Lecture 1.
Introduction to Human Language Technologies

Tomaž Erjavec

Technicalities

- Lecturer: 
  http://nl.ijs.si/et/
  tomaz.erjavec@ijs.si
- Course homepage:
  http://nl.ijs.si/et/teach/mps08-hlt/
- Assessment:
  Seminar work
- Next Wednesday: introduction to datasets
- Exam dates

Introduction to Human Language Technologies

1. Application areas of language technologies
2. The science of language: linguistics
3. Computational linguistics: some history
4. HLT: Processes, methods, and resources
I. Applications of HLT

Speech technologies
- Speech synthesis
- Speech recognition
- Speaker verification (biometrics, security)
- Spoken dialogue systems
- Speech-to-speech translation
- Speech prosody: emotional speech
- Audio-visual speech (talking heads)

Machine translation
Perfect MT would require the problem of NL understanding to be solved first!

Types of MT:
- Fully automatic MT (babelfish, Google translate)
- Human-aided MT (pre and post-processing)
- Machine-aided HT (translation memories)

Problem of evaluation!
- Automatic (BLEU, METEOR)
- Manual (expensive!)
### MT approaches

- **rule based:** rules + lexicons
- **statistical:** parallel corpora

### Statistical MT

- parallel corpora: text in original language + translation
- on the basis of parallel corpora only: induce statistical model of translation
- very influential approach: now used in Google translate

### Information retrieval and extraction

- **Information retrieval (IR)** is the science of searching for documents, for information within documents and for metadata about documents.
  - "bag of words" approach
- **Information extraction (IE)** is a type of information retrieval whose goal is to automatically extract structured information, i.e. categorized and contextually and semantically well-defined data from a certain domain, from unstructured machine-readable documents.
- Related area: **Named Entity Extraction**
  - identify names, dates, numeric expression in text
II. Background: Linguistics

- What is language?
- The science of language
- Levels of linguistics analysis

Language

- Act of speaking in a given situation (parole or performance)
- The abstract system underlying the collective totality of the speech/writing behaviour of a community (langue)
- The knowledge of this system by an individual (competence)

De Saussure
(structuralism ~ 1910) parole / langue
Chomsky
(generative ling. > 1960) performance / competence

What is Linguistics?

The scientific study of language
- Prescriptive vs. descriptive
- Diachronic vs. synchronic
- Performance vs. competence
- Anthropological, clinical, psycho, socio,... linguistics
- General, theoretical, formal, mathematical, computational linguistics
### Levels of linguistic analysis

- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Discourse analysis
- Pragmatics
- + Lexicology

### Phonetics

- Studies how sounds are produced: methods for description, classification, transcription
- Articulatory phonetics (how sounds are made)
- Acoustic phonetics (physical properties of speech sounds)
- Auditory phonetics (perceptual response to speech sounds)

### Phonology

- Studies the sound systems of a language (of all the sounds humans can produce, only a small number are used distinctively in one language)
- The sounds are organised in a system of contrasts; can be analysed e.g. in terms of phonemes or distinctive features
- Segmental vs. suprasegmental phonology
- Generative phonology, metrical phonology, autosegmental phonology, ... (two-level phonology)
Distinctive features

<table>
<thead>
<tr>
<th>Feature</th>
<th>t</th>
<th>z</th>
<th>s</th>
<th>l</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>anterior</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>coronal</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>labial</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>distributed</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>consonantal</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>sonorant</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>voiced</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>approximant</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>continuant</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>nasal</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>strident</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Generative phonology

A consonant becomes devoiced if it starts a word:

[C, +voiced] → [-voiced] / #___

e.g. #vlak# → #flak#

- Rules change the structure
- Rules apply one after another (feeding and bleeding)
- (in contrast to two-level phonology)
Autosegmental phonology

- A multi-layer approach:
  
Morphology

- Studies the structure and form of words
- Basic unit of meaning: morpheme
- Morphemes pair meaning with form, and combine to make words:
  e.g. dogs ← dog/DOG,Noun + -s/plural
- Process complicated by exceptions and mutations
- Morphology as the interface between phonology and syntax (and the lexicon)

Types of morphological processes

- Inflection (syntax-driven):
  run, runs, running, ran
  gledati, gledam, gleda, glej, gledal,...
- Derivation (word-formation):
  to run, a run, runny, runner, re-run, ...
  gledati, zagledati, pogledati, pogled, ogledal,...
- Compounding (word-formation):
  zvezdogled,
  Herzkreislaufwiederbelebung
Inflectional Morphology

- Mapping of form to (syntactic) function
- dogs → dog + s / DOG [N, pl]
- In search of regularities: talk/walk; talks/walks; talked/walked; talking/walking
- Exceptions: take/took, wolf/wolves, sheep/sheep
- English (relatively) simple; inflection much richer in e.g. Slavic languages

Macedonian verb paradigm

The declension of Slovene adjectives
Characteristics of Slovene inflectional morphology

- Paradigmatic morphology: fused morphs, many-to-many mappings between form and function:
  - hodil-a[masculine dual], stok-a[singular, genitive,], sosed-
  - Complex relations within and between paradigms: syncretism, alternations, multiple stems, defective paradigms, the boundary between inflection and derivation,...
- Large set of morphosyntactic descriptions (>1000) Ncsmn, Ncmsg, Ncmpn,...
- MULTTEXT-East tables for Slovene

Syntax

- How are words arranged to form sentences?
  - I milk like
  - I saw the man on the hill with a telescope.
- The study of rules which reveal the structure of sentences (typically tree-based)
- A "pre-processing step" for semantic analysis
- Common terms:
  - Subject, Predicate, Object,
  - Verb phrase, Noun phrase, Prepositional phr., Head, Complement, Adjunct,...

Syntactic theories

- Transformational Syntax
  - N. Chomsky: TG, GB, Minimalism
- Distinguishes two levels of structure: deep and surface; rules mediate between the two
- Logic and Unification based approaches ('80s) : FUG, TAG, GPSG, HPSG, ...
- Phrase based vs. dependency based approaches
Example of a phrase structure and a dependency tree

Semantics

- The study of meaning in language
- Very old discipline, esp. philosophical semantics (Plato, Aristotle)
- Under which conditions are statements true or false; problems of quantification
- The meaning of words - lexical semantics

spinster = unmarried female → * my brother is a spinster

Discourse analysis and Pragmatics

- Discourse analysis: the study of connected sentences - behavioural units (anaphora, cohesion, connectivity)
- Pragmatics: language from the point of view of the users (choices, constraints, effect; pragmatic competence; speech acts; presupposition)
- Dialogue studies (turn taking, task orientation)
Lexicology

- The study of the vocabulary (lexis / lexemes) of a language (a lexical "entry" can describe less or more than one word)
- Lexica can contain a variety of information: sound, pronunciation, spelling, syntactic behaviour, definition, examples, translations, related words
- Dictionaries, mental lexicon, digital lexica
- Plays an increasingly important role in theories and computer applications
- Ontologies: WordNet, Semantic Web

III. The history of Computational Linguistics

- MT, empiricism (1950-70)
- The Generative paradigm (70-90)
- Data fights back (80-00)
- A happy marriage?
- The promise of the Web

The early years

- The promise (and need!) for machine translation
- The decade of optimism: 1954-1966
- *The spirit is willing but the flesh is weak*
- *The vodka is good but the meat is rotten*
- ALPAC report 1966: no further investment in MT research; instead development of machine aids for translators, such as automatic dictionaries, and the continued support of basic research in computational linguistics
- also quantitative language (text/author) investigations
The Generative Paradigm

Noam Chomsky's Transformational grammar: Syntactic Structures (1957)

Two levels of representation of the structure of sentences:
- an underlying, more abstract form, termed 'deep structure',
- the actual form of the sentence produced, called 'surface structure'.

Deep structure is represented in the form of a hierarchical tree diagram, or "phrase structure tree," depicting the abstract grammatical relationships between the words and phrases within a sentence.

A system of formal rules specifies how deep structures are to be transformed into surface structures.

Phrase structure rules and derivation trees

S → NP V NP
NP → N
NP → Det N
NP → NP that S

Characteristics of generative grammar

- Research mostly in syntax, but also phonology, morphology and semantics (as well as language development, cognitive linguistics)
- Cognitive modelling and generative capacity; search for linguistic universals
- First strict formal specifications (at first), but problems of overpremissiveness
Computational linguistics

- Focus in the 70’s is on cognitive simulation (with long term practical prospects.)
- The applied “branch” of CompLing is called Natural Language Processing
- Initially following Chomsky’s theory + developing efficient methods for parsing
- Early 80’s: unification based grammars (artificial intelligence, logic programming, constraint satisfaction, inheritance reasoning, object oriented programming...)

Unification-based grammars

- Based on research in artificial intelligence, logic programming, constraint satisfaction, inheritance reasoning, object oriented programming...
- The basic data structure is a feature-structure: attribute-value, recursive, co-indexing, typed; modelled by a graph
- The basic operation is unification: information preserving, declarative
- The formal framework for various linguistic theories: GPSG, HPSG, LFG,...
- Implementable!

An example HPSG feature structure
Problems

Disadvantage of rule-based (deep-knowledge) systems:
- Coverage (lexicon)
- Robustness (ill-formed input)
- Speed (polynomial complexity)
- Preferences (the problem of ambiguity: "Time flies like an arrow")
- Applicability? (more useful to know what is the name of a company than to know the deep parse of a sentence)
- EUROTRA and VERBMOBIL: success or disaster?

Back to data

- Late 1980's: applied methods based on data (the decade of "language resources")
- The increasing role of the lexicon
- (Re)emergence of corpora
- 90's: Human language technologies
- Data-driven shallow (knowledge-poor) methods
- Inductive approaches, esp. statistical ones (PoS tagging, collocation identification)
- Importance of evaluation (resources, methods)

The new millennium

The emergence of the Web:
- Simple to access, but hard to digest
- Large and getting larger
- Multilinguality

The promise of mobile, ‘invisible’ interfaces;
HLT in the role of middle-ware
<table>
<thead>
<tr>
<th>Processes, methods, and resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>RuslanMitkov (ed.)</td>
</tr>
<tr>
<td>- Text-to-Speech Synthesis</td>
</tr>
<tr>
<td>- Speech Recognition</td>
</tr>
<tr>
<td>- Text Segmentation</td>
</tr>
<tr>
<td>- Part-of-Speech Tagging and Lemmatisation</td>
</tr>
<tr>
<td>- Parsing</td>
</tr>
<tr>
<td>- Word-Sense Disambiguation</td>
</tr>
<tr>
<td>- Anaphora Resolution</td>
</tr>
<tr>
<td>- Natural Language Generation</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Finite-State Technology</td>
</tr>
<tr>
<td>Statistical Methods</td>
</tr>
<tr>
<td>Machine Learning</td>
</tr>
<tr>
<td>Lexical Knowledge Acquisition</td>
</tr>
<tr>
<td>Evaluation</td>
</tr>
<tr>
<td>Sublanguages and Controlled Languages</td>
</tr>
<tr>
<td>Corpora</td>
</tr>
<tr>
<td>Ontologies</td>
</tr>
</tbody>
</table>