• Course outline

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

• Linguistic Engineering
  • Tasks
  • Resources
  • Applications

• Levels of Linguistic processing and description
From NLP to LE

• LE consists of the application of Linguistic Knowledge to the development of computer systems able to recognizing, understanding, interpretation and generation of human language in all its forms.

• LE includes:
  • Methods, Techniques and Tools
  • Resources (Lingware)
  • Applications
Components of the Technology

INPUT

LINGUISTIC RESOURCES

Recognize and Validate

Analyze and Understanding

Apply

Generate

OUTPUT

TEXT

SPEECH

IMAGE

TEXT

SPEECH

IMAGE

Components of the Technology

Recognize and Validate

Analyze and Understanding

Apply

Generate

INPUT

LINGUISTIC RESOURCES

OUTPUT

TEXT

SPEECH

IMAGE
Main used techniques

- Speaker identification, Verification
- Speech recognition
- Character and image recognition
- Language Understanding
- Language generation
- Dialog-based systems
- Speech synthesis
• **Problems:**
  • Unknown speaker
  • Noise (in the environment or in the channel)
  • Temporal changes in the speaker
Speech recognition

- Discourse
  - discret vs contiguous
- Acquisition and use of statistical models of phonemes and words
- Independence of the speaker (without need of training)
- Noise
- Accent, dialects, non-grammaticalities, OOV
Character and image recognition

- Printed Character recognition (OCR)
  - unknown font
  - bad quality of text
- Manual Character recognition (ICR)
- Image recognition
  - Multimedia documents.
    - e.g Web pages
  - Extraction of multimedia fragments (text, photos, ...)
- Text analysis:
  - Titles, headings, captions, paragraphs, etc...
Language Understanding

• Level of understanding

• Incremental analysis:
  • shallow and partial analysis
  • Looking for the interest focus (spotting)
  • In depth analysis of interest focus

• Linguistic, statistical, ML, hybrid approaches

• Semantic models

• Unseen words
Language generation

• Semantic representation of the text
• what to say, how to say
• Content planning, form planning
• Rethorical elements
Dialog-based systems

- Need of a high level of understanding
- Dialog quality
- Identification of the illocutionary content of speaker utterances
- Locutive and non-locutive speech acts
  - assertions, orders, askings, questions, etc...
- Direct and indirect speech acts
Speech synthesis

• Content
• Form: intensity, volume, entonation, duration
• Generation from fragments pre-built
Concatenation
Resources

- 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..
  - Sintax: category, sub-categorization, subcategorización, argument structure, valency, co-occurrence patterns…
  - Semantics: semantic class, selectional restrictions…
  - Pragmatics: use, register, domain, …
Dictionaries

- MRDs
- types: general, normative, learner, mono/bilingual...
- size, content, organization
  - entry, sense, relations, ...
- Lexical databases
  - ej. Acquilex LDB
- Other sources: enciclopaedias, thesaurus,...
Specialized Lexicons

- Onomasticae
- terminolical databases
- Gazetteers
- dictionaries of locutions, idioms,…
- Wordnets
- Acronyms, idioms, jaergon
- Date, numbers, quantities+units, currencies…
morpholexical relations. U. Las Palmas (O. Santana)
Ej: 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 de 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
- Protegé
Raw Corpora

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

- pos tagged (all tags or disambiguated)
- lemma
- sense (granularity of tagset, WN)
- parenthesised
  - parsed
- Paralel 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

- 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 tagged, lemmatized
- Ancora, Spanish & Catalan, Extremely rich annotation, 500Kw
- IEC in the framework of DCC (catalan)
example of Ancora treebank

Surprendente esa actitud activa en un hombre que está alcanzando en esta Vuelta mucho más de lo que nadie, excepto quizá 1, podía presumir.
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)
  - Many pages automatically generated by applications
- Access through
  - known URLs
  - searchers (or meta-searchers) of general purpose
  - specific searchers for a site
- Limitations
  - access (by applications) to HTML codified text (often bad)
  - 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
Person/Machine interaction

- authoring
  - correctors
  - online access to dictionaries, tesaurus, ontologies
  - facilities for the composition, integration, translation, warehouse of documents
- NL interfaces (multilingual, multimodal, user friendly)
automatic collection of linguistic resources

- Aligned corpora (various levels)
- grammars
- gazetteers
- morphology
- selectional restrictions
- Subcategorization patterns
- Topic Signatures
What to read

   - Handbook of Natural Language Processing
     - Marcel Dekker, New York, 2000

   - Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
What to read

- C. Manning, H. Schütze (1999)
  - Foundations of Statistical Natural Language Processing
  - The Oxford handbook of Computational Linguistics
What to read

- Noah A. Smith (2011)
- Clark, A.; Fox, C.; Lappin, S. (eds.) (2012)
- Nitin Indurkhya, Fred J. Damerau (2010)
some useful links

- Organizations
  - Association of Computational Linguistics ACL
    - http://www.aclweb.org/
  - Information Society Technology IST
  - Sociedad Española para el procesamiento del lenguaje natural SEPLN
    - http://www.sepln.org
  - Oficina del Español en la Sociedad de la Información OESI
    - http://www.cervantes.es/default.htm
some useful links

- Organizations
  - TALP (UPC)
    - http://www.talp.cat/talp/
  - Grup de PLN de la UPC
    - http://nlp.lsi.upc.edu
some useful links

- Papers
  - ACL Anthology
    - http://aclweb.org/anthology-new/
  - The Python Papers Anthology
    - http://pythonpapers.org/
some useful links

• Tools
  • N LTK, Natural Language Toolkit
  • Python
    • [http://python.org/](http://python.org/)
  • OpenNLP
  • NLP resources of Stanford University
some useful links

- **Tools**
  - Mallet, toolbox in Java, built by Andrew McCallum for statistical NLP
    - [http://mallet.cs.umass.edu/index.php/Main_Page](http://mallet.cs.umass.edu/index.php/Main_Page)
  - WEKA, ML toolbox
  - Lingpipe
    - [http://alias-i.com/lingpipe/](http://alias-i.com/lingpipe/)
  - Laura Alonso's resource page