Master in Artificial Intelligence

Advanced Human Language Technologies

Introduction

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Human Language Technologies

- **Linguistics** Study of human language. Traditionally by introspection or interviewing native speakers. Today increasingly based on data.
- **Corpus Linguistics** Study of human language using as main information source big amounts of language usage data, either written or spoken (corpus).
- **Computational Linguistics** Study of human language based on the development of formal and computable models for language.
- **Natural Language Processing** (NLP) Development of systems able to automatically process human language (usually regardless of whether they explain language behaviour or not).
- **Human Language Technologies** (HLT) Broader (and fancier) term that embraces NLP, NL generation, speech recognition & synthesis, Information Retrieval, ...

HLT is multidisciplinary

Buildig machines able to interact in human language is a hard (and unsolved) task, and requires inputs from many areas:

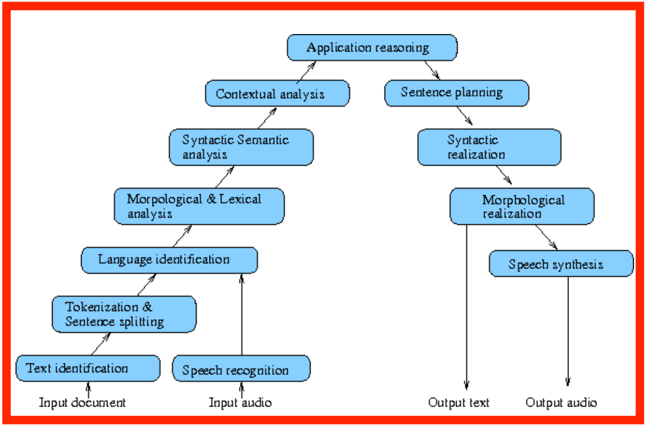
- (Computational) Linguistics
- Artificial Intelligence, Machine Learning
- Phonetics
- Speech Processing
- Cognitive Science, Psycholinguistics

Human Language Technologies at a Glance

As in any other engineering field, the 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 Donald Trump* →
  `president(USA, Donald_Trump)`
- 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

- AI-Completeness: To be able to handle language like a human requires world knowledge and common sense
- Multilinguality: Different languages require different models, resources, and data. Speakers often use words from other languages. *Sayonara, baby.*
- Evaluation: It is not always easy to (automatically) assess the performance of HLT systems. E.g. Correctness/suitability of a translation/summary
- Variability: Many different ways to express the same meaning: *where can I get a map? / I need a map / need map*
- Ambiguity: The same sentence may have different meanings: *I made her duck*

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

HLT Approaches

- **Rule-based systems**: Humans encode knowledge in rules, programs, or databases, which are used by the system to solve the target task.
- **Statistical/Machine Learning systems**: 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.

Rule-based *vs* Statistics/Machine Learning

Language is a collection of statistical distributions:

- Language evolves: (*ale* vs. *eel*, *while* as Adv vs. Noun, *near* as Prep vs. Adj)
- Language varies across locations: Dialect continuum (e.g. Inuit)
- Language varies among individuals: age, education, monolinguism, ...
- Structural ambiguity

  Our company is training workers
  Our problem is training workers
  Our product is training wheels

  Parker saw Mary
  The a are of I

Rule-based *vs* Statistics/Machine Learning

- Rule-based systems are costly (and difficult) to scale up from small/domain specific applications to wide-coverage systems.
- Rule-based systems allow fine-tuning and strict control of system behavior.
- Statistical/ML systems require a lot of training data that may not be available... (and Zipf's Laws make it hard)
- Statistical/ML systems can deal better with ambiguity (since they can compute which interpretation is more likely).

- Rule-based or hybrid systems are a good choice for some applications (e.g. restricted domain chatbots).

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
- . . .

HLT courses in MAI

- **IHLT**: Foundations of NL processing, focusing on possible simple applications (spelling correction, text classification, paraphrase detection, text anonymization, . . .)
- **AHLT**: More in-depth study of ML techniques for NLP interpretation: Classical ML and Dep Learning approaches.
- **HLE**: Review of high-level applications of HLT (MT, IE, QA, Summarization, Dialog, etc.)

AHLT Content (1)

Part I: Classical approaches

- Statistical models of language. MLE Estimation and smoothing. Maximum entropy estimation. Log-linear models
- Similarity models. Lexical semantics. Distributional semantics.
- Sequence prediction. NERC. Local Classifiers, HMMs, Global predictors, Log-linear models, CRFs.
- Sentence level: Constituent parsing, dependency parsing.

AHLT Content (2)

Part II: Deep Learning approaches

- Preliminaries
- Words: Lexical semantics, word embeddings.
- Sequence prediction: PoS, NERC. LSTM, LSTM+CRF
- Sentence level: Recurrent NN. sequence-to-sequence models.
- Document level: CNNs. Document classification, document similarity. Document embeedings.

Evaluation procedure

- 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 + 50% Lab