Advanced Human Language Techonoliges

Introduction

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Outline

Human Language Technologies

Building machines able to interact in human language is a hard (and unsolved) task, which requires inputs from many areas. Largely multidisciplinary.

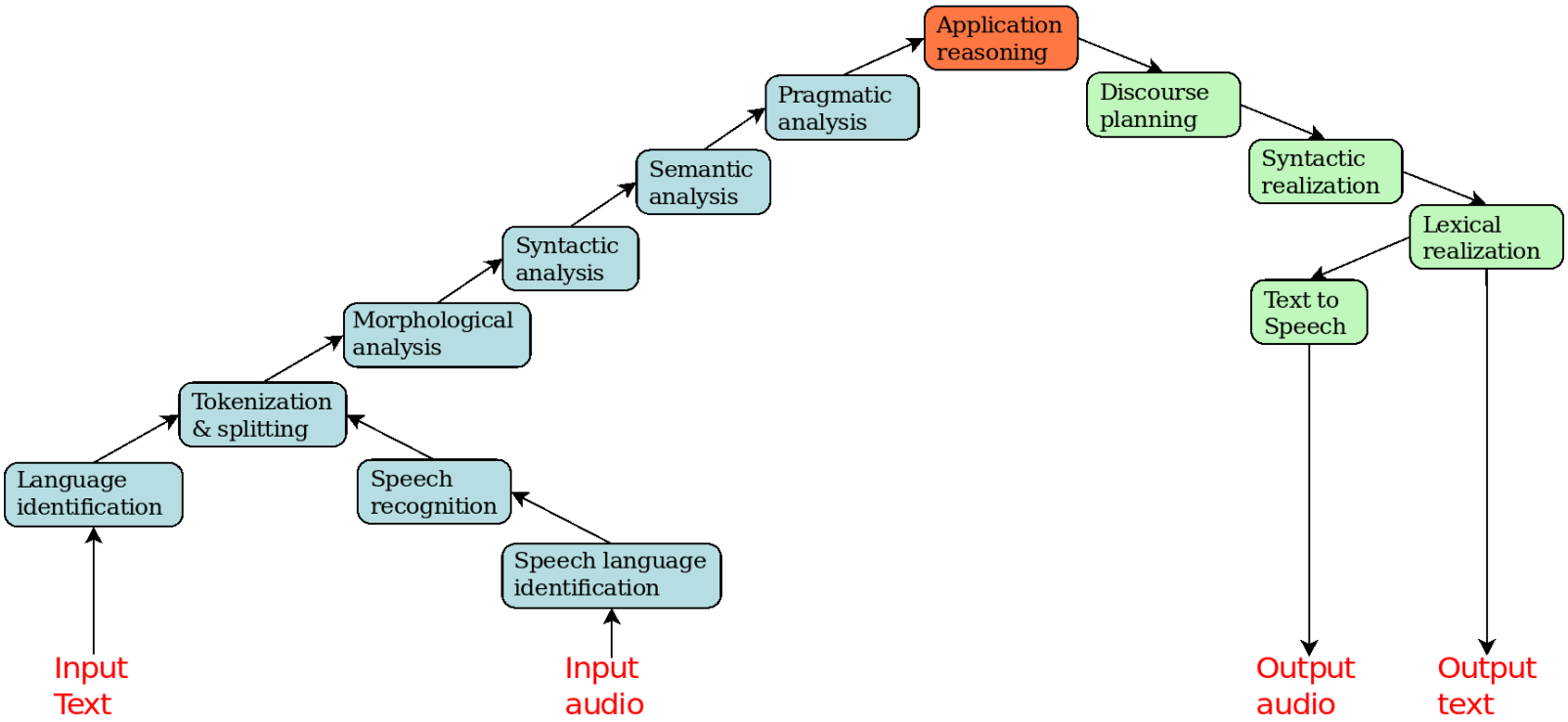
- Linguistics, Corpus Linguistics, Computational Linguistics, Phonetics.
- Artificial Intelligence, Machine Learning, Natural Language Processing.
- Signal Processing, Speech Processing.
- Cognitive Science, Psycholinguistics.
- Neurosciences.

Human Language Technologies at a Glance

As in any other engineering field, the usual approach is dividing the problem in simpler subproblems.

- 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*.
  Sentence word order is S-V-O.
- Semantics: meanings of words and their composition via syntax.
  E.g., *the president of USA is Joe Biden*
  → `president(USA, Joe_Biden)`
- Pragmatics: meaning in the context.
  E.g., **He** *is very well known in* **his country** [sarcasm].
  Could you tell me the time?.

Human Language Technologies at a Glance



- 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)

HLT Challenges: Ambiguity

Most efforts in NLP are devoted to solve different ambiguity levels

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
duck	morphosyntactic	noun / verb
her	syntactic	possessive / dative pronoun
make	semantic	cook / create / cause / convert

Outline

HLT Approaches

- **Rule-based systems**: Humans encode knowledge in rules, programs, or databases, which are used by the system to solve the target task.
- **Data-based systems**: (Statistical/Machine Learning/Neural): Humans provide the system with solved examples of the target task, and the system should infer its own model/rules, later used to solve the task.
- **Hybrid systems**: (Part of) the knowledge is encoded by humans, but the system learns how to use or weight it.

HLT Approaches

Rule-based	⟶	**ML/NN-based**	⟶	**LLM-based**
++ control		- control		- - control
++ precision		+ precision		++ precision
- - recall		+ recall		++ recall

Outline

Examples of applications

- Document similarity / clustering (related news, plagiarism, ...)
- Document classification (e.g. anti-spamming, email routing, sentiment polarity, ...)
- Information Retrieval
- Text correction
- Information Extraction
- Automatic Summarization
- Question Answering
- Machine Translation
- Dialog Systems
- ...

Outline

AHLT Content

Part I: Classical approaches

- Language modelling. Estimation. MLE and MEM models
- Words: Lexical similarity - distributional semantics
- Word Sequences: NERC - CRF
- Sentences: Constituent parsing, dependency parsing.

Part II: Deep Learning approaches

- Words: Lexical semantics, word embeddings.
- Sequence labelling: PoS, NERC. LSTM, LSTM+CRF
- Sentence level: Recurrent NN. sequence-to-sequence models.
- Transformers. Attention.

Evaluation

- Final exam: all the content, exam period
- Lab sessions: groups of 2 students
  - Implementation of three tasks on medical documents.
  - Deliverables: One short report per task
- Final mark = 50% Exam + 50% Lab

https://www.cs.upc.edu/~padro/ahlt/ahlt.html