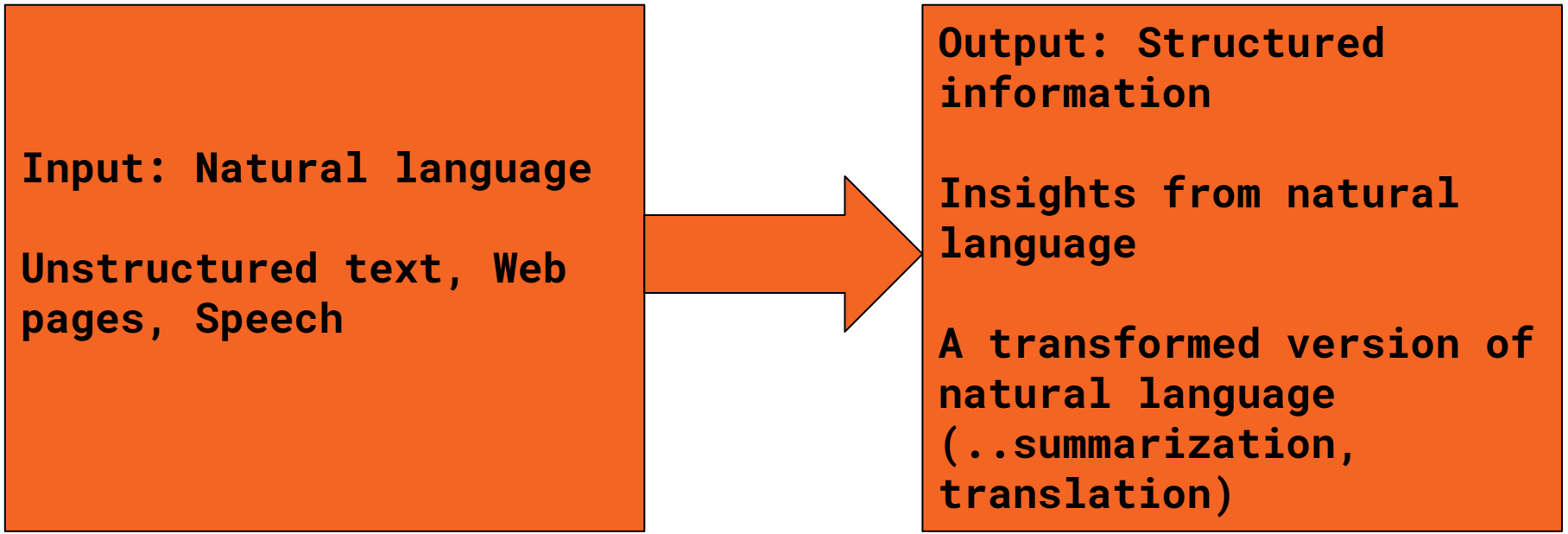
Introduction to NLP

Sandeep Tammu

Natural language processing



Question-Answering (Jeopardy game)

IBM Watson

On Sept. 1, 1715 Louis XIV died in this city, site of a fabulous palace he built.

Versailles

Spam Classification

Black Friday Begins ==>> Mysterical Money link inside..

Killer Mind Control secrets

Let’s meet tomorrow!

Language Technologies

Mostly solved: Spam detection, Parts of speech tagging, Named Entity Recognition

Making good progress: Sentiment analysis, Coreference resolution, Word Sense disambiguation, Parsing, Machine Translation, Information Extraction

Still very hard: Question Answering (QA), Paraphrase detection, Summarization, Dialog

Parts of speech tagging

Colorless green ideas sleep furiously.

ADJ ADJ NOUN VERB ADV

Named Entity Recognition

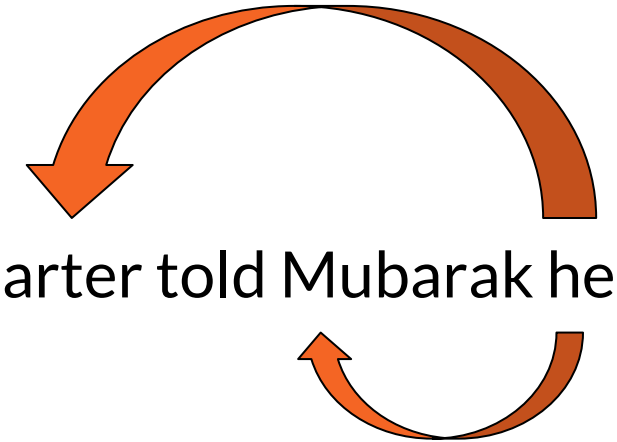
- Identifying Person, Location, Organisation

Einstein met with UN officials in Princeton

PERSON ORG LOC

Person names, Organizations (companies, government organisations, committees, etc), Locations (cities, countries, rivers, etc), Date and time expressions, Other common types: measures (percent, money, weight etc), email addresses, Web addresses, street addresses, etc.,Some domain-specific entities: names of drugs, medical conditions, names of ships, bibliographic references etc.

Coreference resolution: Linking entities



Carter told Mubarak he shouldn’t run again.

Sentiment analysis

the camera really takes good images and you would not be left to desire more.

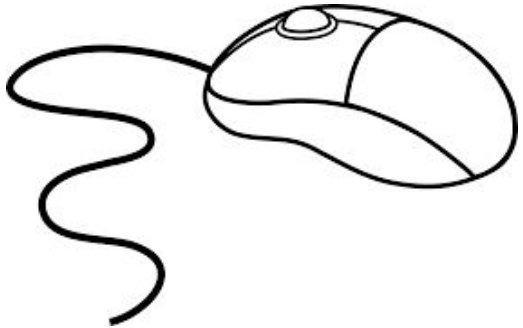
the camera was already "used"

Worste after receiving product after 4 days long lens zoomin not working

it supurib

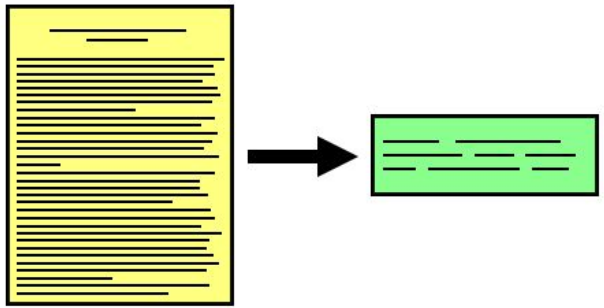
Word Sense disambiguation

I need new batteries for my **mouse**.



Text Summarization

- Summarize longer documents into important sentences
- Needs to understand important keyphrases and concepts in document
- Abstractive, Extractive



Paraphrase detection: Writing in other words

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Dialog and conversing systems

Where is Doctor Strange playing in Bangalore?

INOX at 7:30 pm. Would you like to book a ticket?

Yes/ Ok/ No

I am booking a ticket for the movie Doctor Strange tomorrow at 7:30pm for one person, can you confirm?

Yes.

Challenges in natural language

- Ambiguity: No general rules

  Teacher Strikes Idle Kids, Hospitals Are Sued by 7 Foot Doctors

- Non-standard english: want 2 go, b4 u
- Segmentation issues: #customerexperience
- Idioms: No fixed meaning eg: dark horse
- Neologisms: new words unfriended, BFF
- World knowledge:

  Mary and Sue are sisters, Mary and Sue are mothers.

Tokenization

- Issues in English tokenization

Finland’s capital: Finland, Finlands, Finland’s ?

what’re, I’m, isn’t: What are, I am, is not

Hewlett-Packard state-of-the-art, San Francisco, New York

- Acronyms: m.p.h., PhD.

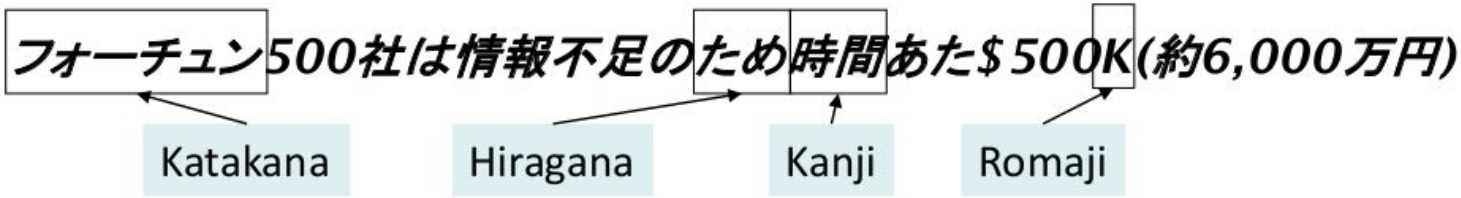
Tokenization

- Other languages: Compound Nouns (German)

  Lebensversicherungsgesellschaftsangestellter

  ‘life insurance company employee’

- Japanese: Further complicated, users can use different kinds of language in a sentence



Lemmatization

- Reduce variations to base form

  car, cars, car's, cars' to car

  organize, organizes, and organizing

- Have to find the correct form in dictionary
- Does things properly using vocabulary and considering the context in which the word is used in.

Stemming

- Reduce the words in a crude manner

Automate (s), automatic, automation to automat

- Rule-based

Caresses: caress, SSES: SS

- Chops off the ends of words
- End user will not be able to interpret the stem all the time

Demo!

NLP in 3 lines: Spacy

- Python package with consistent API
- Very fast, accurate and easy

```
from spacy.en import English
engine = English()
parsedDoc = engine('document')
```

References

- https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html
- https://www.cse.iitb.ac.in/~nlp-ai/WSD.ppt