Welcome to the course!

Introduction to Natural Language Processing (NLP)

Professors: Marta Gatius Vila
Horacio Rodríguez Hontoria

Hours per week: 2h theory + 1h laboratory

Web page:

Main goal
Understand the fundamental concepts of NLP

• Most well-known techniques and theories
• Most relevant existing resources
• Most relevant applications
Welcome to the course!

Introduction to Natural Language Processing

Content

1. Introduction to Language Processing
2. Applications.
3. Language models.
5. Syntactic processing.
7. Generation
Welcome to the course!
Introduction to Natural Language Processing

Assessment

• Exams
  Mid-term exam - November
  End-of-term exam – Final exams period - all the course contents

• Development of 2 Programs – Groups of two or three students

Course grade = maximum ( mid-term exam grade * 0.15 + end-of-term exam grade * 0.45, end-of-term exam grade * 0.6) + assignments grade * 0.4
Welcome to the course!
Introduction to Natural Language Processing

Related (or the same) disciplines:
• Computational Linguistics, CL
• Natural Language Processing, NLP
• Linguistic Engineering, LE
• Human Language Technology, HLT
Linguistic Engineering (LE)

• LE consists of the application of linguistic knowledge to the development of computer systems able to recognize, understand, interpretate and generate human language in all its forms.

• LE includes:
  • Formal models (representations of knowledge of language at the different levels)
  • Theories and algorithms
  • Techniques and tools
  • Resources (Lingware)
  • Applications
Linguistic knowledge levels

- Phonetics and phonology. Language models
- Morphology: Meaningful components of words. Lexicon
  
  *doors is plural*

  
  - Syntax: Structural relationships between words. Grammar
    
    *an utterance is a question or a statement*

  - Semantics: Meaning of words and how they combine. Grammar, domain knowledge
    
    *open the door*

  - Pragmatics: How language is used to accomplish goals. Domain and Dialogue Knowledge
    
    *to be polite*

- Discourse: How single utterances are structured. Dialogue models
Examples of applications involving language models at those different levels

- Intelligent agents (e.g., HAL from the movie 2001: A space Odyssey)
- Web-based question answers
- Machine translation engines

Foundations of LE lie in:

• Linguistics, Mathematics, Electrical engineering and Psychology
Exciting time because of

• The increase of computer resources available
• The rise of the Web (a massive source of information)
• Wireless mobile access
• Intelligent phones

Revolutionary applications are currently in use

• Conversational agents for making travel reservations
• Speech systems for cars
• Cross-language information retrieval and translation
• Automate systems to analyze students essays
Components of the Technology

INPUT

Recognize and Validate → Analyze and Understanding → Apply → Generate

OUTPUT

TEXT SPEECH IMAGE

LINGUISTIC RESOURCES

TEXT SPEECH IMAGE

NLP Introduction
This course is focused on *Language Understanding*

- Different levels of understanding
  - Incremental analysis
  - Shallow and partial analysis
  - Looking for the focus of interest (spotting)
  - In depth analysis of the focus of interest
- Linguistic, statistical, machine learning, hybrid approaches
- Main problems: ambiguity, unseen words, ungrammatical text
Language Generation

• Content planning
  • Semantic representation of the text
  • What to say, how to say

• Form planning
  • Presentation of content
  • Using rhetorical elements
Dialogue

• Need of a high level of understanding
• Involve additional processes
• Identification of the illocutionary content of speaker utterances
• Speech acts
  • assertions, orders, askings, questions, etc.
• Direct and indirect speech acts
NLP Challenges

• Why NLP is difficult?
  • Language is alive (changing)
  • Ambiguity
  • Complexity
  • Knowledge is imprecise, probabilistic, fuzzy

• World knowledge (common sense) is needed
  • Language is embedded into a system of social interaction
NLP Challenges

Ambiguity

• Phonetical ambiguity
• Lexical ambiguity
• Syntactic ambiguity
• Semantic ambiguity
• Pragmatic ambiguity.

References
Resolving ambiguous input

• Multiple alternative linguistic structures can be built
  – *I made her duck*
    • *I cooked waterfowl for her*
    • *I cooked waterfowl belonging to her*
    • *I created the (plaster?) duck she owns*
    • *I caused her to quickly lowed her head or body*
    • *I waved my magic wand and turned her into undifferentiated waterfowl*

  – *Ambiguities in the sentence*
    • *Duck* can be noun(waterfowl) or a verb (go down) -> syntactic and semantic ambiguity
    • *Her* can be a dative pronoun or a possessive pronoun -> syntactic ambiguity
    • *Make* can be create or cook -> semantic ambiguity
LEXICAL AMBIGUITY

• There are several words that have more than one possible meaning (polysemous)
• Frequent words are more ambiguous
SYNTACTIC AMBIGUITY

- Grammars are usually ambiguous
- Usually, more than one parsed tree is correct for a sentence given a grammar
- Some kind of ambiguity (as *pp-attachment*) is at some level predictable
NLP Challenges

SEMANTIC AMBIGUITY

• More than one semantic interpretation is possible for a given sentence

• *Peter gave a cake to the children*
  • One cake for all them?
  • One cake for each?
Pragmatic ambiguity. Reference

- More than one semantic interpretation is possible for a given text. References between sentences.
- *Later he asked her to put it above*
- Later? When?
  - He?
  - Her?
  - It?
  - Above what?
Pragmatic Ambiguity

3. Find x.

Here it is

Ocular Trauma - by Wade Clarke ©2005
Pragmatic Ambiguity(II)

\[ (a + b)^n = \binom{n}{0}a^n b^0 + \binom{n}{1}a^{n-1}b^1 + \binom{n}{2}a^{n-2}b^2 + \ldots + \binom{n}{n}a^0 b^n = a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \ldots + \binom{n}{n}b^n \]

\[ 2^{x+2} = 2^x + 4 \]

"Very funny, Peter."
Which kind of ambiguity?

After explaining to a student through various lessons and examples that:

$$\lim_{x \to 8} \frac{1}{x-8} = \infty$$

I tried to check if she really understood that, so I gave her a different example. This was the result:

$$\lim_{x \to 5} \frac{1}{x-5} = 0$$
Resolving ambiguous input

- Using models and algorithms
- Using data-driven methods
- Semantic-guided processing
  - Restricting the domain. Considering only the language needed for accessing several services
  - Using context knowledge
    (Shallow or Partial analysis)
Two type of models

**Rationalist model.** Noam Chomsky
- Most of the knowledge needed for NLP can be acquired previously, prescripted and used as initial knowledge for NLP.

**Empiricist model.** Zellig Harris
- Linguistic knowledge can be inferred from the experience, through textual corpora by simple means as the association or the generalization.
- Firth “*we can know a word by the company it owns*”
Levels of linguistic description

- Phonetics
- Phonology
- Morphology. Lexical
- Sintax
- Semantics. Logical
- Pragmatics
- Discourse
Several formal models and theories:

- State machine
- Rule systems
- Logic
- Ontologies
- Probabilistic models
- Vector space models
State Machines

• Formal models that consists of state, transitions and input representations

• Variations
  • Deterministic/non deterministic
  • Finite-state automata
  • Finite-state transducers
Rule Systems

• Grammar formalisms
  • Regular grammars
  • Context free grammars
  • Feature grammars

• Probabilistics variants of them
• Used for phonology, morphology and syntax
Logic

• First order logic (Predicate calculus)
• Related formalism
  • Lambda calculus
  • Features structures
  • Semantic primitives
• Used for modelling semantics and pragmatics but also for lexical semantics
Probabilistic Models

• State machine, rule systems and logic systems can be augmented with probabilities.

• State machine augmented with probabilities become Markov model and hidden Markov model.
  • Used in different processes: part-of-speech tagging, speech recognition, dialogue understanding, text-to-speech and machine translation.
• Ability to solve ambiguity problems
Vector-space Models

- Based on linear algebra
- Underlies information retrieval and many treatment of word meaning
Architecture of NLP systems

• Architecture based on layers
  ✷ Each layer owns specific classes in charge of solving some problems.
  ✷ The objects of a layer request services to other objects from the same layer or from the layer of the immediate inferior level.
  ✷ The objects of a layer provide services to other objects from the same layer or from the layer of the immediate superior level.

• Architecture based on pipes & filters
  ✷ Each filter enriches the input stream and send it to the output stream
NLP Tasks

• Three levels of increasing complexity
  • **Basic level.** Basic tasks: (paragraph, sentence) Segmenters, language identifiers, Name Entity Recognizer (NER), Name Entity Classification (NEC), NERC, tokenizers, morphological analyzers, taggers, parsers, chunkers, semantic analyzers.

  • **Intermediate level.** Tasks implying the performance of basic components:
    • Document classification, summarization, IE, IR.

  • **Application level.**
    • Conversational systems
Language Processing and Intelligence

The ability to process language is related to the intelligence of the machines.

Turing test (1950) consists of convincing the interrogator the machine is a person (The machine tries to answer questions as a human would).

Q: Please write me a sonnet on the topic of the Forth Bridge

A: Count me on this one. I never could write poetry.

Q: Add 34957 to 70764

A: (Pause) 105621
Language Processing and Intelligence

- Eliza program (Weizenbaum, 1966)
- Very simple program (based on pattern-matching).
- I does not understand humans but it seems it does.

User: You are like my father in some ways
Eliza: what resemblance you see
User: you don’t argue with me
Eliza: Why do you think I don’t argue with you
User: You are afraid of me
Eliza: Does it please to belive I am afraid of you?
Loebner Prize competition based on Turing test. Some programs fool judges some of the time (Shieber, 1994)

There are fun web robots based trying to look human (Alice)

There are dialogue systems (conversational systems) helping people use many different type of applications
Relevant Resources

• Conference and journals focus on LE: ACL, EACL, COLING, AI conferences.
• Competitions: TREC, CLEF, MUC, ACE, TAC
• Available resources:
  – Corpora, Ontologies
  – WordNet, EuroWordNet, Balkanet,
  – FrameNet, VerbNet, Propbank, OntoNotes
Resources for Language Understanding

• General Lexicons
• Dictionaries
• Specialized Lexicons
• Ontologies
• Grammars
• Textual Corpora
• Internet as an information source
General Lexicons

• Word repositories
  • Lemmaries, formaries, lists of words, phrasal lexicons

• Knowledge on words
  • Phonology
  • Morphology: POS, agreement
  • Syntax: category, sub-categorization, subcategorization, argument structure, valency, co-occurrence patterns

• Semantics: semantic class, selectional restrictions
  – Pragmatics: use, register, domain
Dictionaries

• MRDs (Machine Readable Dictionaries)
• Types: general, normative, learner, mono/bilingual
• Size, content, organization
  • entry, sense, relations,
• Lexical databases
  • e.g. Acquilex LDB
• Other sources: enciclopaedias, thesaurus
Specialized Lexicons

• Onomasticae
• Terminological databases
• Gazetteers
• Dictionaries of locutions, idioms
• Wordnets
• Acronyms, idioms, jaergon
• Date, numbers, quantities+units, currencies
Morpholexical Relations
U. Las Palmas (Santana)
Example: Using Gazetteers in Q&A systems

- Multitext (U.Waterloo)
  - Clarke et al, 2001, 2002
    - Structured data
      - Biographies (25,000), Trivial Q&A (330,000), Country locations (800), acronyms (112,000), cities (21,000), animals (500), previous TREC Q&A (1393), ...
    - 1 Tb of Web data
  - Altavista
- AskMSR (Microsoft)
  - Brill, 2002
Grammars

• Morphological Grammars
• Syntactic Grammars
  • constituent
  • dependency
  • case
  • transformational
  • systemic
• Phrase-structure vs Unification Grammars
• Probabilistic Grammars
• Coverage, language, tagsets
Ontologies

• Lexical vs conceptual ontologies
• General vs domain restricted ontologies
• Task ontologies, metaontologies
• Content, granularity, relations
• Interlinguas: KIF, PIF
• CYC, Frame-Ontology, WordNet, EuroWordNet, GUM, MikroKosmos
• Protegeé
Raw Corpora

• Textual vs Speech
• Size (1Mw - 1Gw - 1TW)
• Few structure (if any)
• Provide information not available in a more treatable way:
  • colocations, argumental structure, context of occurrence, grammatical induction, lexical relations, selectional restrictions, idioms, examples of use
Tagged Corpora

• Pos tagged (all tags are disambiguated)
• Lemma
• Sense (granularity of tagset, WN)
• Parenthised
  • parsed
• Parallel corpora
• Balanced, pyramidal, opportunistic corpora
Some examples of Corpora

- Brown Corpus
- ACL/DCI (Wall Street Journal, Hansard, ...)
- ACL/ECI (European Corpus Initiative)
- USA-LDC (Linguistic Data Consortium)
- LOB (ICAME, International Computer Archive of Modern English)
- BNC (British National Corpus)
- SEC (Lancaster Spoken English Corpus)
- Penn Treebank
- Susanne
- SemCor
- Trésor de la Langue Française (TLF)
Some examples of Spanish Corpora 2

- Oficina del Español en la Sociedad de la Información OESI
  - http://www.cervantes.es/default.htm
- CREA, RAE. 200 Mw.
- CRATER, (sp, en, fr), U.A.Madrid. 5.5Mw. aligned, POS tagged
- ALBAYZIN. Speech, isolated sentences, queries to a geographic database
- LEXESP, 5Mw, Pos taged, lemmatized
- Ancora, Spanish & Catalan, Extremely rich annotation, 500Kw
- IEC in the framework of DCC (catalan)
Internet as an information source

- Huge volume
  - > 2,000 Million pages, tenths of Tb,
  - expansion (doubles size each two years)
- Heterogeneity
  - content, language (70% English), formats
  - redundancy
  - Hidden Web
- General Information servers
  - (Medialinks)
  - 14,000 servers (5,000 newspapers, 70 in Spain)
Internet as an information source

- Internet today
  - Documents HTML
  - Built for human use (visualization)
  - Pages automatically generated by applications
  - Access through
    - known URLs
    - searchers of general purpose
    - specific searchers for a site

- Limitations
  - Access (by applications) to HTML codified text
  - Building (and maintaining!) wrappers
Internet as an information source

- Web2.0
- Software agents
  - crawlers, spiders, softbots, infobots ...
- Wacki
  - Baroni, 2008
- Wikipedia
Applications

• Two main areas
  • Massive management of textual information sources
    • for human use
    • for automatic collection of linguistic resources
  • Person/Machine interaction
Massive management of textual information sources

- Machine Translation
- Information Management
  - Automatic Summarization
  - Information \{Retrieval, Extraction, Filtering, Routing, Harvesting, Mining\}
  - Document Classification
  - Question Answering
  - Conceptual searchers
Automatic collection of linguistic resources

• Aligned corpora (various levels)
• Grammars
• Gazetteers
• Resources including
  – Morphology bases
  – Selectional restrictions
  – Subcategorization patterns
  – Topic Signatures