CMSC 325
Computational Linguistics

Fall 2022
Deepak Kumar

Administrivia

• CMSC325 Computational Linguistics (see course web page)
• Instructor: Deepak Kumar (dkumar@brynmawr.edu)
• Lectures: MW 10:10 to 11:30a
• Weekly Lab (optional): M 11:40a to 1:00p in Park 231
• Text: *Speech and Language Processing, 3rd Edition*
  Daniel Jurafsky & James Martin
  *Natural Language processing with Python –
Analyzing Text with the Natural Language
Toolkit (NLTK)*
  Steven Bird, Ewan Klein, and Edward Loper.
• Software: Python 3.0 + NLTK
Computational Linguistics

• Study what goes into getting computers to perform useful and interesting tasks involving human languages

• Also concerned with the insights that such computational work gives us into human processing of language

Why care?

• Enormous amount of knowledge is now available in machine readable form as natural language text.

• Conversational agents are becoming common: Siri, Google Voice, Alexa, etc.

• Much of human communication is now mediated by computers.
Some Common Applications

• Google Search

• Machine Translation
  • Google Translate
  • Phone apps – iTranslate (Demo)
  • Real-time language/voice translation (Demo)

• Q & A

• Web Analytics
  Data mining of blogs, discussion forums, message boards, user groups, social media, etc.
  for...
  • Product marketing information
  • Political opinion tracking
  • Social network analysis
  • Buzz analysis
  • Etc.

Google Translate: Buying Lentils in Italy!
Ricetta tipica

Ingredienti per 4 persone:
400 g di lenticchie, 1 litro d'acqua, 1 spicchio d'aglio, 1 gambo di sedano, sale e pepe. Versare le lenticchie su un tegame possibilmente di coccio, aggiungere l'acqua, l'aglio e il sedano: far cuocere per 20-30 minuti circa. A cottura quasi ultimata aggiungere sale e olio crudo. Servire con pane tostato e olio.

Importante la lenticchia non va tenuta a bagno.
Si consiglia la pulitura a "dito".

Produzione e Confezionato dall'Azienda Agricola
Salvatori Regina
Viale XX Settembre, 29
06046 Norcia (PG)
0743.816523

Da consumare preferibilmente entro 2 anni
Machine Translation: Progress

December 2009

Some Common Applications

• Google Search
• Machine Translation
  • Google Translate (Demo)
  • Phone apps – iTranslate (Demo – Deepak’s phone)
  • Real-time language/voice translation – Microsoft Research English to Chinese (Demo start at 5:25)
• Q & A (IBM Watson Jeopardy!, 2011) – Demo (https://www.youtube.com/watch?v=P18EdAKuC1U)
• Web Analytics
  Data mining of blogs, discussion forums, message boards, user groups, social media, etc. for...
  • Product marketing information
  • Political opinion tracking
  • Social network analysis
  • Buzz analysis
  • Etc.
More Applications

- Text/Document Classification
- Document Summarization
- Question/Answering
- Language Modeling
- Speech Recognition
- Caption Generation
- Text generation from a prompt
- Image Generation from a caption/description
Dall-e Mini

- https://www.craiyon.com/

Topics

- Words
- Syntax
- Meaning
- Discourse
Topics

• Words
• Syntax
• Meaning
• Discourse

Applications exploiting each

Applications – Language Processing versus Data Processing?

• An application that requires the use of knowledge about human languages

Example: Is Linux/Unix wc (word count) an example of a language processing application?
Applications – Language Processing versus Data Processing?

• An application that requires the use of knowledge about human languages

Example: Is Linux/Unix wc (word count) an example of a language processing application?

• When it counts words:
  • To count words you need to know what a word is. That is knowledge of language.

• When it counts lines and bytes: No
  • Lines and bytes are computer artifacts, not linguistic entities.
Some big applications requiring knowledge of language

• Question answering
• Conversation agents
• Summarization
• Machine Translation

These require a tremendous amount of knowledge of language.

Example

• Siri:
  What is the population of Bryn Mawr?
  What should I eat today?
  Tell me a joke.
What knowledge is needed?

• Speech recognition & synthesis
  Knowledge of English words (e.g. what they mean,...)

• How groups of words “clump”
  • What the clumps mean?

Course Content

• Linguistic topics
  • Phonology, morphology, syntax, discourse structure

• Formal Systems
  • Regular languages, context-free grammars, logic

• Applications
The Pipeline

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Ambiguity

- Computational Linguists are obsessed with ambiguity
- It is a fundamental problem of computational linguistics
- Resolving ambiguity is a crucial goal
Ambiguity

• Find at least five meanings of this sentence:

I made her duck.

- I cooked duck for her (to eat)
- I cooked the duck she owned
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into a duck
  ...

Ambiguity

• Find at least five meanings of this sentence:

I made her duck.
Ambiguity is Pervasive

I made her duck.

- I caused her to quickly lower her head or body
  - **Lexical category:** “duck” can be a N or V
- I cooked the duck she owned
  - **Lexical category:** “her” can be a possessive (“of her”) or a dative (“for her”)
- I created the (plaster?) duck she owns
  - **Lexical semantics:** “make” can mean “create” or “cook”

Ambiguity is Pervasive

- **Phonology**
  - I mate or duck
  - I'm eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I'm aid her duck
  - I mate her duck
  - I'm ate her duck
  - I'm ate or duck
  - I mate or duck
Dealing with ambiguity

- **Tightly coupled** interaction among processing levels; Knowledge from other levels can help resolve ambiguity.

- Ignore ambiguity as it occurs and hope that other levels can help resolve it – **Pipeline processing**

- Make the most likely choices – **probabilistic approaches**

- Don’t do anything, maybe it won’t matter

---

Models & Algorithms

- **Models** – formalisms that are used to capture the various kinds of linguistic knowledge that we need.
  
  State machines, Rule-based approaches, Logical formalisms, Probabilistic models, etc.

- **Algorithms** – used to manipulate the knowledge representations
  
  Transducers/filters, state-space search, dynamic programming, classifiers, etc.
The Pipeline

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Resources