

Master in Artificial Intelligence

Introduction

Human
Language
Technology
courses in
MAI

Introduction to Human Language Technologies



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat d'Informàtica de Barcelona



Outline

Introduction

Human
Language
Technology
courses in
MAI

1 Introduction

- What is Human Language Technology?
- Which is the general strategy for computing Human Language?
- Why is Human Language difficult to be processed?
- Examples of applications

2 Human Language Technology courses in MAI

- HLT branch
- IHL

Outline

Introduction

What is Human
Language
Technology?

Human
Language
Technology
courses in
MAI

1 Introduction

- What is Human Language Technology?
- Which is the general strategy for computing Human Language?
- Why is Human Language difficult to be processed?
- Examples of applications

2 Human Language Technology courses in MAI

- HLT branch
- IHL

Definition

Introduction

What is Human
Language
Technology?

Human
Language
Technology
courses in
MAI

- HLT is the technology focused on the study of human language from a computational point of view.
- HLT comprises computational methods, resources and models specifically designed to deal with all kind of text:
 - list of words
 - question in natural language
 - document in electronic format (e.g., plain text, web page, sms, tweet, oral transcriptions)
 - **corpus**: collection of documents in electronic format

Definition

Introduction

What is Human
Language
Technology?

Human
Language
Technology
courses in
MAI

- HLT is a multidisciplinary area:
 - **Natural Language Processing (NLP)**
 - Computational Linguistics
 - Artificial Intelligence
 - Speech Processing
 - Cognitive Science, Psychology
 - Logic, Mathematics

Outline

Introduction

Which is the general strategy for computing Human Language?

Human Language Technology courses in MAI

1 Introduction

- What is Human Language Technology?
- Which is the general strategy for computing Human Language?
- Why is Human Language difficult to be processed?
- Examples of applications

2 Human Language Technology courses in MAI

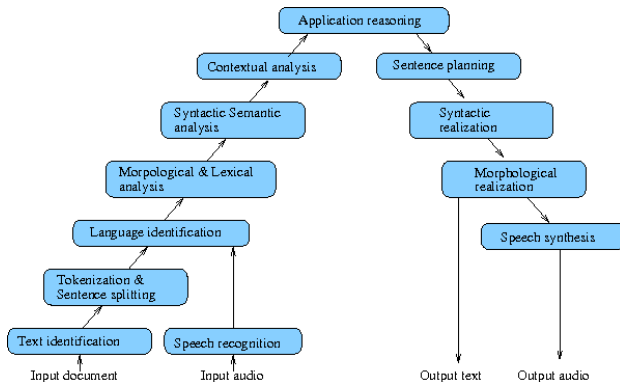
- HLT branch
- IHL

Definitions

The general strategy follows the standard subareas of linguistics:

- Phonetics: sounds of human speech.
E.g., *infrequent* → /ɪn'frikwənt/
- Morphology: structural formation of words.
E.g., *in-frequent-ly*.
- Syntax: structural relations between words in sentences.
E.g., *a determiner is followed by a common noun*.
- Semantics: meanings 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

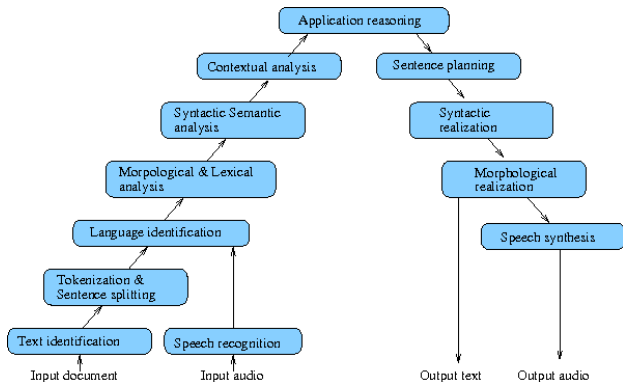


Introduction

Which is the general strategy for computing Human Language?

Human Language Technology courses in MAI

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)

Introduction

Which is the general strategy for computing Human Language?

Human Language Technology courses in MAI

Outline

Introduction

Why is Human
Language difficult to
be processed?

Human
Language
Technology
courses in
MAI

1 Introduction

- What is Human Language Technology?
- Which is the general strategy for computing Human Language?
- Why is Human Language difficult to be processed?
- Examples of applications

2 Human Language Technology courses in MAI

- HLT branch
- IHL

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''

Introduction

Why is Human
Language difficult to
be processed?

Human
Language
Technology
courses in
MAI

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 pragmatic | possessive or dative pronoun Esther or Alice |
| duck | synt-sem | noun or verb |

Outline

Introduction

Examples of
applications

Human
Language
Technology
courses in
MAI

1 Introduction

- What is Human Language Technology?
- Which is the general strategy for computing Human Language?
- Why is Human Language difficult to be processed?
- **Examples of applications**

2 Human Language Technology courses in MAI

- HLT branch
- IHL

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)

Introduction

Examples of
applications

Human
Language
Technology
courses in
MAI

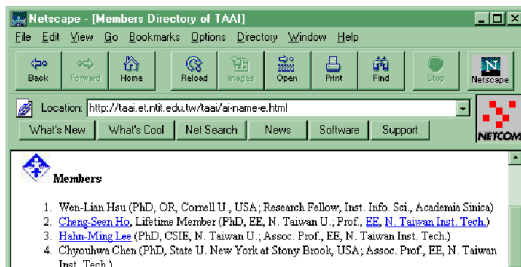
- E.g.: Searchers (Google, Yahoo, ...)
- Given a corpus, $D = \{D_i\}$, and a user query (list of words), Q , provide $\hat{D} \subset D$ that better match Q .
- $\text{sim}(v(Q), v(D_i))$, where $v(X)$ represents X in a vector space
- What vector space seems better?
 - words? $Q = \text{"window"}$, $D_i = \text{"... he closed the windows..."}$
 - lemmas? $Q = \text{"window"}$, $D_i = \text{"... he closed Windows..."}$
 - compounds? $Q = \text{"Energie"}$, $D_i = \text{"... Sonnenenergie..."}$
 - ...
 - 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-Lian 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 |
| ... | | | |

Introduction

Examples of
applications

Human
Language
Technology
courses in
MAI

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

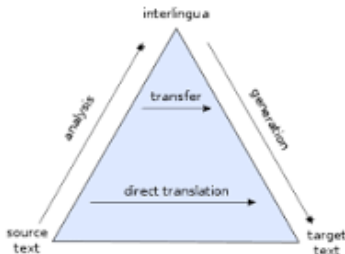
- E.g.: Generate biographies, minutes of a meeting, abstracts or extracts of written documents
- Given a document or a corpus, generate an extract or an abstract consisting of the most relevant content.
- Abstractive methods:
 - Generate new text from the conceptual representation of the important information 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.
- Different MT models differ from the level of NLP they use:



- Transfer model is the most frequently used
- 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 níscolo se cría bajo pinos

EN: **The níscolo** grows under pines

ok: Red pine mushroom grows under pines

ES: Los níscolos se crían bajo pinos

EN: **The chanterelles are raised** under pines

ok: Red pine mushrooms grow under pines

Dialog Systems

Introduction

Examples of
applications

Human
Language
Technology
courses in
MAI

- 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

Introduction

Human
Language
Technology
courses in
MAI

1 Introduction

- What is Human Language Technology?
- Which is the general strategy for computing Human Language?
- Why is Human Language difficult to be processed?
- Examples of applications

2 Human Language Technology courses in MAI

- HLT branch
- IHL

Outline

Introduction

Human
Language
Technology
courses in
MAI

HLT branch

- 1 Introduction
 - What is Human Language Technology?
 - Which is the general strategy for computing Human Language?
 - Why is Human Language difficult to be processed?
 - Examples of applications

- 2 Human Language Technology courses in MAI
 - HLT branch
 - IHL

IHLT AHLT HLE

Introduction

Human
Language
Technology
courses in
MAI

IHLT branch

- **IHLT**: the foundations of NLP interpretation, focusing on possible simple applications (spelling correction, text classification, paraphrase detection, text anonymization, ...)
- **AHLT**: more in-depth study of ML techniques for NLP interpretation (especially for syntactic and semantic parsing)
- **HLE**: review of complex applications of HLT (MT, IE, QA, Summ, Dialog)

Outline

Introduction

Human
Language
Technology
courses in
MAI
IHL

1 Introduction

- What is Human Language Technology?
- Which is the general strategy for computing Human Language?
- Why is Human Language difficult to be processed?
- Examples of applications

2 Human Language Technology courses in MAI

- HLT branch
- IHL

Content

Introduction

Human
Language
Technology
courses in
MAI
IHL

Here you can find the schedule of the course. [▶ LINK](#)
You can also find the URL of this schedule in the Racó.

Evaluation procedure

Introduction

Human
Language
Technology
courses in
MAI
IHL

- Final exam: all the content, exam period
- Lab sessions: groups of 2 students
 - Development of one project
 - Some deliverables of lab exercises
- Final mark = 50% Exam + 40% Project + 10% Lab deliverables