

Computational Linguistics

Rules and Statistics

Kenneth Steimel



What is Computational Linguistics?

- Using principles from computer science to analyze and create language
 - Programming
 - Linear algebra
 - Graph Theory
 - Statistics

Okay, but what **is** Computational Linguistics?

- How many of you have used a virtual assistant? (Alexa, Google Assistant, Cortana, etc.)
- How many of you have used Google translate?
- How many of you have a spam filter on your email?

These are all applications of computational linguistics

Back to virtual assistants

- How many of you have issues with using a virtual assistant?
- How many of you never have any problems at all?
- Let's try some things...
 - » "Alexa, play the Mountain Goats"
 - » "Alexa, thank you"
 - » "Alexa, who is the president of Indiana University"
 - » "Alexa, where did he graduate from"
 - "Alexa, tell me about my schedule at 3"

Alexa, play the Mountain Goats

- What responses did you get?
- What things does a computer need to figure out to start playing this?

Alexa, play the Mountain Goats

- What responses did you get?
- What things does a computer need to figure out to start playing this?
- The Mountain Goats refers to a band not an animal
- The user intends for me to open some music app to play that band

Alexa, thank you

- What did you want to happen, given the last thing you said?
- What did your assistant do?

Alexa, thank you

- What did you want to happen, given the last thing you said?
 - I wanted Alexa to stop playing the Mountain Goats
- What did your assistant do?
 - “You’re welcome”

Alexa, who is the President of IU

- IU is an acronym that stands for Indiana University
- Did this work?

Alexa, where did he graduate from

- 'he' refers to the answer to the last question
- Did any of you have an assistant that failed at this?

Alexa, tell me about my schedule at 3

- What are the two structural ambiguities here?
- Which one did your assistant pick?

How does this work?

- Computational linguists primarily use two different types of approaches
 - Rule-based
 - Statistical

Rule-based Methods

- This is similar to what you've been doing in this class so far
- We write morphological rules, syntactic rules and phonological rules
- We use different algorithms to apply these rules
 - CYK and Earley parsers create trees out of syntactic rules
- But why does this matter?
 - Alexa, tell me my schedule at 3
 - » Like we learned earlier, the syntactic structure is related to the semantics of a sentence
 - » This allows tools like virtual assistants, search engines, etc. to determine the multiple meanings that a sentence might have and then determine the most likely interpretation from that.
 - “I think that Nicholas Cage stole the Declaration of Independence”
 - » “Nicholas Cage stole the Declaration of Independence” is not a fact here
 - » We need to know that the S inside the CP isn't a factual

Statistical Methods

- Statistical methods take language data (typically text) and do an analysis based on probabilities.
- Some types of analysis involve making models that learn from labels and then label completely new data (Supervised Learning)
- Other types of analysis involve grouping the data (Unsupervised learning)
- One application for this is called word sense disambiguation
 - This is a way of resolving lexical ambiguity using probabilities
 - “We can go to the movies”
 - “We can green beans every August”

Statistical Methods

- Statistical methods take language data (typically text) and do an analysis based on probabilities.
- Some types of analysis involve making models that learn from labels and then label completely new data (Supervised Learning)
- Other types of analysis involve grouping the data (Unsupervised learning)
- One application for this is called word sense disambiguation
 - This is a way of resolving lexical ambiguity using probabilities
 - “We can go to the movies”
 - “We can green beans every August”
 - Can is more likely to be an auxiliary when it occurs before a verb
 - We make a probabilistic model that takes this into account.

Computational Linguists develop language resources

- Lexicons: Word lists
- Knowledge graphs: Entities and relations between entities
 - <https://www.visual-thesaurus.com/wordnet.php?link=102084071>
- Treebanks: Parsed sentences
- Bi-texts: Pairs of sentences that mean the same thing in different languages

What classes on Computational Linguistics are available at IU?

- L245: Language and Computers
- L415: Corpus Linguistics
- L435: Foundational Skills in CL
- L445 Introduction to CL

Can I major in CL?

- Yes you can!
- We have a B.S. program for undergrads
- M.A. or M.S. in CL
- We also have a 5 year combined B.S. M.S. program