Mining Unstructured Data

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Outline

Outline

What is unstructured data?

- Information which is not organised following a pre-defined model
- This data may be from:
  - Human language (text/speech): collections of well written documents (articles, books, legal notes,...), collections of non-standard textual documents (sms, tweets, opinions, webpages, health records, chats, speech transcripts...)
  - Audio: space exploration recordings, ...
  - Image/video: digital photos (face images,...) or videos (military tracking, athmospheric movements, ...)

What is unstructured data?

- Information which is not organised following a pre-defined model
- This data may be from:
  - Human language (text/speech): collections of well written documents (articles, books, legal notes,...), collections of non-standard textual documents (sms, tweets, opinions, webpages, health records, chats, speech transcripts...)
  - Audio: space exploration recordings, ...
  - Image/video: digital photos (face images,...) or videos (military tracking, athmospheric movements, ...)

  This course focuses on **data from human language**, as it is the type most frequently used for unstructured data mining

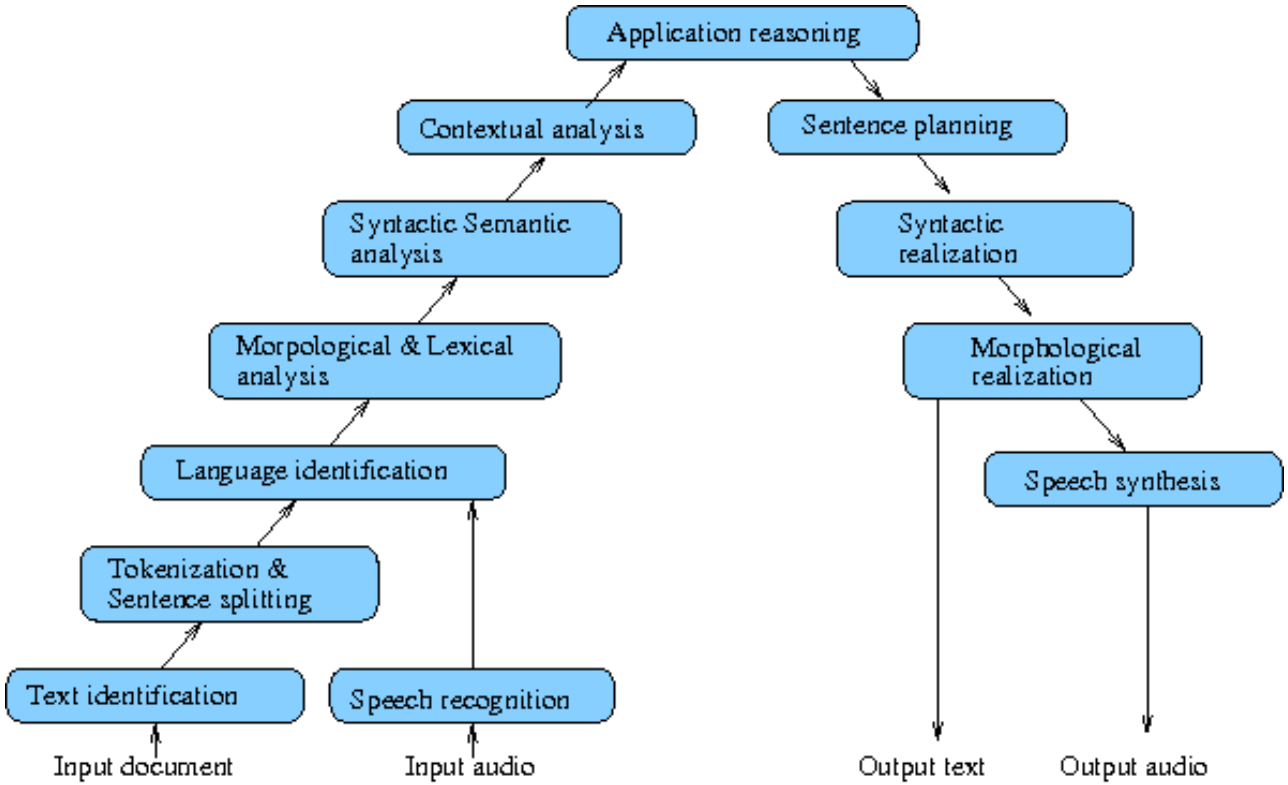
Outline

Definitions

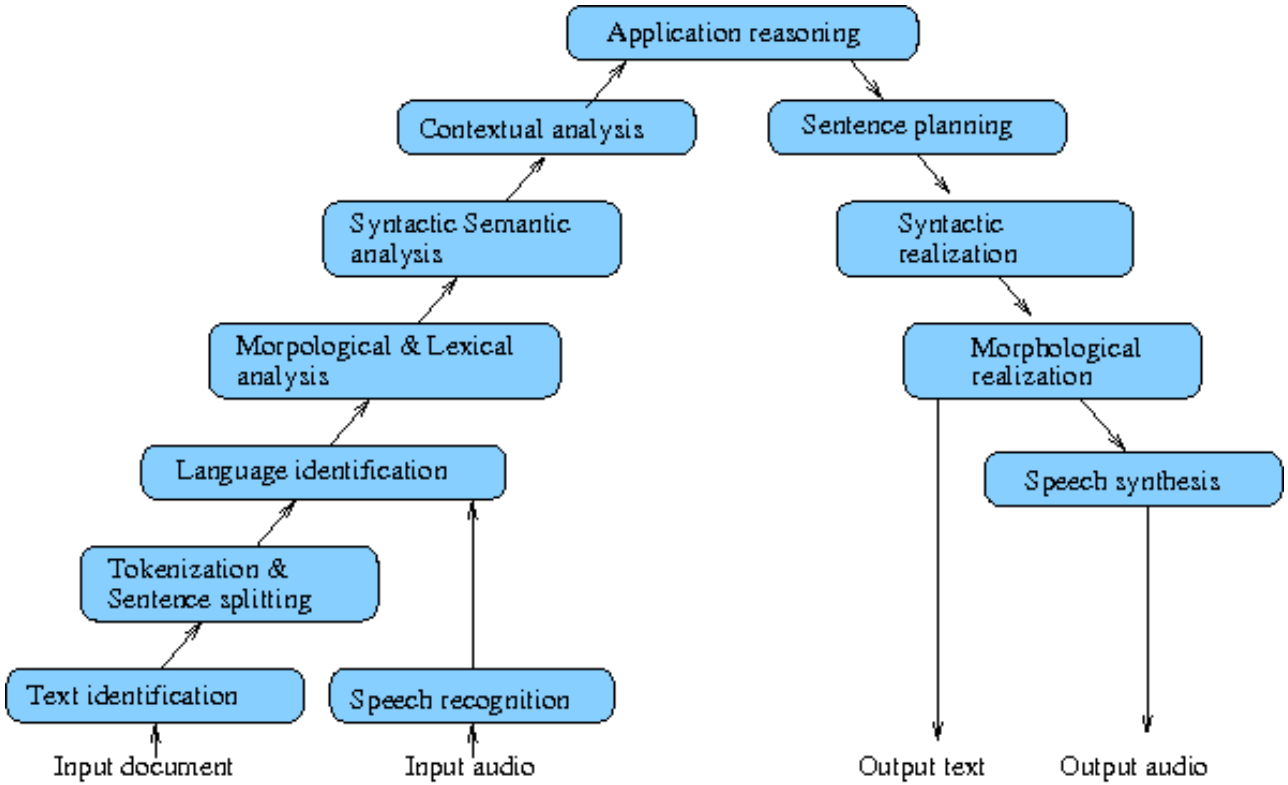
The general strategy follows the standard subareas of linguistics:

- Phonetics: sounds of human speech.
  E.g., *infrequent* → `/ɪn'frikwənt/`
- Morphology: structural formation and categorisation of words.
  E.g., *in-frequent-ly*, *'the' is Determiner*.
- Syntax: structural relations between words in sentences.
  E.g., *a determiner is followed by a common noun*.
- Semantics: meenings of words and their composition via syntax.
  E.g., *the president of USA is Donald Trump* → `president(USA, Donald_Trump)`
- Pragmatics: meaning in the context.
  E.g., **He** *is very well known in* **his country** [sarcasm]

General architecture



General architecture



- Branches: NL Understanding and NL Generation.
- Approaches: Knowledge-based *vs.* Statistical-based.
- Shallow methods (lexical overlap, pattern matching) *vs.* Deep methods (semantic analysis, logical inference)

Outline

Problems

- World-knowledge
  - Representing world-knowledge is mandatory for understanding NL (AI-completeness)
    e.g., Yago - facts, OpenCyc - common sense
- Multilinguality
  - Different languages require different models and resources
  - Use of words from other languages
    `Estoy a full!` (non-standard Spanish text)
- Evaluation
  - Correctness/suitability of a translation/summary
- Variability
  - Different sentences refer to one meaning
    `Where can I get a map?`
    `I need a map`
    `need map` (non-standard text)
- Ambiguity
  - One sentence refers to different meanings
    `Esther said about Alice: ''I made her duck''`

Ambiguity

E.g., Esther said about Alice: ''I made her duck''

- I cooked waterfowl for her
- I cooked the waterfowl she owned
- I created the duck she owns
- I caused her to quickly lower her head or body
- I turned her into waterfowl

Word	Ambiguity	Alternatives
make	semantic	cook or create
her	syntactic	possessive or dative pronoun
	pragmatic	Esther or Alice
duck	synt-sem	noun or verb

Outline

Examples of applications

- Document clustering
- Document classification (e.g. anti-spamming, email routing, sentiment polarity, language identification)
- Information Retrieval
- Text correction
- Plagiarism detection
- Information Extraction
- Automatic Summarization
- Question Answering
- Machine Translation
- Dialog Systems

...

Information Retrieval (IR)

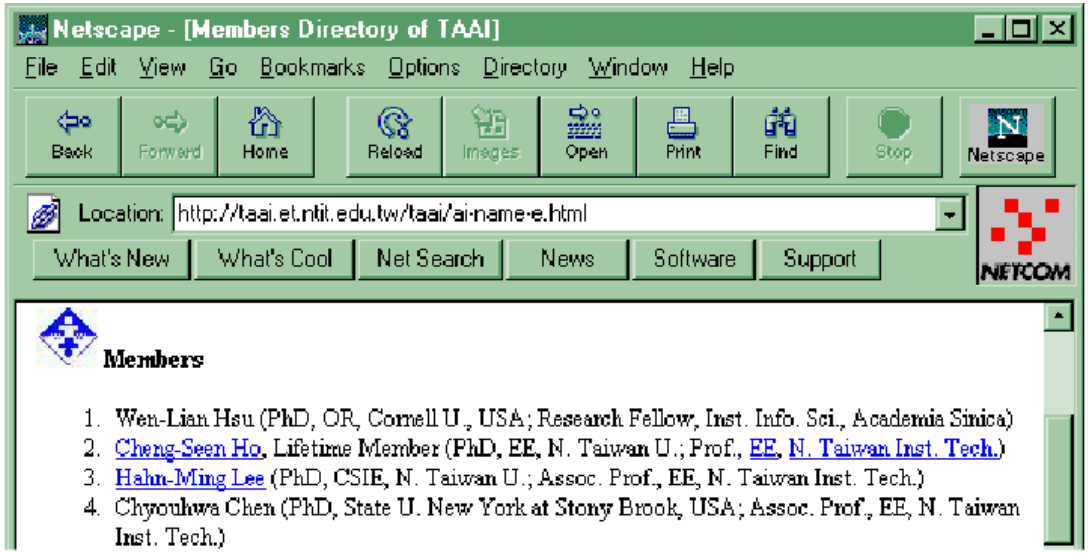
- E.g.: Searchers (Google, Yahoo, ...)
- Given a corpus, $D = \{D_i\}$, and a user query (list of words), Q , provide $\hat{D} \subset D$ that better match Q.
- $sim(v(Q), v(D_i))$, where $v(X)$ represents X in a vector space
- What vector space seems better?
  - words? $Q =$"window", $D_i =$"...he closed the windows..."
  - lemmas? $Q =$"window", $D_i =$"...he closed Windows..."
  - compounds? $Q =$"Energie", $D_i =$"...Sonnenenergie..."
  - ...
  - In-depth NLP seems not productive

Information Extraction (IE)

- E.g.: Enriching DBs or KBs with new content. Document collection indexing. Sentiment analysis.
- Extract the **relevant information** contained in text (entities, properties, relationships and events).
- Main subtasks:
  - Named Entity Recognition and Classification (NERC)
  - Slot Filling
  - Relationship Extraction
  - Event Extraction
- Depending on the specific task, more in-depth NLP is required (syntax, semantics, pragmatics, world-knowledge), as well as ML techniques.

Information Extraction (IE)

- Example 1: Member Name, Degree, School and Affiliation from WEB pages.



Name	Degree	Affiliation	School
Wen-Lan Hsu	PhD, OR, Cornell U., USA	Research Fellow	Inst. Info. Sci. Academia Sinica
Chen-Seen Hu	PhD, EE, N. Taiwan U.	Prof.	EE, N. Taiwan Inst. Tech
Hahn-Ming Lee	PhD, CSIE, N. Taiwan U.	Prof.	EE,N. Taiwan Inst. Tech
...			

Information Extraction (IE)

- Example 2: incidents from free text (type of incident, perpetrator, target, date, location, effects, instrument).

```
At 5pm on  Thursday , a  white Fiat van  veered off the road and into a
crowd outside the Plaça de Catalunya metro station in Barcelona.   The  van
continued down  Las Ramblas  for more than 500 metres while  crashing  into
 pedestrians .   13 people have been killed .   100 people were injured  and
 15 are in serious condition .  Las Ramblas attacker  Younes Abouyaaqoub
 was  killed in Subirats.
```

Information Extraction (IE)

- Example 2: incidents from free text (type of incident, perpetrator, target, date, location, effects, instrument).

At 5pm on Thursday, a white Fiat van veered off the road and into a crowd outside the Plaça de Catalunya metro station in Barcelona. The van continued down Las Ramblas for more than 500 metres while crashing into pedestrians. 13 people have been killed. 100 people were injured and 15 are in serious condition. Las Ramblas attacker Younes Abouyaaqoub was killed in Subirats.

type of incident = crash

location = Las Ramblas (Barcelona)

date = 17/8/2017

perpetrator = Younes Abouyaaqoub

target = pedestrians

instrument = white Fiat van

effects = 13 people killed, 100 people injured, 15 people in serious condition

Automatic Summarization

- E.g.: Generate biographies, minutes of a meeting, abstracts or extracts of written documents
- Given a document or a corpus, generate an extract or an abstract consisting of the most relevant content.
- Abstractive methods:
  - Generate new text from the conceptual representation of the important information contained in the input text.
  - Require language understanding and generation
- Extractive methods:
  - Select the most important sentences in the input text and produce a summary.
  - The set of sentences should maximize overall importance and coherency and minimize the redundancy.
- How are *importance* and *redundancy* computed?
- Semantics and ML techniques help

Question Answering (QA)

- E.g.: Questions answered by intelligent cars and rooms.
- Given a corpus, $D = \{D_i\}$, and a question, Q, extract the exact answer for Q from D.
  - Factoid QA: answers are exact facts

    E.g.: `Who was the president of the USA in 1987?`
  - Non-factoid QA: a definition, an explanation of how or why, a biography summary, ...

    E.g.: `Tell me what has been said so far in the meeting`
- Main subtasks:
  - Document indexing
  - Question processing (question type, question focus)
  - Answer extraction
- more in-depth NLP is required as well as ML techniques. Information extraction and Automatic Summarization help.

Machine Translation (MT)

- E.g.: Translation of written documents, help in human-human communication by mobile, online translation of broadcast news.

- Classical MT (Rule-based MT or Statistical-based MT):
  - Breaks input sentence into either words or phrases
  - Maps words to words or phrases to phrases using short context for taking decisions
- Neural MT:
  - Translates sentence to sentence
  - Uses the broader context of words and phrases at each step

- In general, the results are not comparable to human translation

Machine Translation (MT)

Examples of drawbacks: (with Google Translate)

- Working sentence by sentence: lack of context

  ES: Ana no aprobó el examen. Su amigo sí.
  EN: Ana did not pass the exam. **Your** friend **yes**.
  ok: Ana did not pass the exam. Her friend did.

- Lack of world-knowledge: Named entities

  ES: Disfrutar es el mejor nuevo restaurante de Europa
  EN: **Enjoy** is the best new restaurant in Europe
  ok: Disfrutar is the best new restaurant in Europe

- Restricted domains: terminology

  ES: El níscalo se cría bajo pinos
  EN: **The níscalo** grows under pines
  ok: Red pine mushroom grows under pines

  ES: Los níscalos se crían bajo pinos
  EN: **The chanterelles are raised** under pines
  ok: Red pine mushrooms grow under pines

Dialog Systems

- E.g.: chatbots, dialog-driven QA in smart cars and rooms, health-care assistance
- Help users to achieve specific goals by means of natural language interaction
- Main subtasks:
  - Interpreting user intervention
  - Determining the next system's action considering the user intention (answer a question, ask for more info, suggest alternatives, ...)
  - Generating system's intervention
- High complexity: Natural language understanding and generation is required

Outline

Schedule and Evaluation procedure

You can find the schedule at Racó, or directly here.

- Final exam: all the content, exam period
- Lab sessions:
  - Groups of 2 students (mandatory)
  - Deliverables for 5 tasks
- Final mark = 50% Exam + 50% Lab deliverables