Ubiquity of Communication Devices

Figure 1: Computers are everywhere and processing/generating data.
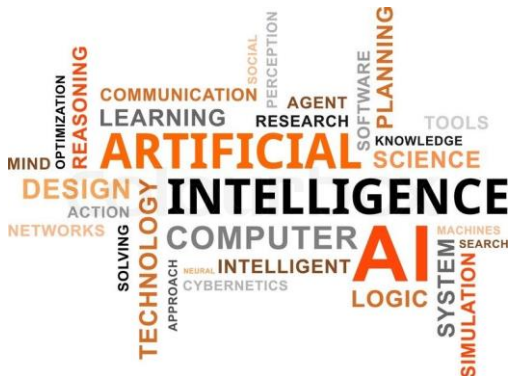
Text

Figure 3: A big chunk of data out there is text.

Why is NLP Hard?

Natural Language is highly Ambiguous

d Use of the same word to mean different things

d Ambiguity facilitates shorter expressions

d Relies on peoples ability to use their knowledge

I saw a man on the hill with a telescope.

Why is NLP Hard?

Natural Language is highly Ambiguous

d Use of the same word to mean different things

d Ambiguity facilitates shorter expressions

d Relies on peoples ability to use their knowledge

I saw a man on the hill with a telescope.

Why is NLP Hard?

Natural Language is highly Ambiguous
- *d* Use of the same word to mean different things
- *d* Ambiguity facilitates shorter expressions
- *d* Relies on peoples ability to use their knowledge
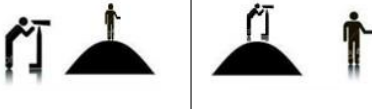
I saw a man on the hill with a telescope.

Why is NLP Hard?

Natural Language is highly Ambiguous

d Use of the same word to mean different things

d Ambiguity facilitates shorter expressions

d Relies on peoples ability to use their knowledge

I saw a man on the hill with a telescope.

Why is NLP Hard?

Natural Language is highly Ambiguous

d Use of the same word to mean different things

d Ambiguity facilitates shorter expressions

d Relies on peoples ability to use their knowledge

I saw a man on the hill with a telescope.

Why is NLP Hard?

Natural Language is highly Ambiguous

d Use of the same word to mean different things

d Ambiguity facilitates shorter expressions

d Relies on peoples ability to use their knowledge

I saw a man on the hill with a telescope.

NLP Challenges

mostly solved

Spam detection

| One Day Trip to TOLI PEER! | ✔ |
| Your Assistance is Urgently Needed ... | ✘ |

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV

Colorful green students sleep furiously.

Named entity recognition (NER)

PERSON   ORG   LOC

Imran met with UN officials in Islamabad

NLP Challenges

making good progress

Sentiment analysis

Best biryani in Rawalpindi!

The waiter ignored us for 20 minutes.

Coreference resolution

Imran told Shahbaz he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my *mouse*.

Parsing

I can see Faisal Masjid from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Conference…

Information extraction (IE)

You're invited to our welcome party, Friday Sep 27 at 6:30

Party
Sep 27
add

NLP Challenges

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

Google acquired Kaggle last year

Kaggle has been taken over by Google

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨ Economy is good

Dialog

What time is dinner at Monal?

At 7:30. Do you want to make a booking?

Word Segmentation

- Breaking a string of characters (graphemes) into a sequence of words.
- In some written languages (e.g. Urdu) words are not separated spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ( )]

Word Segmentation

- Breaking a string of characters (graphemes) into a sequence of words.
- In some written languages (e.g. Urdu) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ( )]

بسم الله الرّحمن الرّحيم

Morphological Analysis

- Morphology is the field of linguistics that studies the internal structure of words.
- A morpheme is the smallest linguistic unit that has semantic meaning
    - e.g. *carry, pre, ed, ly, s*
- Morphological analysis is the task of segmenting a word into its morphemes
    - carried → carry + ed (past tense)
    - independently → in + (depend + ent) + ly
    - Googlers → (Google + er) + s (plural)
    - unlockable → un + (lock + able) ?
    - → (un + lock) + able ?

Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech
- Useful for subsequent syntactic parsing and word sense disambiguation

Strong stocks are good
ADJ N V ADJ

Big weak stocks are bad
ADJ ADJ N V ADJ

Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings.

    - Atqa has a strong *interest* in computational linguistics.
    - Sarah pays a large amount of *interest* on her credit card.

- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

  agent patient source destination instrument
  - Hasan drove Omer from Lahore to Islamabad in his Toyota Prius.
  - The hammer broke the window.

- Also referred to a *case role analysis, thematic analysis,* and *shallow semantic parsing*

Semantic Parsing

- A semantic parser maps a natural-language sentence to a complete, detailed semantic representation (logical form).
- For many applications, the desired output is immediately executable by another program.
- Example: Mapping an English database query to Prolog:
  - How many cities are there in the Pakistan?
    ```
    answer(A, count(B, (city(B), loc(B, C),
            const(C, countryid(Pakistan))),
          A))
    ```

Textual Entailment

- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.

Umar broke the window	The window broke
She is a Pakistani; she is therefore brave	Being brave is a consequence of being Pakistani
Even Ahsen could solve that problem	Ahsen is the last person you'd expect to be able to solve the problem
Since its formation in 1948, Israel fought many wars with neighboring Arab countries	Israel was established in 1948

Anaphora Resolution — Co-Reference

- Determine which phrases in a document refer to the same underlying entity.
    - Hammad put the chip on the plate and ate it.
    - Bush started the war in Iraq. But the president needed the consent of Congress.
- Some cases require difficult reasoning.
    - Today was Ayesha's birthday. Hina and Sara went to the store. They were going to get presents. Sara decided to get a handbag. "Don't do that," said Hina. "Ayesha has a handbag. She will make you take it back."

Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.

- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.

  people organizations places

  Asad Umar was the CEO of Engro Corporation and lives in Rawalpindi.

- Relation extraction identifies specific relations between entities.

  Asad Umar was the CEO of Engro Corporation and lives in Rawalpindi.

Question Answering

Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the web).

- When was Imran Khan born? (factoid)
  October 5, 1952
- Who was the King of Saudi Arabia when Imran Khan was born?
  AbdulAziz

Reading Comprehension

Read a passage of text and answer questions about it.

Airport
The Stanford Question Answering Dataset

An **airport** is an aerodrome with **facilities** for **flights** to take off and **land**. Airports often have **facilities** to store and maintain aircraft, and a control tower. An **airport** consists of a **land**ing area, which comprises an aerially accessible open space including at least one operationally active surface such as a runway for a plane to take off or a helipad, and often includes adjacent utility buildings such as control towers, hangars and terminals. Larger airports may have fixed base operator services, **airport** aprons, air traffic control centres, passenger **facilities** such as restaurants and lounges, and emergency services.

What is an aerodrome with facilities for flights to take off and land?
airport

What is an aerially accessible open space that includes at least one active surface such as a runway or a helipad?
landing area

What is an airport?
aerodrome with facilities for flights to take off and land

Text Summarization

- Produce a short summary of a longer document or article.
  - **Article**: With a split decision in the final two primaries and a flurry of superdelegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary Rodham Clinton that will make him the first African American to head a major-party ticket.
  Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against Sen. John McCain, the presumptive Republican nominee.
  - **Summary**: Senator Barack Obama was declared the presumptive Democratic presidential nominee.

Machine Translation (MT)

Translate a sentence from one natural language to another.

کوئی بات ایسی اگر ہوئی، کہ تمہارے جی کو بری لگی

Machine Translation (MT)

- Translate a sentence from one natural language to another.

کوئی بات ایسی اگر ہوئی، کہ تمہارے جی کو بری لگی

Anyway if it happened, that was bad for you