

Ling 201 – Computational Linguistics

Jirka Hana – May 22, 2006

Overview of topics

1. What is it?
2. Where is it used?
3. Theory driven vs. Machine learning

1 What is it?

Computational linguistics – study of how to process natural language with computers.

Purpose:

- practical: computers can help us with many tasks involving language
- theoretical: linguistics can test its theories

Methods used and developed in CL are used in other areas, too (DNA decoding, recognition of faces on photographs, etc.)

2 Where is it used?

2.1 Applications

Some of the following applications are already available (to a certain extent) some are still waiting to be made real:

- Machine translation from one language to another
 - MT can be fully or partly automatic
 - important especially for multilingual countries (Canada, India, Switzerland, etc.), international institutions (U.N., IMF, etc.), multinational companies, exporters
 - Currently, the European Union has more than 20 official languages. All federal laws and other documents have to be translated into all languages.
- Searching the internet or a database for relevant documents
 - Google, a library system, searching for law precedents
- Spell-checking & Grammar checking
 - e.g. **goed*, **sentance*, **speach*; **this books*, **a man are*; *to vs. too vs. two*

- Discovering plagiarism
 - analyzing text to find whether it is composed from blocks written by different people,
 - automatically check whether one text is not a rewritten version of another text
- Text summarization
 - e.g. creating a 3 page summary of a 1000 page book
- Dictation system, automatic automatic closed captioning
- Beeping out four letter-words
- Reading a written text aloud, e.g. for blind people
- Automatic customer service via phone or web in natural languages (instead of infinite menus)
 - e.g. You would say: *I need some help with the form I-765.*
- Adding diacritics to a text without it
 - e.g. *Munchen* → *München*, *Ceske Budejovice* → *České Budějovice*
-
-
-
-
-
-
-
- etc.

2.2 Parts

Things that can be used in the applications above

- Speech recognition/synthesis
- Morphological analysis
 - was* is a past tense of *be*, 1st or 3rd person sg.
- Tagging – determining the word classes of words
 - What are the word classes in *Can he can me for kicking a can?*
- Understanding dates:
 - 12th of April 2002; April 12, 2002; 04/12/2002; 04/12/02; 04/12; April 12; 2002-04-12

- Parsing – determining the syntactic structure(s) of a sentence
- Word sense disambiguation
What does *pen* mean in *Put some ink in the pen*, and what in *Put the pig in the pen*?
- What does a pronoun refer to.
What does *they* refer to in the following sentences:
The officials forbade the celebrations, because they were afraid of riots.
The officials forbade the celebrations, because they tend to be violent.
- Determining language of a text (after that you can run the appropriate spelling/grammar checker, translator, etc.)
- etc.

3 Theory driven vs. Machine learning

3.1 Theory driven approach

In this approach, (computational) linguists explicitly encode the knowledge about a particular language (its grammar) into a computer program.

- + Exploiting knowledge about language accumulated over the centuries
- Language grammars tend to be extremely complicated and fuzzy
- Grammar is not enough to understand sentences – world knowledge is also necessary
- Sufficient linguistic research has been done only for major languages

3.2 Machine learning approach

In this approach, a computer ‘learns’ about a particular language (or linguistic problem) from a corpus. The designer has to create the learning algorithm, but then the program learn merely by exposure to data.

Linguistic corpus – a large set of sentences from a particular language or languages. Sometimes it is accompanied with some linguistic information (e.g. each word comes with its word class assigned, morphology information; each sentence comes with its syntactical structure, etc.).

For example, the British National Corpus (BNC) contains 100M words in both written and spoken language (<http://www.hcu.ox.ac.uk/BNC/>)

- + It is very good in fuzzy areas of language which are hard to capture with rules.
- + The learning algorithm is usually language independent.
- It needs large corpora, often with a lot of additional information (e.g. syntactic trees)

- It does ‘stupid’ errors that the theory driven approach does not have problems with e.g. Could claim that *can* is a preposition.

(Probable) solution: Combination of both approaches – tell to the computer everything that linguistics knows, then let the computer learn the rest from corpora. However, it is extremely complicated.

4 See also

- Try speech recognition: 1-800-555-TELL (try for example news, weather)
- Foundations of Statistical Natural Language Processing: <http://www-nlp.stanford.edu/fsnlp/>
- Machine Translation: An Introductory Guide: <http://www.essex.ac.uk/linguistics/clmt/MTbook/>
- Microsoft NLP: <http://www.research.microsoft.com/research/nlp/>
- CL at MIT: <http://cognet.mit.edu/MITECS/Entry/joshi>