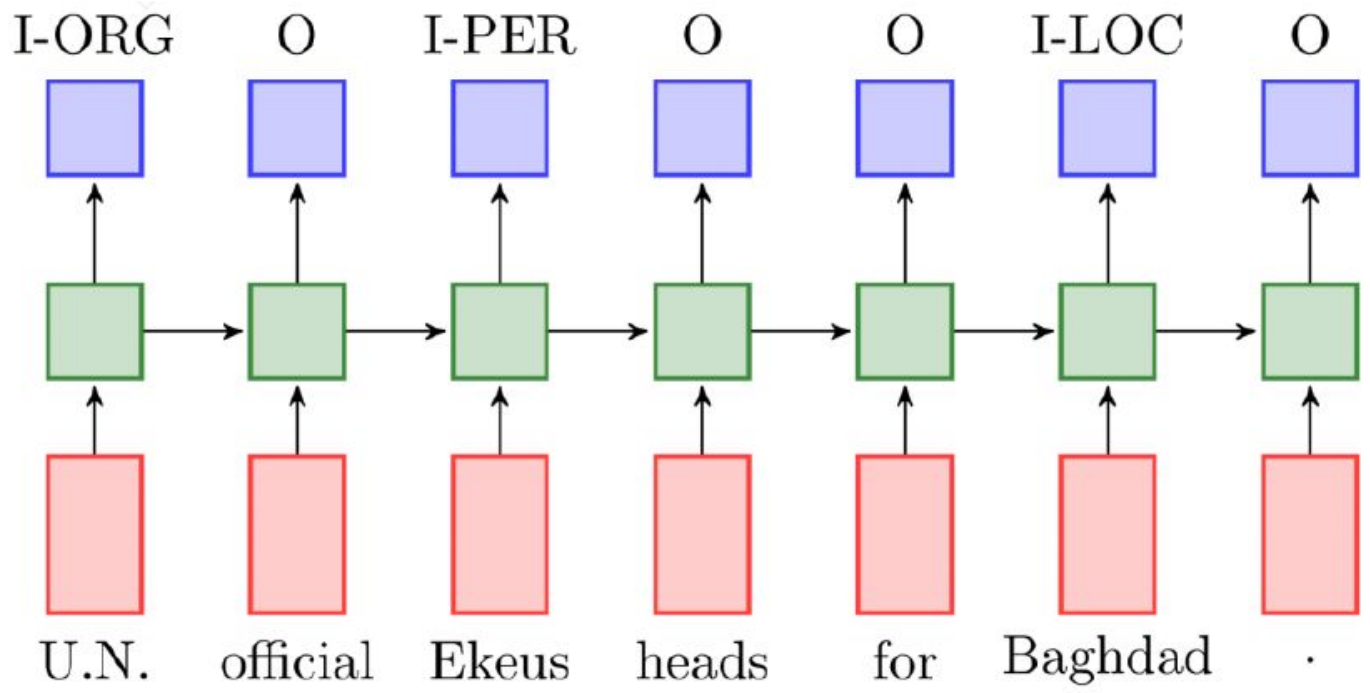
- Where **x** consists of the sequence of tokens from input text.

Example: Part-of-Speech Tagging



Sequential Model

- Hidden Markov Model
- Conditional Random Field
- RNN



From [Guide to Sequence Tagging with Neural Network in Python]

Notebook

- How to use Spacy for NLP tasks