What is Computational Linguistics?

Thomas Graf

Stony Brook University
mail@thomasgraf.net

Two Kinds of Computational Linguistics

Two Kinds of Computational Linguistics

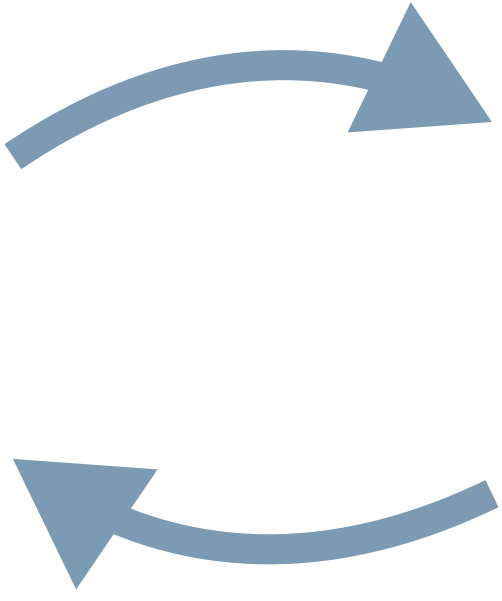
doing language with computers

Two Kinds of Computational Linguistics

doing language with computers

how humans compute language

Two Kinds of Computational Linguistics



doing language with computers

how humans compute language

Language Technology: The State of the Art

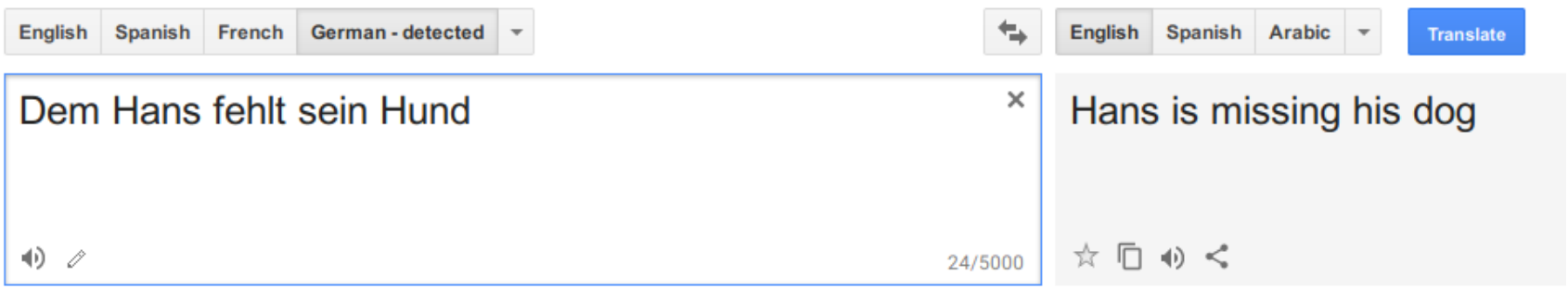
- It seems that modern technology has mastered language:
  - machine translation
  - text generation
  - chat bots
  - word prediction
- But is that true? Let's do some **experiments**!

Google Mis-Translate

1. Go to translate.google.com.
2. Enter *Dem Hans fehlt sein Hund*,
   'Hans is missing his dog'.
3. Now enter *Der Hans fehlt seinem Hund*,
   'Hans is missed by his dog'.
4. Does Google get the difference?

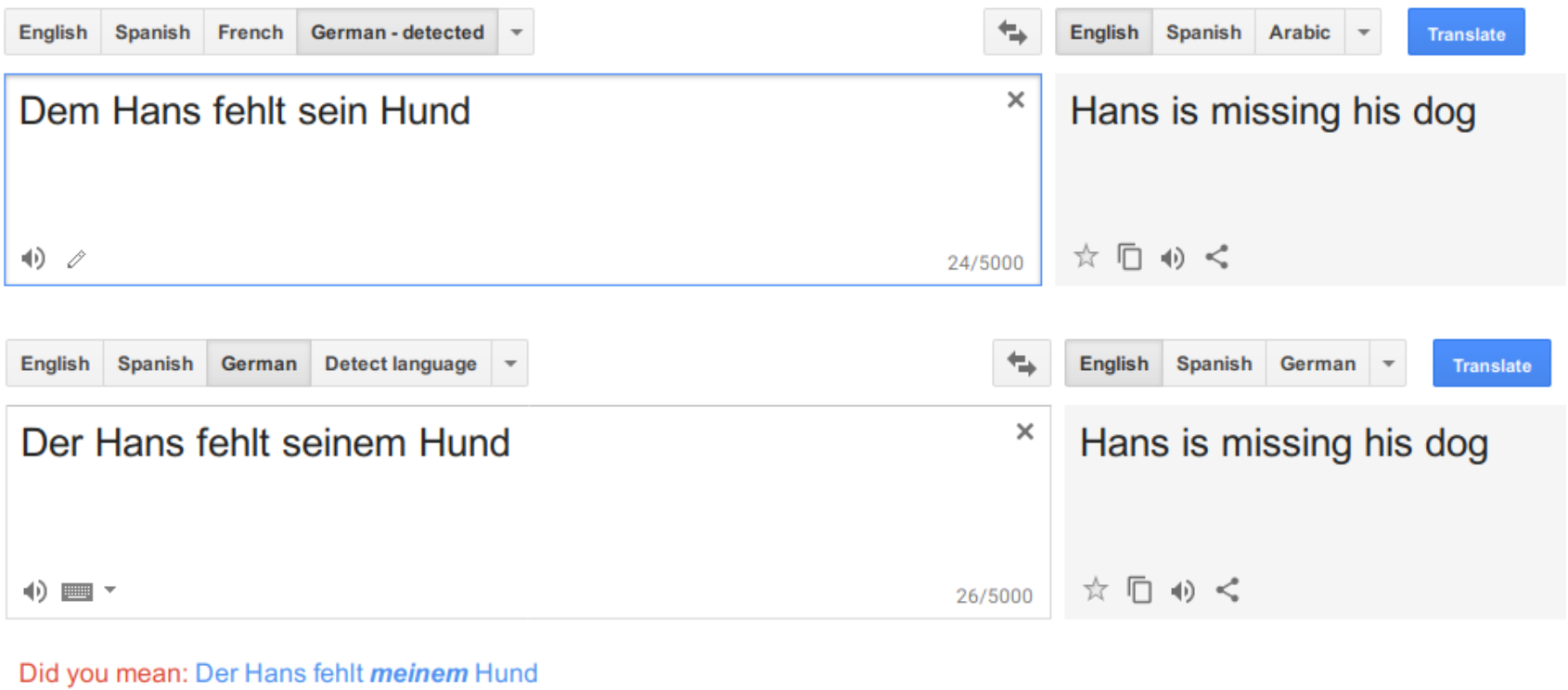
Google Mis-Translate

1. Go to translate.google.com.
2. Enter *Dem Hans fehlt sein Hund*, ‘Hans is missing his dog’.
3. Now enter *Der Hans fehlt seinem Hund*, ‘Hans is missed by his dog’.
4. Does Google get the difference?



Google Mis-Translate

1. Go to translate.google.com.
2. Enter *Dem Hans fehlt sein Hund*, 'Hans is missing his dog'.
3. Now enter *Der Hans fehlt seinem Hund*, 'Hans is missed by his dog'.
4. Does Google get the difference?

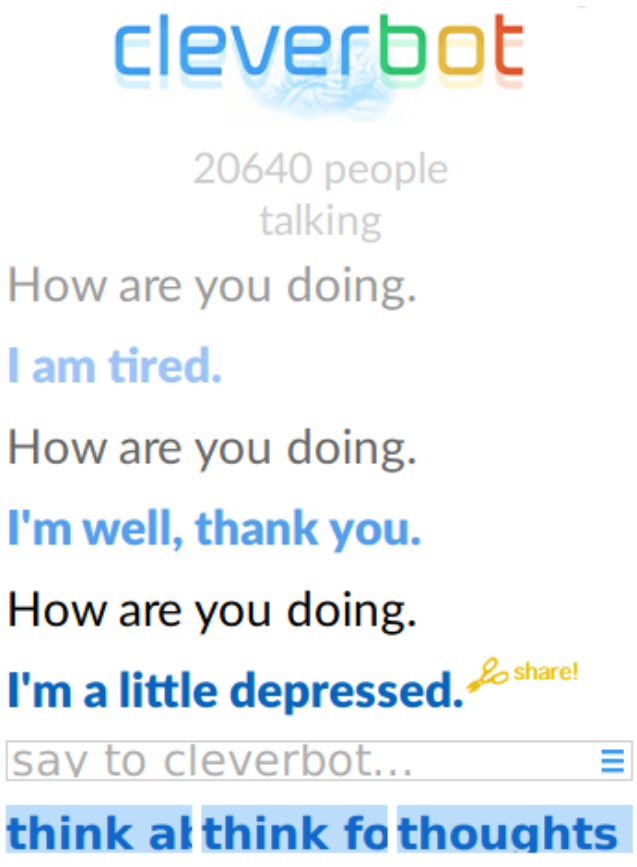


Fooling the Chatbot

1. Go to cleverbot.com.
2. Immediately enter *How are you doing*.
3. Keep entering it and look at the replies you get.
4. Do you still think you're talking to a human?

Fooling the Chatbot

1. Go to cleverbot.com.
2. Immediately enter *How are you doing*.
3. Keep entering it and look at the replies you get.
4. Do you still think you're talking to a human?



The Word Prediction Loop

1. Open some chat or messaging app on your phone.
2. Don't type anything.
3. Instead, click the second word suggestion (the one in the middle).
4. Keep doing this.
5. Did you get a reasonable sentence of English?

The Word Prediction Loop

1. Open some chat or messaging app on your phone.
2. Don't type anything.
3. Instead, click the second word suggestion (the one in the middle).
4. Keep doing this.
5. Did you get a reasonable sentence of English?

I am a beautiful person who is the best of luck to you by the way to get the best of luck to you by the way to get the best of luck to you by the way to get the . . .

It's All Pattern Matching

- Computers have no understanding of language,
  they just do **pattern matching**.
- They're like a student who memorized the math solutions
  without understanding them.
- So how does the pattern matching work?

n-Grams for Word Prediction

- The word prediction looped because it does not look at the whole sentence.
- Instead, only consecutive chunks of words are considered
  $\Rightarrow$ **n-grams**

Example

- **Sentence:** John really likes Mary.
- **Bigrams (n=2):**
  John really really likes likes Mary Mary .
- **Trigrams (n=3):**
  John really likes really likes Mary likes Mary .

You Do it!

- **Sentence:**
  When buffalo buffalo buffalo buffalo buffalo buffalo.
- **Bigrams:**

- **Trigrams:**

You Do it!

- **Sentence:**
  When buffalo buffalo buffalo buffalo buffalo buffalo.
- **Bigrams:**
  When buffalo
  buffalo buffalo
  buffalo .
- **Trigrams:**

You Do it!

- **Sentence:**
  When buffalo buffalo buffalo buffalo buffalo buffalo.
- **Bigrams:**
  When buffalo
  buffalo buffalo
  buffalo .
- **Trigrams:**
  When buffalo buffalo
  buffalo buffalo buffalo
  buffalo buffalo .

Frequencies for Prediction

- Suppose you have a large **corpus**.
  e.g. the collected writings of the Wall Street Journal
- Then you can extract **transition probabilities**:
  1. Extract all bigrams/trigrams.
  2. Count how often each one occurs.
  3. Convert that to a percentage.
- Use those percentages to predict the next likely word.

When buffalo buffalo buffalo buffalo buffalo buffalo

Bigram	Count	Percentage
When buffalo	1	
buffalo buffalo	5	
buffalo .	1	

Input: When buffalo

Possible completions ranked by likelihood:

1. buffalo
2. .

Frequencies for Prediction

- Suppose you have a large **corpus**.
  e.g. the collected writings of the Wall Street Journal
- Then you can extract **transition probabilities**:
  1. Extract all bigrams/trigrams.
  2. Count how often each one occurs.
  3. Convert that to a percentage.
- Use those percentages to predict the next likely word.

When buffalo buffalo buffalo buffalo buffalo buffalo

Bigram	Count	Percentage
When buffalo	1	$\frac{1}{7} = 14\%$
buffalo buffalo	5	
buffalo .	1	

Input: When buffalo

Possible completions ranked by likelihood:

1. buffalo
2. .

Frequencies for Prediction

- Suppose you have a large **corpus**.
  e.g. the collected writings of the Wall Street Journal
- Then you can extract **transition probabilities**:
  1. Extract all bigrams/trigrams.
  2. Count how often each one occurs.
  3. Convert that to a percentage.
- Use those percentages to predict the next likely word.

When buffalo buffalo buffalo buffalo buffalo buffalo

Bigram	Count	Percentage
When buffalo	1	$\frac{1}{7} = 14\%$
buffalo buffalo	5	$\frac{5}{7} = 72\%$
buffalo .	1	

Input: When buffalo

Possible completions ranked by likelihood:

1. buffalo
2. .

Frequencies for Prediction

- Suppose you have a large **corpus**.
  e.g. the collected writings of the Wall Street Journal
- Then you can extract **transition probabilities**:
  1. Extract all bigrams/trigrams.
  2. Count how often each one occurs.
  3. Convert that to a percentage.
- Use those percentages to predict the next likely word.

When buffalo buffalo buffalo buffalo buffalo buffalo

Bigram	Count	Percentage
When buffalo	1	$\frac{1}{7} = 14\%$
buffalo buffalo	5	$\frac{5}{7} = 72\%$
buffalo .	1	$\frac{1}{7} = 14\%$

Input: When buffalo

Possible completions ranked by likelihood:

1. buffalo
2. .

This is the Core of Current Language Technology

- Gather lots of data.
- Extract chunks of structure (e.g. n-grams).
- Determine their frequency, and use that for probabilistic reasoning.

Applications

- word predictions
- search engines
- machine translation
- optical character recognition
- stylistic analysis
- text generation

This is no match for how humans use language!

This is the Core of Current Language Technology

- Gather lots of data.
- Extract chunks of structure (e.g. n-grams).
- Determine their frequency, and use that for probabilistic reasoning.

Applications

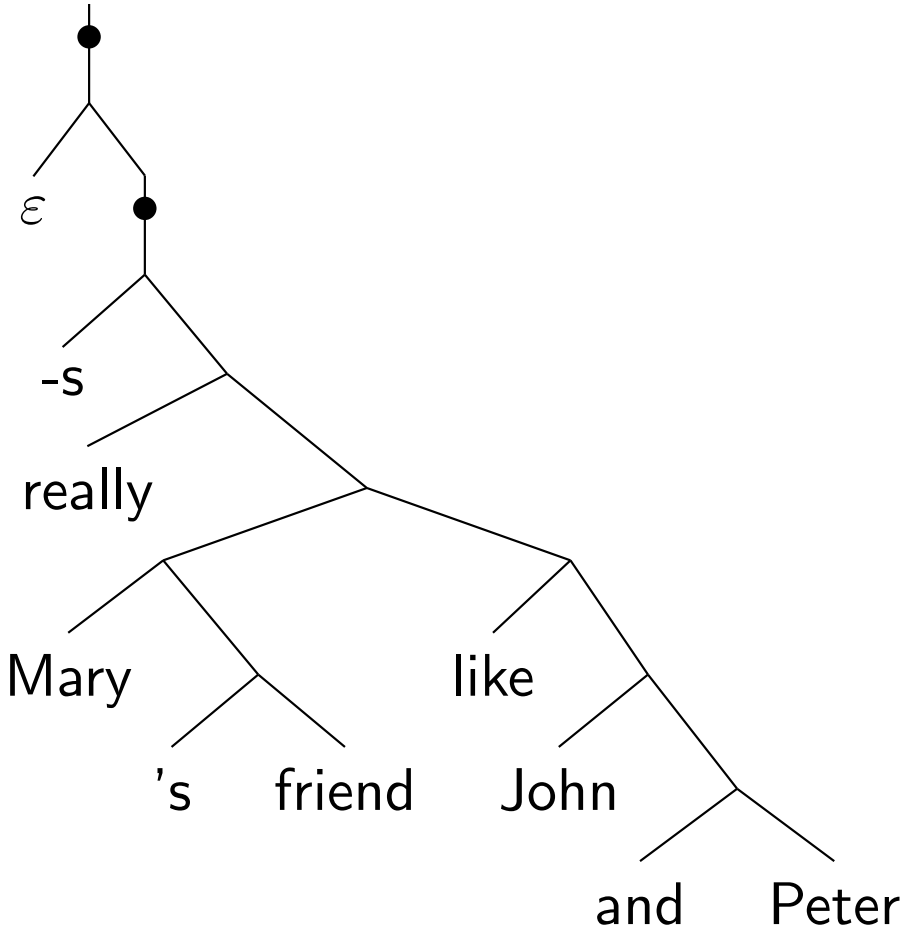
- word predictions
- search engines
- machine translation
- optical character recognition
- stylistic analysis
- text generation

This is no match for how humans use language!

A Brief Glimpse At Linguistics

- **Sentences** are structurally complex, like **molecules**.

John and Peter, Mary's friend really likes



like(x,John) & like(x,Peter) & friend(Mary,x)

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself/himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself/himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself/himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself/himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him/himself.

Pitfalls of Language I

- The interpretation of pronouns is very tricky.
- Yet you all have perfect command of this complex system.

(1) a. Hugo likes himself.
b. Hugo likes him.
c. Hugo knows that Bill likes him/himself.

(2) a. Hugo introduced himself to Bill.
b. Hugo introduced Bill to himself/himself.

(3) Hugo looked at his contemporaries — less clever than him/himself — and saw them outstrip him.

Pitfalls of Language II

- Even contractions can have an effect on meaning.
- Every native speaker has somehow mastered these rules.

(4) Who do you want to leave?

(5) Who do you wanna leave?

Pitfalls of Language II

- Even contractions can have an effect on meaning.
- Every native speaker has somehow mastered these rules.

(4) Who do you want to leave?

Answer 1: I want to leave Bill.
Answer 2: I want Bill to leave.

(5) Who do you wanna leave?

Pitfalls of Language II

- Even contractions can have an effect on meaning.
- Every native speaker has somehow mastered these rules.

(4) Who do you want to leave?

Answer 1: I want to leave Bill.
Answer 2: I want Bill to leave.

(5) Who do you wanna leave?

Answer 1: I want to leave Bill.
Answer 2: impossible

Pitfalls of Language III

Adjectives usually modify nouns.

(6) retired physicist = physicist who is retired

But sometimes they can only modify a subpart.

(7) a. nuclear physicist ≠ physicist who is nuclear
    b. nuclear physicist = a scientist working in nuclear physics

And sometimes they can do both.

(8) a. radical feminist = a feminist who is radical
    b. radical feminist = an advocate of radical feminism

Pitfalls of Language III

Adjectives usually modify nouns.

(6) retired physicist = physicist who is retired

But sometimes they can only modify a subpart.

(7) a. nuclear physicist ≠ physicist who is nuclear

b. nuclear physicist = a scientist working in nuclear physics

And sometimes they can do both.

(8) a. radical feminist = a feminist who is radical

b. radical feminist = an advocate of radical feminism

Pitfalls of Language III

Adjectives usually modify nouns.

(6) retired physicist = physicist who is retired

But sometimes they can only modify a subpart.

(7) a. nuclear physicist $\neq$ physicist who is nuclear

b. nuclear physicist = a scientist working in nuclear physics

And sometimes they can do both.

(8) a. radical feminist = a feminist who is radical

b. radical feminist = an advocate of radical feminism

Pitfalls of Language III

Adjectives usually modify nouns.

(6) retired physicist = physicist who is retired

But sometimes they can only modify a subpart.

(7) a. nuclear physicist $\neq$ physicist who is nuclear

b. nuclear physicist = a scientist working in nuclear physics

And sometimes they can do both.

(8) a. radical feminist = a feminist who is radical

b. radical feminist = an advocate of radical feminism

Pitfalls of Language III

Adjectives usually modify nouns.

(6) retired physicist = physicist who is retired

But sometimes they can only modify a subpart.

(7) a. nuclear physicist ≠ physicist who is nuclear
b. nuclear physicist = a scientist working in nuclear physics

And sometimes they can do both.

(8) a. radical feminist = a feminist who is radical
b. radical feminist = an advocate of radical feminism

Firestorm, nuclear physicist

Pitfalls of Language IV

Even when the meaning of a phrase is clear,
it can still be ungrammatical.

(9) a. a rusty car

b. a car that is rusty

(10) a. a former president

b. ?? a president that is former

And the presence of a single word can make a sentence ungrammatical, even if it is fine in very similar sentences.

(11) a. Who do you think Bill likes?

b. Who do you think likes Bill?

(12) a. Who do you think that Bill likes?

b. ?? Who do you think that likes Bill?

Pitfalls of Language IV

Even when the meaning of a phrase is clear,
it can still be ungrammatical.

(9) a. a rusty car
b. a car that is rusty

(10) a. a former president
b. ?? a president that is former

And the presence of a single word can make a sentence
ungrammatical, even if it is fine in very similar sentences.

(11) a. Who do you think Bill likes?
b. Who do you think likes Bill?

(12) a. Who do you think that Bill likes?
b. ?? Who do you think that likes Bill?

The Human Mystery

- We all have complete mastery of these rules even though
  - we were never told about them (not even in grammar courses)
  - we have no conscious knowledge of them.
- Humans don't learn language, we acquire it naturally, without explicit instruction.
- Even a five year old is better at language than computers.

The Big Questions of (Computational) Linguistics

- How are humans so good at language?
- What does their knowledge look like, and how is it acquired?
- How can we get computers to work this way?

The Human Mystery

- We all have complete mastery of these rules even though
  - we were never told about them (not even in grammar courses)
  - we have no conscious knowledge of them.
- Humans don't learn language, we acquire it naturally, without explicit instruction.
- Even a five year old is better at language than computers.

The Big Questions of (Computational) Linguistics

- How are humans so good at language?
- What does their knowledge look like, and how is it acquired?
- How can we get computers to work this way?

We Need More Complex Models

- Language is all about hidden structure.
- Humans are **incredibly good** at inferring these structures and reasoning with them.
- Without this knowledge, computers will never have human-like performance.
- But we **can't get it to work** right now.

Two Problems

- Complex models are harder to work with.
  You already know enough Python for working with n-grams!
- Complex models are much harder to compute with.
  They are still too demanding for most applications.

Take-Home Messages

1. Don't be too impressed by current language technology.
   It's all very simplistic and no match for even a 4-year old.
2. For real progress, we need linguistically informed models.
3. But there's still many practical hurdles to overcome.
   Don't expect talking robots anytime soon.