Introduction to Computational Linguistics

Sandra Kübler

LSA Summer Institute 2017
Computational linguistics is the scientific study of language from a computational perspective.

Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be ”knowledge-based” (”hand-crafted”) or ”data-driven” (”statistical” or ”empirical”).
Work in computational linguistics is in some cases motivated from a scientific perspective in that one is trying to provide a computational explanation for a particular linguistic or psycholinguistic phenomenon; and in other cases the motivation may be more purely technological in that one wants to provide a working component of a speech or natural language system.

http://www.aclweb.org/archive/what.html
Computational linguistics is the application of linguistic theories and computational techniques to problems of natural language processing.

http://www.ba.umist.ac.uk/public/departments/registrar/academicoffice/uga/lang.htm
Computational linguistics is the science of language with particular attention given to the processing complexity constraints dictated by the human cognitive architecture. Like most sciences, computational linguistics also has engineering applications.

http://www.cs.tcd.ie/courses/csll/CSLLcourse.html
Computational linguistics is the study of computer systems for understanding and generating natural language.

Short History

- 1950s: Machine Translation
- 1960s: Chomsky
- 1964: ALPAC Report
- 1966: ELIZA
- 1973: SHRDLU
- 1980s: knowledge-based CL
- 1990s: statistical / machine learning approaches in CL
Early 1950s

- Machine Translation (MT): one of the earliest applications of computers
- Major players: US and USSR
- Russian to English and reverse
- Georgetown University, Washington system:
  - translated sample texts in 1954
  - euphoria - a lot of funding, many groups in US, USSR
  - BUT: the system could not scale up
ALPAC Report

- assessed research results of groups working on MTs
- conclusions:
  - MT not possible in near future
  - funding should cease for MT!
  - basic research should be supported
- word by word translation does not work, linguistic knowledge is needed
ELIZA

- the first chatterbot – a computer program that mimics human conversation
- author: Joseph Weizenbaum – Massachusetts Institute of Technology
- simulation of a (Rogerian) therapist
- user types in some statement or set of statements in natural language
- ELIZA then analyzes the user’s statement and generates some response
- basic technology: pattern matching

USER: You don’t argue with me.
ELIZA: WHY DO YOU THINK I DON’T ARGUE WITH YOU
ELIZA

- the first chatterbot – a computer program that mimics human conversation
- author: Joseph Weizenbaum – Massachusetts Institute of Technology
- simulation of a (Rogerian) therapist
- user types in some statement or set of statements in natural language
- ELIZA then analyzes the user’s statement and generates some response
- basic technology: pattern matching

USER: You don’t argue with me.
ELIZA: WHY DO YOU THINK I DON’T ARGUE WITH YOU

Try it out: http://nlp-addiction.com/chatbot/eliza/
SHRDLU

- author: Terry Winograd – MIT
- interaction with a robot in a block world.
- the user can:
  - ask the robot to manipulate the blocks
  - ask it about the blocks configurations
  - ask it about its reasoning
  - update facts
- "understands" language in a limited domain by using syntactic parsing and semantic reasoning
- large scale grammar of English + parser
- procedural semantics for words and phrases
Knowledge-Based CL

- proof of concept
- manually written rules
- later: development of linguistic formalisms (Lexical Functional Grammar, Head-Driven Phrase Structure Grammar, Tree Adjoining Grammar etc.)
- linguistic/logic paradigm extensively pursued
- not robust enough
- few applications
- not scalable
Statistical / Machine Learning Approaches

- instead of writing rules, have computer learn rules / regularities
- approach massive ambiguity problem by probabilities
- need annotated data for training
- data sparseness problem
- unsupervised learning does not help: no linguistically relevant rules
- mildly supervised approaches
CL Applications

- spelling and grammar checking
- speech recognition
- Dialog Systems
- Machine Translation
- translation memories
- text summarisation
- information retrieval and extraction
- question answering
Machine Translation

2 approaches:

- deep linguistic processing
- statistical approach

systems:

- TAUM-Meteo
- Systran / Babelfish
- VerbMobil
- Google
Linguistic levels

- phonetics / phonology
- morphology
- POS annotation
- syntax
- lexical semantics
- discourse
CL Analysis

- finite-state morphology (analysis + generation)
- POS tagging
- parsing
- word sense disambiguation
- detect selectional restrictions (kill, murder, assassinate)
- shallow inference (X killed Y ⇒ Y is dead)
- anaphora / coreference resolution
Concepts Borrowed from Computer Science

- finite-state automata / transducers
- search: divide and conquer, beam search, nondeterminism, guides and oracles
- parsing (compilers)
- dynamic programming
- machine learning approaches: decision trees, $k$-nearest neighbors, clustering, support vector machines, ...