Introduction to Human Language Technologies

Tomaž Erjavec

Karl-Franzens-Universität Graz

Lecture 1: Overview
9.11.2007

Overview

1. a few words about me
2. a few words about you
3. introduction to HLT
4. lab work: first steps with Python

Lecturer

- Tomaž Erjavec
  Department of Knowledge Technologies
  Jožef Stefan Institute
  Ljubljana
- http://nl.ijs.si/et/
- tomaz.erjavec@ijs.si
- Work: corpora and other language resources, standards, annotation, text-critical editions
- Web page for this course: http://nl.ijs.si/teach/graz07/hlt/
- assessment
Students

- background: field of study
- exposure to
  - linguistics?
  - corpus linguistics?
  - programming?
- emails

Overview of the course

1. Introduction
2. Basic processing of text
3. Working with corpora
4. Multilingual applications
5. Lexical semantics
6. ...

Lectures + work with NLTK
Computer processing of natural language

- Computational Linguistics:
  - a branch of computer science, that attempts to model the cognitive faculty of humans that enables us to produce/understand language
- Natural Language Processing:
  - a subfield of CL, dealing with specific methods to process language
- Human Language Technologies:
  - (the development of) useful programs to process language

Languages and computers

How do computers “understand” language?
- (written) language is, for a computer, merely a sequence of characters (strings)
  - words are separated by spaces
  - words are separated by spaces or punctuation
  - words are separated by spaces or punctuation and space
  - (2, 3H)dexamethasone, $4,000.00, pre-and post-natal, etc.

Problems

Languages have properties that humans find easy to process, but are very problematic for computers
- Ambiguity: many words, syntactic constructions, etc. have more than one interpretation
- Vagueness: many linguistic features are left implicit in the text
- Paraphrases: many concepts can be expressed in different ways

Humans use context and background knowledge; both are difficult for computers
• Time flies like an arrow.
• I saw the spy with the binoculars. He left the bank at 3 p.m.

The dimensions of the problem

Many applications require only a shallow level of analysis.

Structuralist and empiricist views on language

• The structuralist approach:
  – Language is a limited and orderly system based on rules.
  – Automatic processing of language is possible with rules.
  – Rules are written in accordance with language intuition.

• The empirical approach:
  – Language is the sum total of all its manifestations (written and spoken).
  – Generalisations are possible only the basis of large collections of language data, which serve as a sample of the language (corpora).
Other names for the two approaches

- rationalism vs. empiricism
- competence vs. performance
- deductive vs. inductive
- Deductive method: from the general to specific; rules are derived from axioms and principles; verification of rules by observations
- Inductive method: from the specific to the general; rules are derived from specific observations; falsification of rules by observations

Empirical approach

-Describing naturally occurring language data
-Objective (reproducible) statements about language
-Quantitative analysis: common patterns in language use
-Creation of robust tools by applying statistical and machine learning approaches to large amounts of language data
-Basis for empirical approach: corpora
-Empirical turn supported by rise in processing speed of computers and their amount of storage, and the revolution in the availability of machine-readable texts (the word-wide web)

The history of Computational Linguistics

- MT, empiricism (1950-70)
- Structuralism: 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

\[
\begin{align*}
S & \rightarrow NP \ V \ NP \\
NP & \rightarrow N \\
NP & \rightarrow \text{Det} \ N \\
NP & \rightarrow \text{NP that } S
\end{align*}
\]

\[
\begin{align*}
\text{S} & \rightarrow \text{NP } \ V \ \text{NP} \\
\text{NP} & \rightarrow \text{N} \\
\text{NP} & \rightarrow \text{Det} \ N \\
\text{NP} & \rightarrow \text{NP that } S
\end{align*}
\]

\[
\text{S} \rightarrow \text{NP V NP} \\
\text{NP} \rightarrow \text{N} \\
\text{NP} \rightarrow \text{Det N} \\
\text{NP} \rightarrow \text{NP that S}
\]
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,..)

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 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, Candide)
• 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

HLT applications

• Speech technologies
• Machine translation
• Question answering
• Information retrieval and extraction
• Text summarisation
• Text mining
• Dialogue systems
• Multimodal and multimedia systems
• Computer assisted: authoring; language learning; translating; lexicology; language research
HLT applications II.

• Corpus tools
  • concordance software
  • tools for statistical analysis of corpora
  • tools for compiling corpora
  • tools for aligning corpora
  • tools for annotating corpora

• Translation tools
  • programs for terminology databases
  • translation memory programs
  • machine translation

HLT research fields

• Phonetics and phonology: speech synthesis and recognition
• Morphology: morphological analysis, part-of-speech tagging, lemmatisation, recognition of unknown words
• Syntax: determining the constituent parts of a sentence (NP, VP) and their syntactic function (Subject, Predicate, Object)
• Semantics: word-sense disambiguation, automatic induction of semantic resources (thesauri, ontologies)
• Multilingual technologies: extracting translation equivalents from corpora, machine translation
• Internet: information extraction, text mining, advanced search engines

Processes, methods, and resources

The Oxford Handbook of Computational Linguistics, Ruslan Mitkov (ed.)

• Text-to-Speech Synthesis
• Speech Recognition
• Text Segmentation
• Part-of-Speech Tagging and lemmatisation
• Parsing
• Word-Sense Disambiguation
• Anaphora Resolution
• Natural Language Generation

• Finite-State Technology
• Statistical Methods
• Machine Learning
• Lexical Knowledge Acquisition
• Evaluation
• Sublanguages and Controlled Languages
• Corpora
• Ontologies
Further reading

• Language Technology World
  http://www.lt-world.org/

• The Association for Computational Linguistics
  http://www.aclweb.org/ (c.f. Resources)

• Interactive Online CL Demos
  http://www.ifi.unizh.ch/CL/InteractiveTools.html

• Natural Language Processing – course materials
  http://www.cs.cornell.edu/Courses/cs674/2003sp/