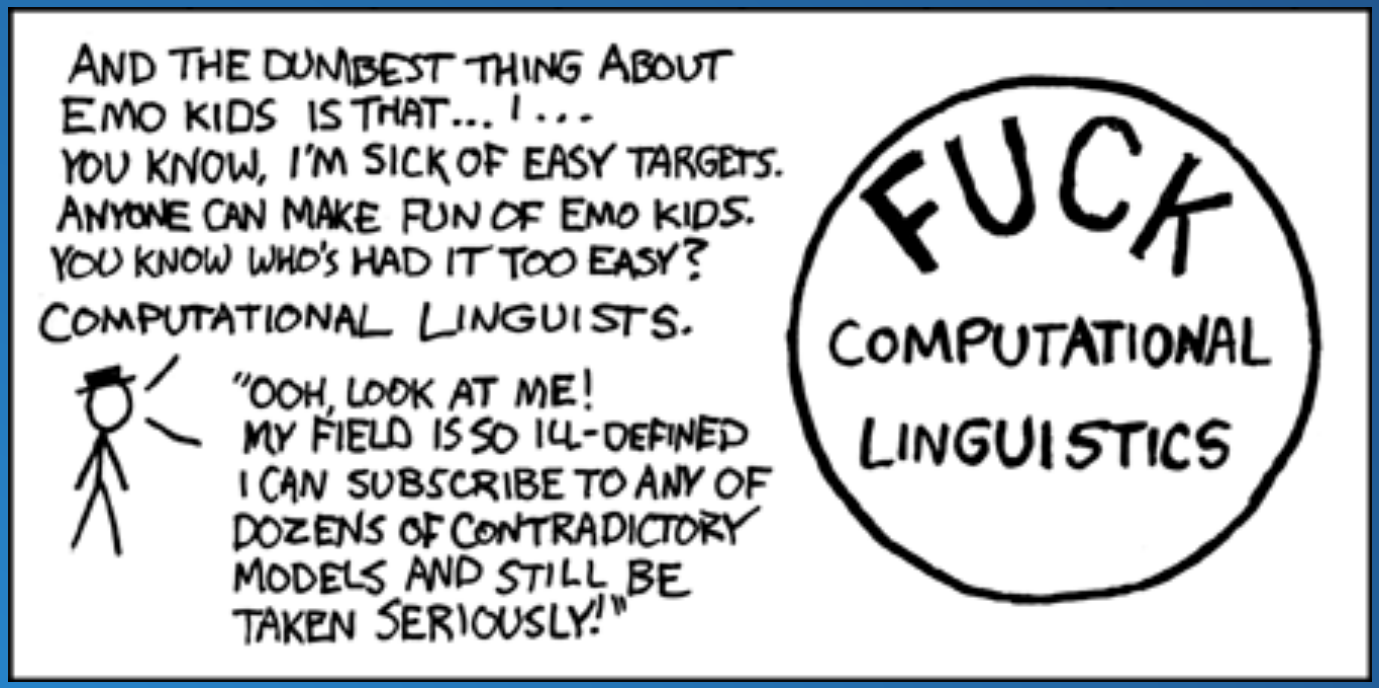
Computational Linguistics

by Caitlin Harding

with material from Jurafsky and Martin



credit: xkcd, http://xkcd.com/114

What is computational linguistics?

What is computational linguistics?

- Using computers/computational methods to model and/or process natural language

What is computational linguistics?

- Using computers/computational methods to model and/or process natural language
- Interdisciplinary
  - linguistics, computer science, artificial intelligence, mathematics, logic, philosophy, cognitive science, psychology, neurology, etc.

What does this look like in the real world?

What does this look like in the real world?

SIRI

What does this look like in the real world?

SIRI

chat bots
(ELIZA)

What does this look like in the real world?

SIRI

Watson

chat bots
(ELIZA)

What does this look like in the real world?

SIRI

Watson

chat bots
(ELIZA)

Google
translate

What does this look like in the real world?

SIRI

Watson

chat bots
(ELIZA)

Google
translate

Vocaloid

What does this look like in the real world?

SIRI

Watson

chat bots
(ELIZA)

language
analysis

Google
translate

Vocaloid

What does this look like in the real world?

SIRI

Watson

chat bots
(ELIZA)

language
analysis

Google
translate

Vocaloid

spam
detection

What does this look like in the real world?

SIRI

Watson

chat bots (ELIZA)

language analysis

Google translate

Vocaloid

voice automated services

spam detection

What does this look like in the real world?

SIRI

Watson

chat bots (ELIZA)

language analysis

Google translate

Vocaloid

voice automated services

spam detection

robots

Let's break this down!

Let's break this down!

- Words

Let's break this down!

- Words
  - regex, automata, n-grams, part-of-speech (POS) tagging

Let's break this down!

- Words
  - regex, automata, n-grams, part-of-speech (POS) tagging
- Speech

Let's break this down!

- Words
  - regex, automata, n-grams, part-of-speech (POS) tagging
- Speech
  - speech synthesis, voice recognition

Let's break this down!

- Words
  - regex, automata, n-grams, part-of-speech (POS) tagging
- Speech
  - speech synthesis, voice recognition
- Syntax

Let's break this down!

- Words
  - regex, automata, n-grams, part-of-speech (POS) tagging
- Speech
  - speech synthesis, voice recognition
- Syntax
  - grammars, syntactic parsing, statistical parsing

Let's break this down!

- Words
  - regex, automata, n-grams, part-of-speech (POS) tagging
- Speech
  - speech synthesis, voice recognition
- Syntax
  - grammars, syntactic parsing, statistical parsing
- Semantics

Let's break this down!

- Words
  - regex, automata, n-grams, part-of-speech (POS) tagging
- Speech
  - speech synthesis, voice recognition
- Syntax
  - grammars, syntactic parsing, statistical parsing
- Semantics
  - logic, word-sense disambiguation, word relations, semantic role labeling

Words

Words

- regex

Words

- regex
  - short for "regular expressions"

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines
- automata

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines
- automata
  - finite state automata, finite state machines

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines
- automata
  - finite state automata, finite state machines
  - often used for modelling morphological processes

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines
- automata
  - finite state automata, finite state machines
  - often used for modelling morphological processes
- n-grams

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines
- automata
  - finite state automata, finite state machines
  - often used for modelling morphological processes
- n-grams
  - probability of a given word/character given the word(s) /character(s) preceding it

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines
- automata
  - finite state automata, finite state machines
  - often used for modelling morphological processes
- n-grams
  - probability of a given word/character given the word(s) /character(s) preceding it
  - unigram, bigram, trigram, etc.

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines
- automata
  - finite state automata, finite state machines
  - often used for modelling morphological processes
- n-grams
  - probability of a given word/character given the word(s) /character(s) preceding it
  - unigram, bigram, trigram, etc.
  - spelling correction, language identification, etc.

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines
- automata
  - finite state automata, finite state machines
  - often used for modelling morphological processes
- n-grams
  - probability of a given word/character given the word(s) /character(s) preceding it
  - unigram, bigram, trigram, etc.
  - spelling correction, language identification, etc.
  - Claude Shannon

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines
- automata
  - finite state automata, finite state machines
  - often used for modelling morphological processes
- n-grams
  - probability of a given word/character given the word(s) /character(s) preceding it
  - unigram, bigram, trigram, etc.
  - spelling correction, language identification, etc.
  - Claude Shannon
- POS tagging

Words

- regex
  - short for "regular expressions"
  - used by a lot of search engines
- automata
  - finite state automata, finite state machines
  - often used for modelling morphological processes
- n-grams
  - probability of a given word/character given the word(s) /character(s) preceding it
  - unigram, bigram, trigram, etc.
  - spelling correction, language identification, etc.
  - Claude Shannon
- POS tagging
  - rule-based or stochastic (statistical)

Speech

Speech

- Closer to hard sciences/lab work

Speech

- Closer to hard sciences/lab work
- speech synthesis and text-to-speech

Speech

- Closer to hard sciences/lab work
- speech synthesis and text-to-speech
  - many stages

Speech

- Closer to hard sciences/lab work
- speech synthesis and text-to-speech
  - many stages
    - text -> segmented text -> phones (sounds) -> syllable boundaries -> intonation, pitch, and length -> F0 (fundamental frequency)

Speech

- Closer to hard sciences/lab work
- speech synthesis and text-to-speech
  - many stages
    - text -> segmented text -> phones (sounds) -> syllable boundaries -> intonation, pitch, and length -> F0 (fundamental frequency)
  - Vocaloid

Speech

- Closer to hard sciences/lab work
- speech synthesis and text-to-speech
  - many stages
    - text -> segmented text -> phones (sounds) -> syllable boundaries -> intonation, pitch, and length -> F0 (fundamental frequency)
  - Vocaloid
- speech recognition

Syntax

Syntax

- grammar

Syntax

- grammar
  - model for a given language that is to be used by a computer for language processing and/or generation

Syntax

- grammar
  - model for a given language that is to be used by a computer for language processing and/or generation
- syntactic and statistical parsing

Syntax

- grammar
  - model for a given language that is to be used by a computer for language processing and/or generation
- syntactic and statistical parsing
  - attempt to clarify structural ambiguities ("Mary saw the man with binoculars.")

Syntax

- grammar
  - model for a given language that is to be used by a computer for language processing and/or generation
- syntactic and statistical parsing
  - attempt to clarify structural ambiguities ("Mary saw the man with binoculars.")
  - label roles, POS

Semantics

Semantics

- "Colorless green ideas sleep furiously."
  - syntactically correct
  - semantically incorrect - how can we determine that computationally?

Semantics

- "Colorless green ideas sleep furiously."
  - syntactically correct
  - semantically incorrect - how can we determine that computationally?
- logic
  - language used to talk about language

Semantics

- "Colorless green ideas sleep furiously."
  - syntactically correct
  - semantically incorrect - how can we determine that computationally?
- logic
  - language used to talk about language
- word-sense disambiguation
  - "I went fishing for some bass." vs "The bass in this song is awesome."

Semantics

- "Colorless green ideas sleep furiously."
  - syntactically correct
  - semantically incorrect - how can we determine that computationally?
- logic
  - language used to talk about language
- word-sense disambiguation
  - "I went fishing for some bass." vs "The bass in this song is awesome."
- word relations
  - synonyms, antonyms, etc.

Semantics

- "Colorless green ideas sleep furiously."
  - syntactically correct
  - semantically incorrect - how can we determine that computationally?
- logic
  - language used to talk about language
- word-sense disambiguation
  - "I went fishing for some bass." vs "The bass in this song is awesome."
- word relations
  - synonyms, antonyms, etc.
- semantic role labeling
  - "Mary gave the book to John." (Mary = agent, book = theme, John = recipient)

Semantics

- "Colorless green ideas sleep furiously."
  - syntactically correct
  - semantically incorrect - how can we determine that computationally?
- logic
  - language used to talk about language
- word-sense disambiguation
  - "I went fishing for some bass." vs "The bass in this song is awesome."
- word relations
  - synonyms, antonyms, etc.
- semantic role labeling
  - "Mary gave the book to John." (Mary = agent, book = theme, John = recipient)
- General meaning
  - sometimes becomes a philosophical question - what really is a table?
  - emotion - how do you determine if something is sad, happy, etc.?

NLP

NLP

- natural language processing

NLP

- natural language processing
- concerned with the interactions between computers and natural human languages, specifically in extracting meaningful information from natural language or producing meaningful natural language

NLP

- natural language processing
- concerned with the interactions between computers and natural human languages, specifically in extracting meaningful information from natural language or producing meaningful natural language
- related to artificial intelligence, machine learning, machine translation, etc.

NLP

- natural language processing
- concerned with the interactions between computers and natural human languages, specifically in extracting meaningful information from natural language or producing meaningful natural language
- related to artificial intelligence, machine learning, machine translation, etc.
- deals with many open problems

Growing popularity

Growing popularity

- many open problems (problems that have not been completely solved or have room for improvements)

Growing popularity

- many open problems (problems that have not been completely solved or have room for improvements)
- a lot of interest in solving these problems, both academically and commercially - we live in the age of "big data"

Growing popularity

- many open problems (problems that have not been completely solved or have room for improvements)
- a lot of interest in solving these problems, both academically and commercially - we live in the age of "big data"
- UW CS department just hired four of the biggest names in machine learning and "big data"

Growing popularity

- many open problems (problems that have not been completely solved or have room for improvements)
- a lot of interest in solving these problems, both academically and commercially - we live in the age of "big data"
- UW CS department just hired four of the biggest names in machine learning and "big data"
- need hand annotated data

I'm sold! Now what?

- CSE472/LING472 - Intro to Computational Linguistics
- UW Professional Master's in Computational Linguistics
- Coursera - Stanford's NLP Course
- CS Department Website (research, talks, etc.)
- Advisors