Ling/CSE 472: Introduction to Computational Linguistics

3/30/12

Introduction, overview, preview

Overview

- What is Computational Linguistics
- Syllabus
- Who’s here
- Showing the computer who’s boss
- Preview: Regular languages

What is Computational Linguistics?

- Getting computers to deal with human languages
    - ... for practical applications (examples?)
    - ... for linguistic research (examples?)

Linguistic research

- Searching large corpora for patterns of use and linguistic examples
- Creating structured databases of information for typological research (Autotyp, ODIN)
- Creating ontologies for interoperable markup of linguistic resources (GOLD)
- Modeling human linguistic competence and performance (computational psycholinguistics, grammar engineering)
- Software to facilitate language documentation (Elan, FIELD, SIL FieldWorks, Grammar Matrix)

Practical applications

- Speech recognition
- Speech synthesis
- Machine translation
- Information retrieval
- Natural language interfaces to computers
- Dialogue systems

Practical applications

- Computer-assisted language learning (CALL)
- Grammar checkers
- Spell checkers
- OCR (optical character recognition)
- Handwriting recognition
- Augmentative and assistive communication

Practical applications

- BioMedical NLP: Matching patients to clinical trials
- BioMedical NLP: Flagging electronic health records for urgent tests
- BioMedical NLP: Assistance in coding for insurance billing
- BioMedical NLP: Searching the biomedical literature for untested but promising things to study
- Legal domain: Electronic discovery

Practical applications

- B2B: Sentiment analysis for brand tracking
- Context-aware advertising
- Intelligence/national security: Monitoring social media, news, intercepted email/voice traffic
- ...

End-to-end applications are constructed from components that handle subtasks

- Each subtask has input and output
- Each subtask can be evaluated
    - precision, recall
    - intrinsic and extrinsic evaluation
- Output from one subtask is input to the next
- Many subtasks have “analysis” and “generation” variants
- Examples of subtasks?

Subtasks
(What's the input? What's the output?)

- Part of Speech tagging
- Named Entity Recognition
- Lemmatization
- Morphological analysis
- Parsing (constituent structure, dependency structure)
- Coreference resolution
- Word sense disambiguation
- Event detection
- Dialog act labeling
- Language modeling
- Alignment (of bitexts)
- ...

Statistical v. symbolic methods

- Still a hot topic
- Statistical methods involve *training a stochastic model* on a body of data so it can predict the most probable label/structure/etc for new data
    - Knowledge comes from implicit patterns in naturally occurring language (unsupervised learning) or from hand-labeled data (supervised learning)
- Symbolic methods involve *knowledge engineering*, or hand-coding of linguistic knowledge which is then applied to tasks
- Statistical methods provide *robustness*, symbolic methods *precision*
- Statistical and symbolic methods can be combined

Goals of this course

- Midway between “Language and Computers” and “X Methods for NLP”
- Familiarity with computational linguistic resources and how they are applied in research in computational linguistics and other subfields
- A rough sense of the state of the art (what can we do with language on computers anyway?)
- Ability to conceptualize problems from the perspective of computational linguistics

Syllabus

- Web page: http://courses.washington.edu/ling472
- Slides will be posted (often before lecture)
- Using Canvas (http://uw.instructure.com) and Tegrity (links will be included on Canvas page)
- Lab meetings (Fridays)

Course requirements

- Homework assignments (5 total, turned in via Canvas): 45%
- Midterm exam (4/25): 20%
- Final project: 30%
- Class participation: 5%
    - including Twitter assignment: 2%
- Get set up: see course web page for server cluster accounts, lab access, reading assignments, *link to first day WebQ*, etc.

@everyone: Tweet the class! #cl472

- http://courses.washington.edu/ling472/twitter.html
- What subcategory hashtags should we use?

Who’s here?

- A good class to work together --- everyone brings different skills
- I’m going to bring a lot to this class because...
- This is going to stretch me because...

Letting the computer know who's boss

- Computer 'literacy' is really a combination of experience and attitude

- Experience gives you the answers to many questions and a sense of what the possible space of answers is

- The important attitude boils down to confidence in one's ability to find the answer to a new question

- There are always new questions because:

    - The technology is always developing

    - There is too much for any one person to know it all

Letting the computer know who's boss

- Keep in mind:
  - It's always obvious once you know the answer
  - All pieces of software were designed by some person or people with some functionality in mind
- Places to look for answers:
  - On-line documentation (man, info, help)
  - Product websites (esp. discussion forums)
  - Google: websites, and especially newsgroups
  - Off-line documentation (i.e., books!)
  - Work together!
    - ... and post to the discussion boards in Canvas
  - 10 minute rule
    - It's ~~okay~~ critically important to ask questions!

Questions?

Overview

- What is Computational Linguistics
- Syllabus
- Who’s here
- Showing the computer who’s boss
- Preview: Regular languages