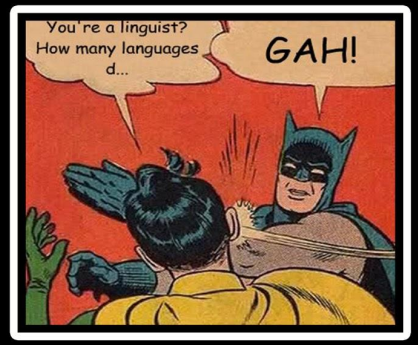
Computational linguistics as a profession

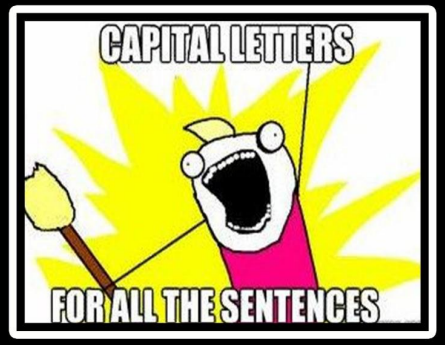
Mariana Romanyshyn
Computational Linguist at Grammarly, Inc.

COMPUTATIONAL LINGUIST

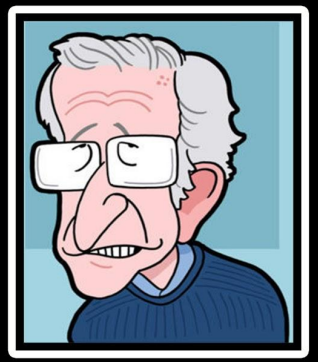


WHAT MY FRIENDS THINK I DO

WHAT MY MOTHER THINKS I DO



WHAT SOCIETY THINKS I DO



WHAT I THINK I DO

```
def generate_sentence(first_word):
        """Generate a sentence using the first word."""
        ngram_list = trigrams(brown.words(categories = "adventure"))
        fd = FreqDist(ngram_list)
        sentence = [first_word]
        while len(sentence) < 40:
                list_of_nexts = []
                for (i, j, k) in fd.keys():
                        if i == first_word:
                                list_of_nexts.append(j)
                        if len(list_of_nexts) == 10:
                                break
                if len(list_of_nexts) == 0:
                        return "The sentence cannot be generated."
                first_word = random.choice(list_of_nexts)
                sentence.append(first_word)
                if first_word in [".", "!", "?", "..."]:
                        break
        return ' '.join(sentence)
```

WHAT I REALLY DO

Contents

1. NLP applications in our world
2. Competencies of a computational linguist
3. Grammarly and CompLing Summer School

1. NLP applications in our world

The Goal of NLP

Goal:

have computers ***understand*** natural language in order to perform ***useful*** tasks

How:

transform free-form text into structured data and back

What NLP applications do you know?

Types of NLP Applications

- Analysis
- Transformation
- Generation

Types of NLP Applications

ANALYSIS

Spam Filtering

…

Types of NLP Applications

ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

- Quora: Insincere Questions (2019)
- Jigsaw: Toxic Comments (2018)
- Workshop on Abusive Language Online (2017-2019)

…

Types of NLP Applications

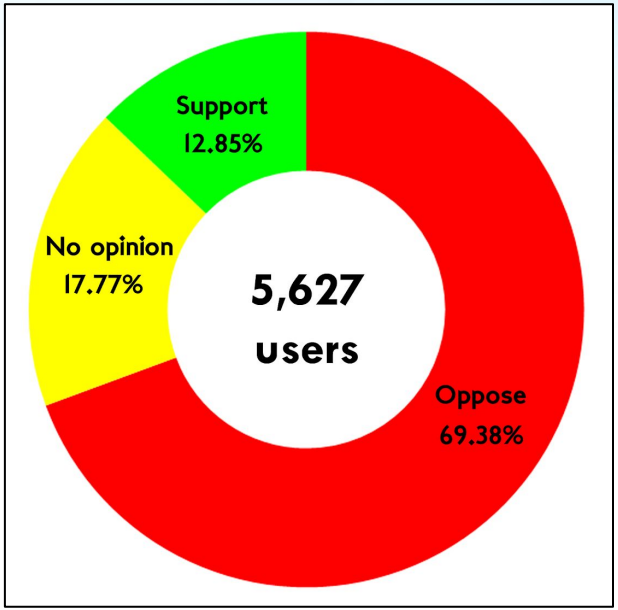
ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

…



Sentiment maps

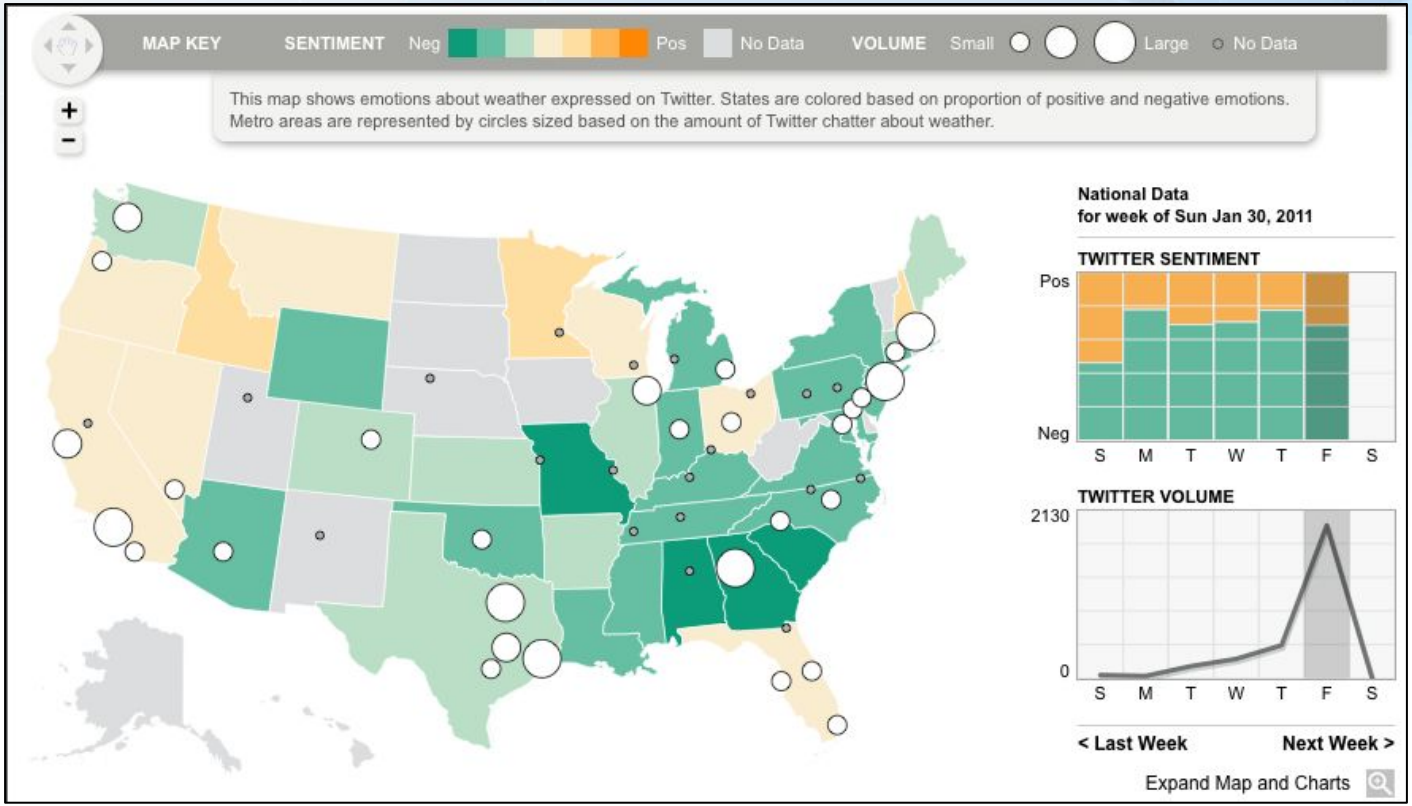


Image from http://www.dialogueearth.org/

Sentiment Analysis

It tastes amazing!
It tastes horrible!
Nothing special.
Cola tastes much better than Pepsi.

Sentiment Analysis

It tastes amazing!
It tastes horrible!
Nothing special.
Cola tastes much better than Pepsi.

Sentiment Analysis

It tastes like beer!
It tastes interesting!
It tastes like my mom said it would!
If it was served with milk, it would taste great!

Sentiment Analysis

It tastes like beer!
It tastes interesting!
It tastes like my mom said it would!
If it was served with milk, it would taste great!

The good taste was **no** surprise.
If only it tasted good!
It was **not only** good but also cheap!

Sentiment Analysis

Types of NLP Applications

ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

- sentiment classes or sentiment scale
- objects of the sentiment
- type of emotion
- subjectivity

Types of NLP Applications

ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

Sarcasm/Irony Detection

Humor Detection

…

Cognitive features

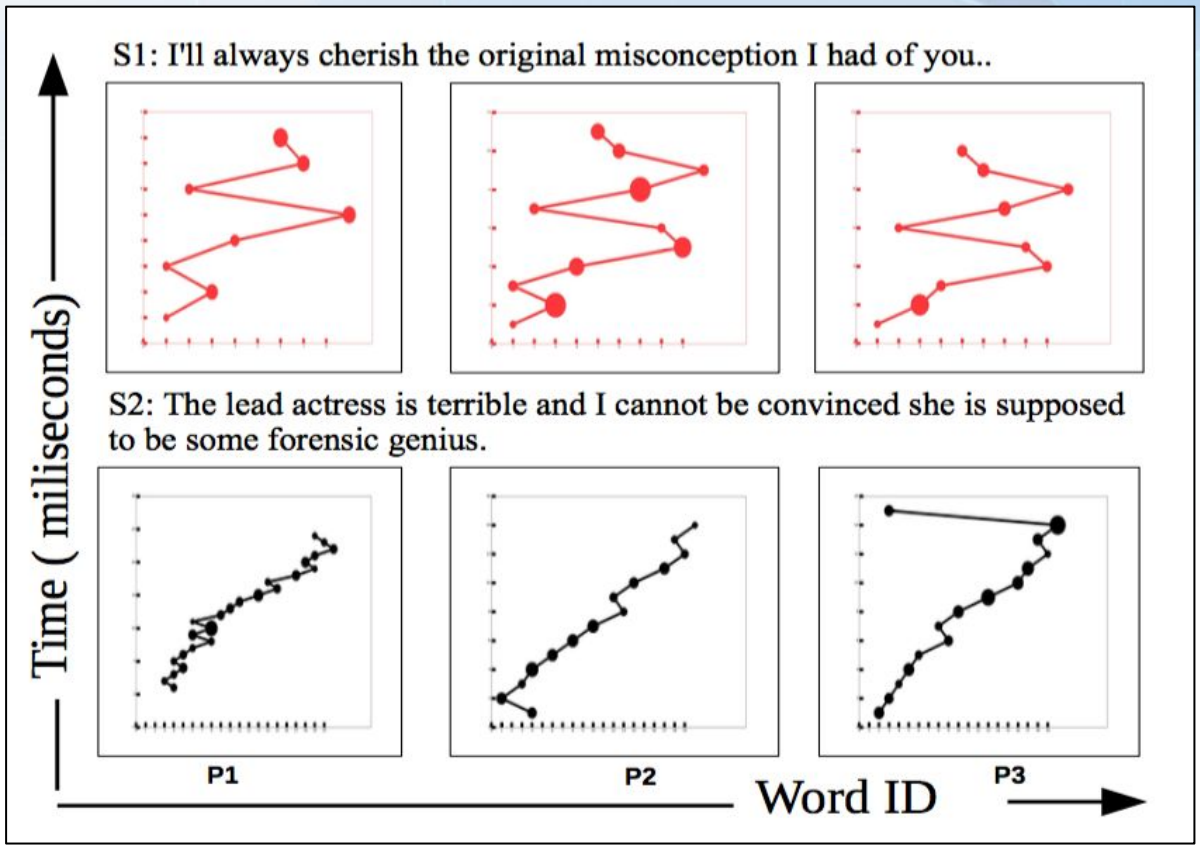


Image from Mishra A. et al (2016)

Types of NLP Applications

ANALYSIS

Spam Filtering

Abusive/Toxic Language Det

Sentiment Analysis

Sarcasm/Humor Detection

Text Grading

…

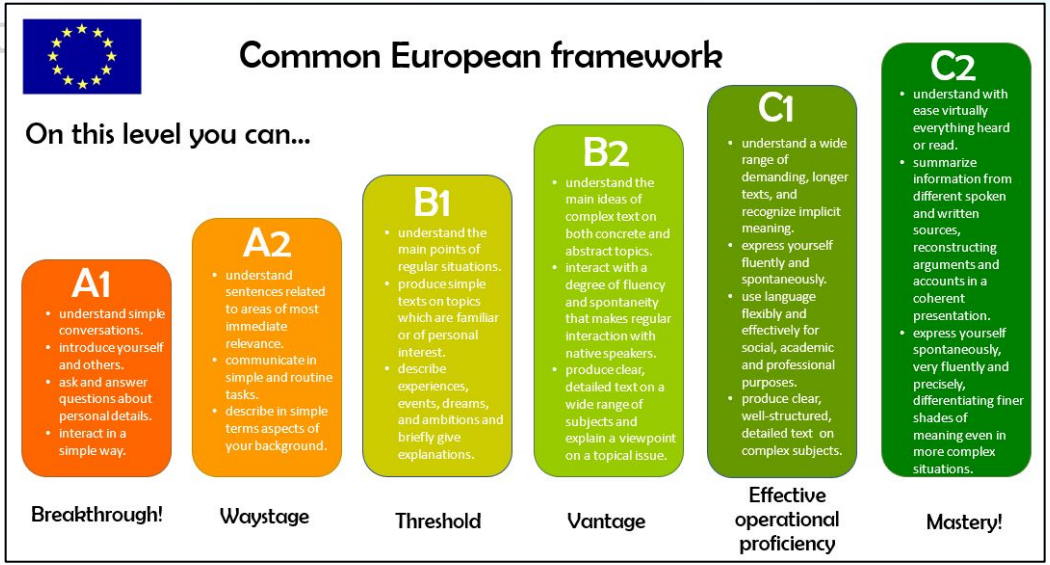


Image from http://mayraspanishschool.com/

Types of NLP Applications

ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

Sarcasm/Humor Detection

Text Grading

Text Mining

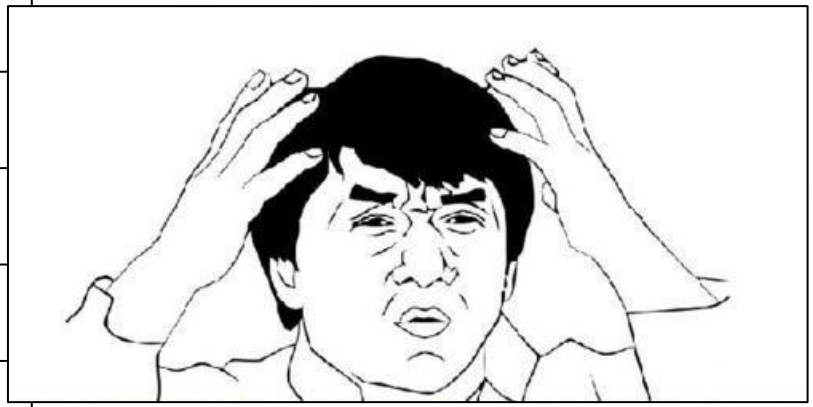
…

Text Mining

Image from www.restaurantlechristine.com/

Text Mining

Name	Le Christine	
Email		
Address	1 rue Christine, 75006 Paris	
Phone	+33 1 40 51 71 64	
Opening hours	Mon	12:00-14:30, 18:30-00:00
	...	12:00-14:30, 18:30-00:00
	Fri	12:00-14:30, 18:30-00:00
	Sat	18:30-00:00
	Sun	18:30-00:00



Types of NLP Applications

ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

Sarcasm/Humor Detection

Text Grading

Text Mining

Fact/Event Extraction…

Fact Extraction

Bloomberg

Cantor Fitzgerald Sued by Partners Who Moved to Reorient

China Lawsuit

In 2011 Cantor filed a lawsuit in China against Boyer, Ainslie and other traders who left its Hong Kong office, accusing them of breaching their employment agreements and causing a 29 percent drop in average monthly revenue at the branch. Two years later, Cantor officials settled their claims against the former executives, according to filings with the Hong Kong Stock Exchange. The terms weren't made public.

Sheryl Lee, a Cantor spokeswoman, said today by phone that the company has a policy of not commenting on litigation.

Image from www.bloomberg.com/

Fact Extraction

Bloomberg

Cantor Fitzgerald Sued by Partners Who Moved to Reorient

China Lawsuit

In 2011 Cantor filed a lawsuit in China against Boyer, Ainslie and other traders who left its Hong Kong office, accusing them of breaching their employment agreements and causing a 29 percent drop in average monthly revenue at the branch. Two years later, Cantor officials settled their claims against the former executives, according to filings with the Hong Kong Stock Exchange. The terms weren't made public.

Sheryl Lee, a Cantor spokeswoman, said today by phone that the company has a policy of not commenting on litigation.

Image from www.bloomberg.com/

Types of NLP Applications

TRANSFORMATION

Machine Translation

…

Transformations in MT

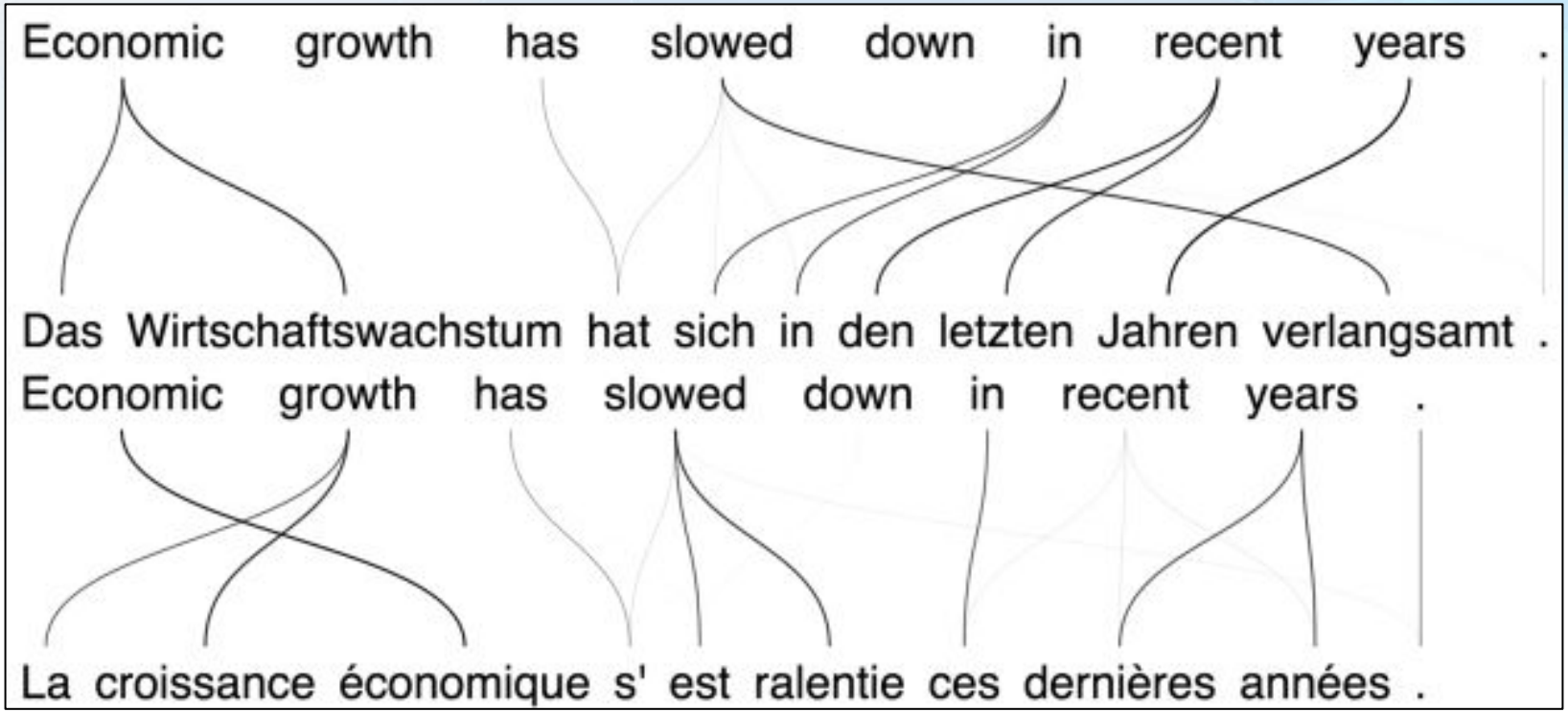


Image from Kyunghyun Cho (2015)

Types of NLP Applications

TRANSFORMATION

Machine Translation

Error Correction

…

Types of NLP Applications

TRANSFORMATION

Machine Translation

Error Correction

Speech to Text / Text to Speech

…

Types of NLP Applications

TRANSFORMATION

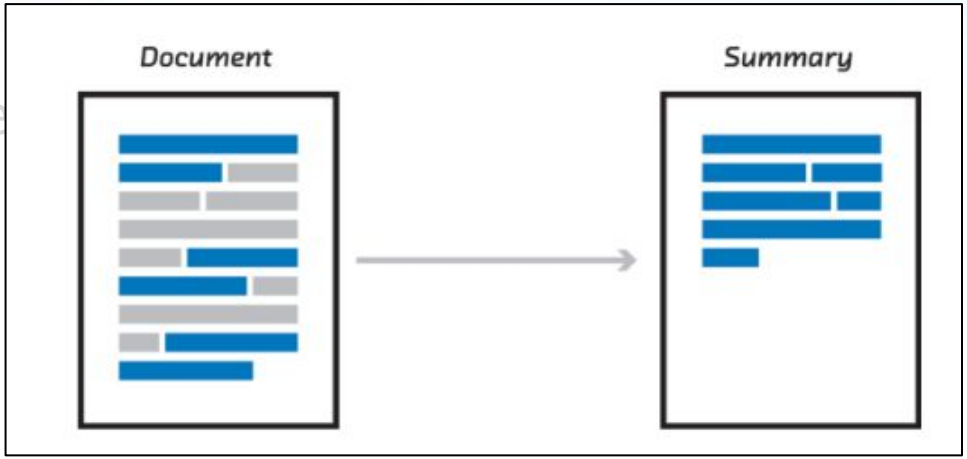
Machine Translation

Error Correction

Speech to Text / Text to Spee

Text Summarization

...



Types of NLP Applications

TRANSFORMATION

Machine Translation

Error Correction

Speech to Text / Text to Speech

Text Summarization

Text Simplification

...

Text Simplification

They are humid, prepossessing Homo Sapiens with full-sized aortic pumps.

Text Simplification

They are humid, prepossessing Homo Sapiens with full-sized aortic pumps.

They are warm, nice people with big hearts.

Types of NLP Applications

TRANSFORMATION

Machine Translation

Error Correction

Speech to Text / Text to Speech

Text Summarization

Text Simplification

Text Anonymization

...

Image from www.ntt-review.jp

Text Anonymization

Original:

> Jack and Jill Robinson bought a car at Toyota Motor for $400K on May 13th, 2011.

Text Anonymization

Original:

Jack and Jill Robinson bought a car at Toyota Motor for $400K on May 13th, 2011.

Anonymized:

Boris and Althea Stephanopoulos bought a car at BimBom Motor for €120K on March 21st, 2001.

Types of NLP Applications

Generation

Conversational Agents

…

Siri

*"I remember the first time we loaded these data sources into Siri. I typed **"start over"** into the system, and Siri came back saying,*

"Looking for businesses named 'Over' in Start, Louisiana."

— Adam Cheyer

Taken from https://medium.com/swlh/the-story-behind-siri-fbeb109938b0

The story of Tay

Types of NLP Applications

Generation

Conversational Agents

Story Cloze Task

…

Story Cloze

Tom and Sheryl have been together for two years. One day, they went to a carnival. Tom won Sheryl several stuffed bears. When they reached the Ferris wheel, he got down on one knee.

Which ending is more probable?

- *Tom asked Sheryl to marry him.*
- *He wiped mud off of his boot.*

Taken from Mostafazadeh N. et al. (2016)

Types of NLP Applications

Generation

Conversational Agents

Story Cloze Task

Computer-Generated Text

…

It is now exactly 05:00 in Samarkand. In some ramshackle dwelling a person who is called Gang, who is on the small side, reads an entirely made-up word on a box of breakfast cereal. He turns entirely around.

It is now right about 18:01 in Matamoros. In some dim yet decent structure a man named Tao, who is no larger or smaller than one would expect, reads a tiny numeric code from a recipe clipping. He smiles a tiny smile.

It is now as it happens 19:02 in Grand Turk. In some sturdy yet undistinguished habitat a youth named Peng, who is quite sizable and imposing, reads a stained card. He sits up straight.

It is now only a moment before 02:03 in Windhoek. In some suitable structure someone named Ezra, who is significantly smaller than others of the same age, reads a canary-colored manuscript. He hums quietly.

Taken from Nick Montfort's http://nickm.com/code/world_clock.py (2013)

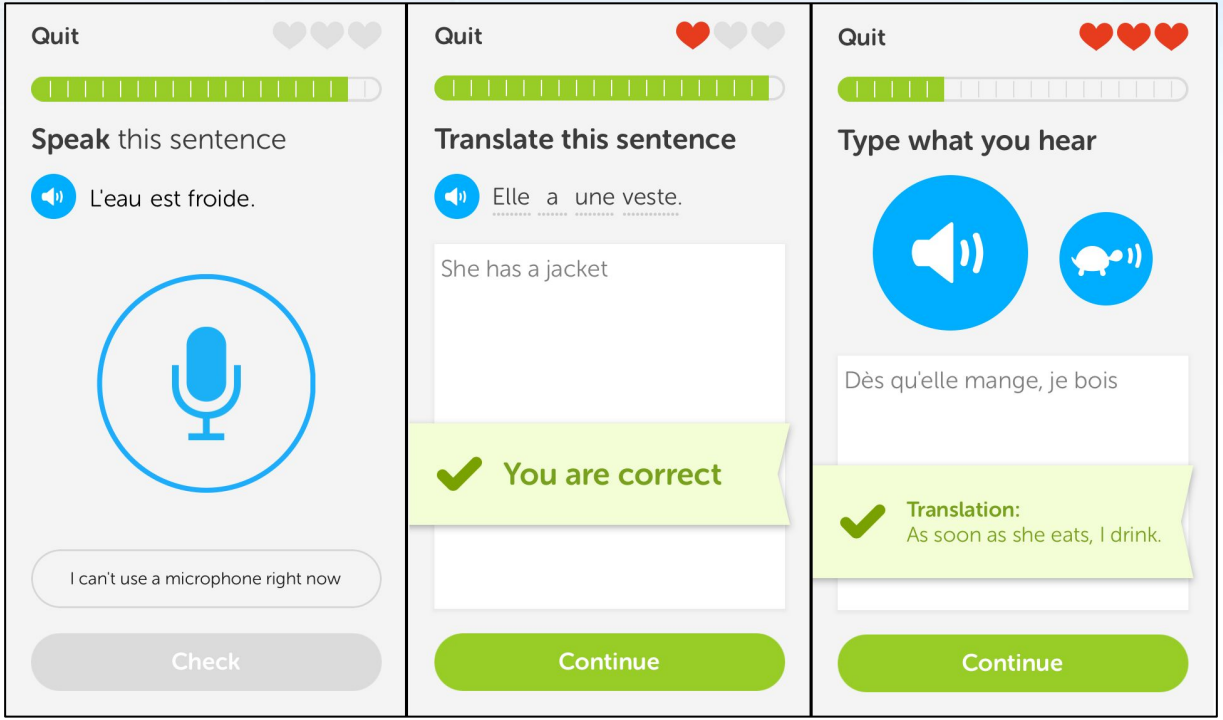
Computer-Generated Text

OpenAI's latest language model (2019)

SYSTEM PROMPT (HUMAN-WRITTEN)	*A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.*
MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)	The incident occurred on the downtown train line, which runs from Covington and Ashland stations. In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief. "The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."

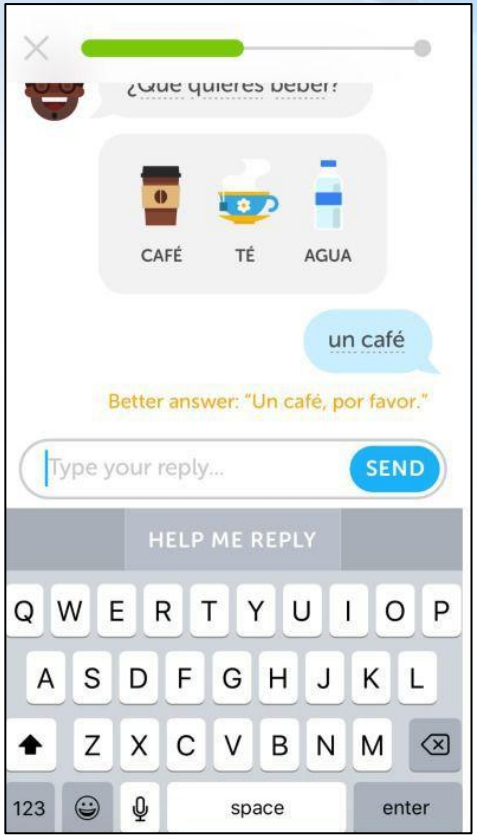
Types of NLP Applications

Language Learning

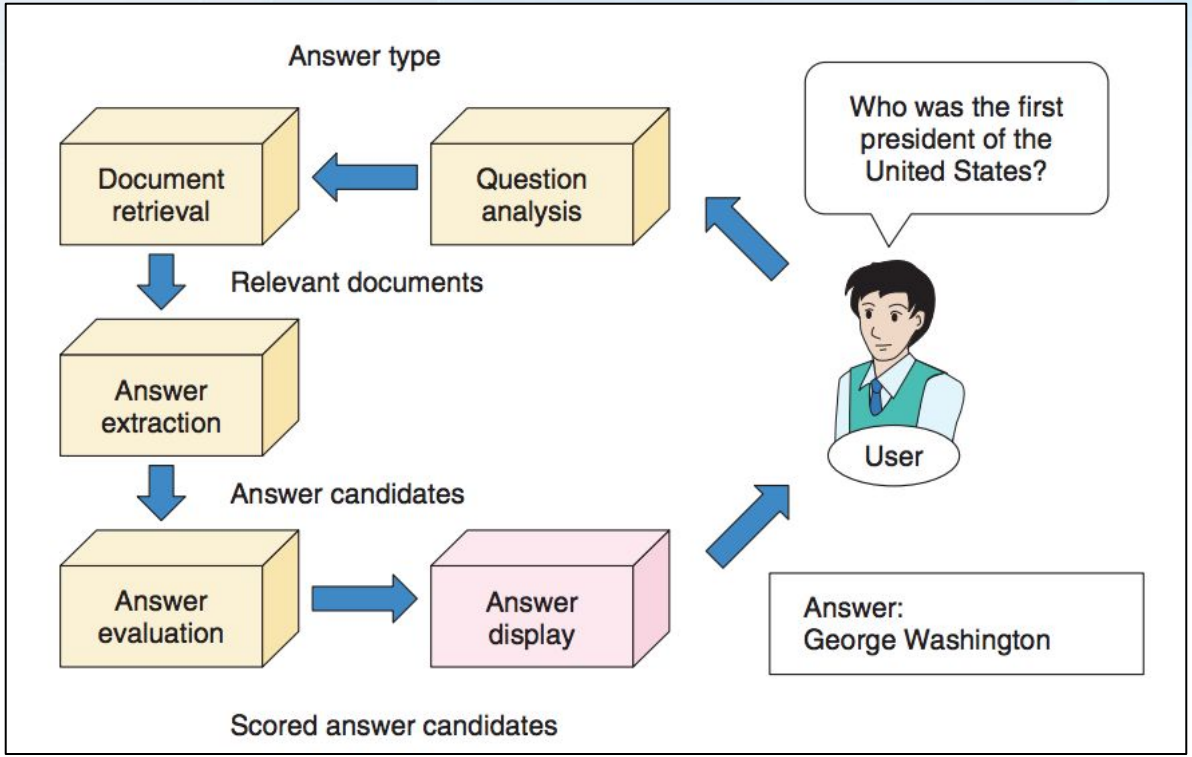


Duolingo


Hola
¿Cómo estás?
Bien tú?
¡Qué bien! Yo estoy bien.
Conoce a mis amigos. Primero, ¡dile "hola" a Roberto!
Type in Spanish
SEND
Somebody else is typing...
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123
space
return


CAFÉ
TÉ
AGUA
un café
Better answer: "Un café, por favor."
Type your reply...
SEND
HELP ME REPLY
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123
space
enter

IBM Watson


Answer type
Who was the first president of the United States?
Document retrieval
Question analysis
Relevant documents
Answer extraction
Answer candidates
User
Answer evaluation
Answer display
Answer: George Washington
Scored answer candidates

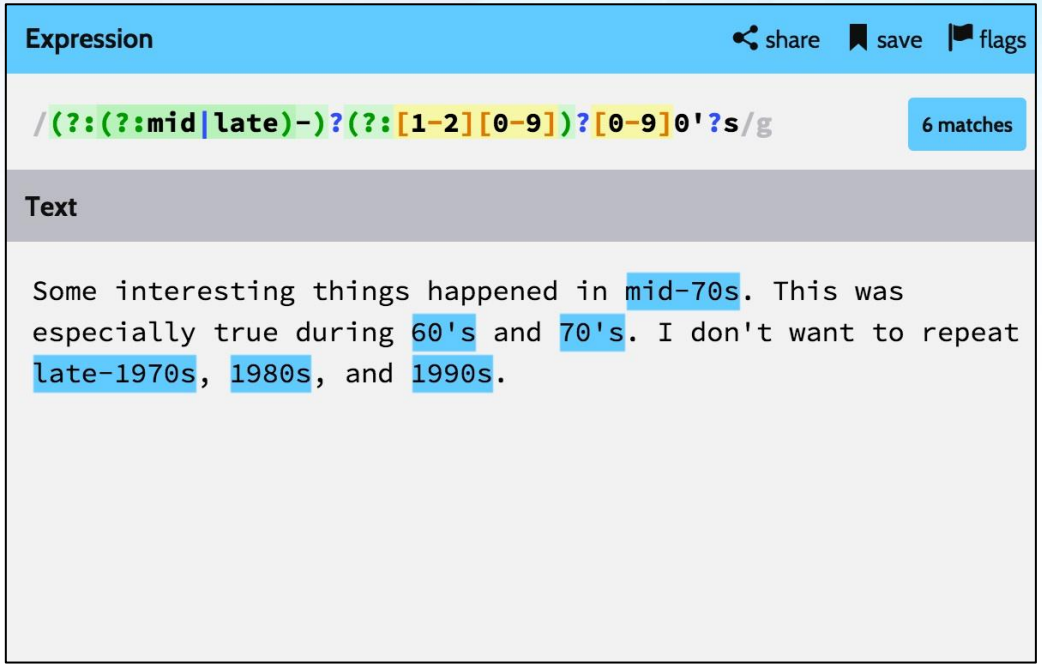
2. Competencies of a computational linguist

Competencies

- Basic tech skills
- Linguistics
- Computer Science
- NLP technologies

Basic tech skills

- Regular expressions
- Shell commands
- Smart text editors
  - Sublime Text 3
  - Emacs
  - Atom
  - Notepad++



Linguistics

- Pattern matching
- Structural linguistics
- Linguistic ambiguities

Structural linguistics

Claims:

- ***language*** is an object that can be described and decomposed
- ***language*** has clear structure and levels

The task:

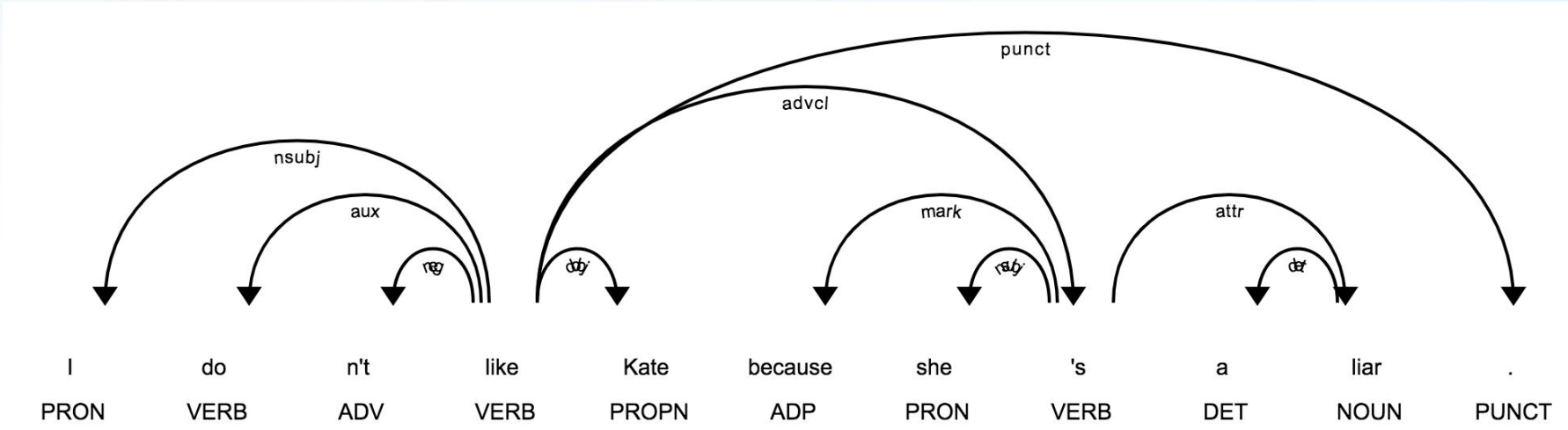
- develop algorithms to extract ***features*** from language
- develop algorithms that use the extracted ***features*** to solve the broader task

For example

I don't like Kate because she's a liar.

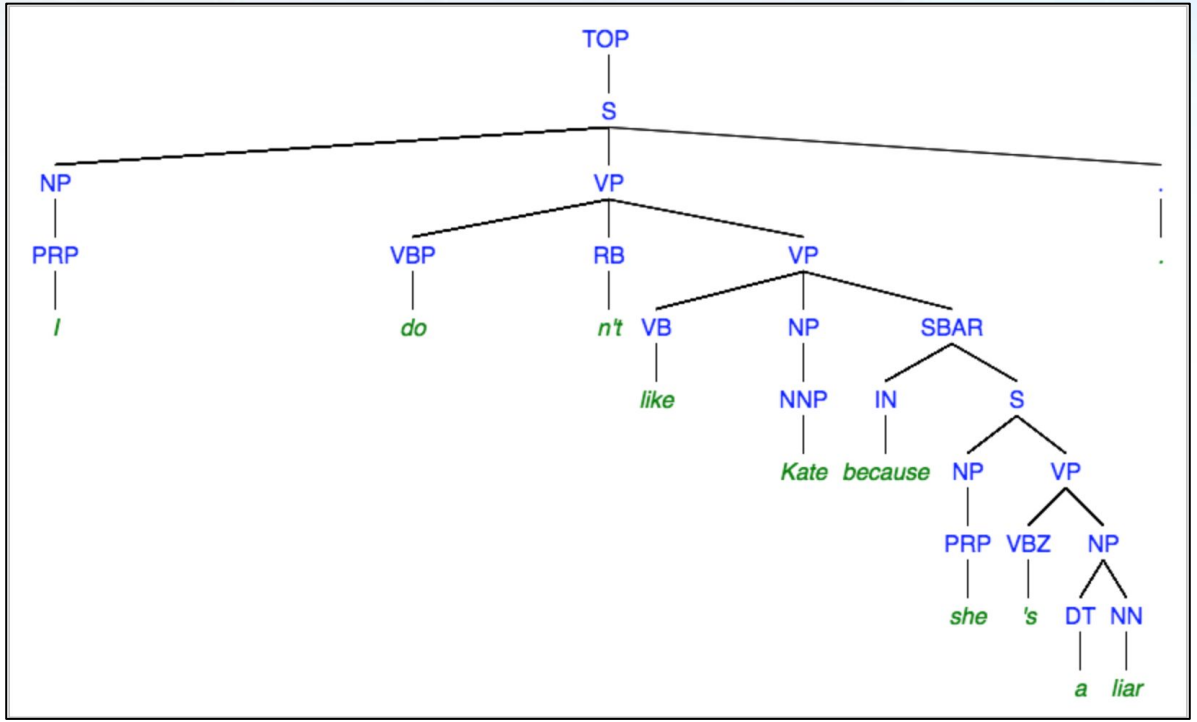
For example

Dependency tree:



For example

Constituency tree:



For example

Named Entities:

I do n't like ***Kate_PERSON*** because she 's a liar .

For example

Coreference:

I do n't like ***Kate*** because ***she*** 's a ***liar*** .

For example

Semantic roles:

I do n't like ***Kate*** because she 's a liar .

I - agent

Kate - patient

And many-many more language features...

Linguistic ambiguities

At every level:

- phonetics

Linguistic ambiguities

At every level:

- phonetics
- morphology

an *un·ion·ized*

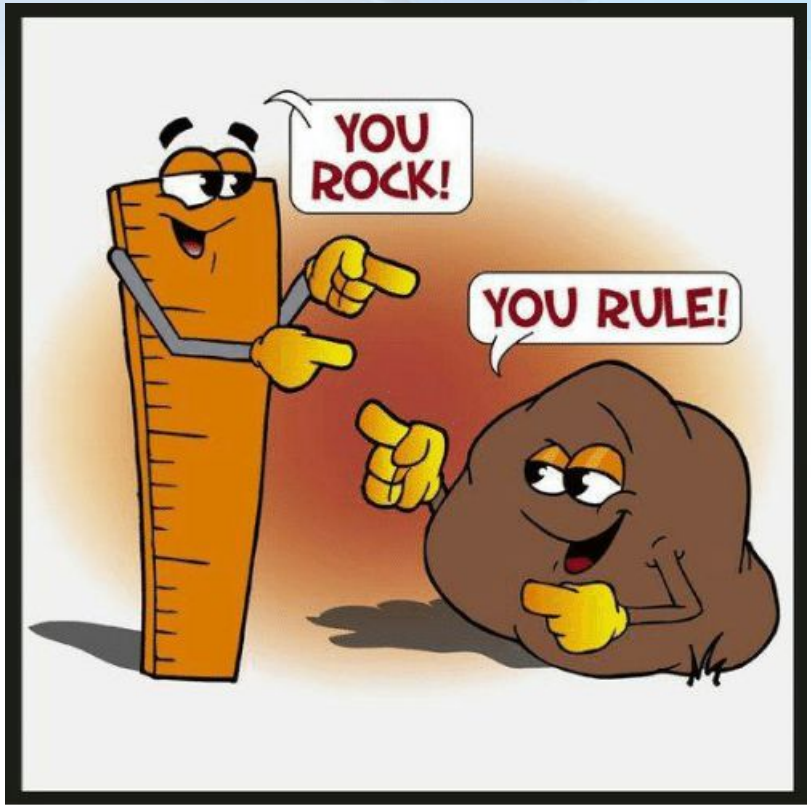
vs.

a *union·ized*

Linguistic ambiguities

At every level:

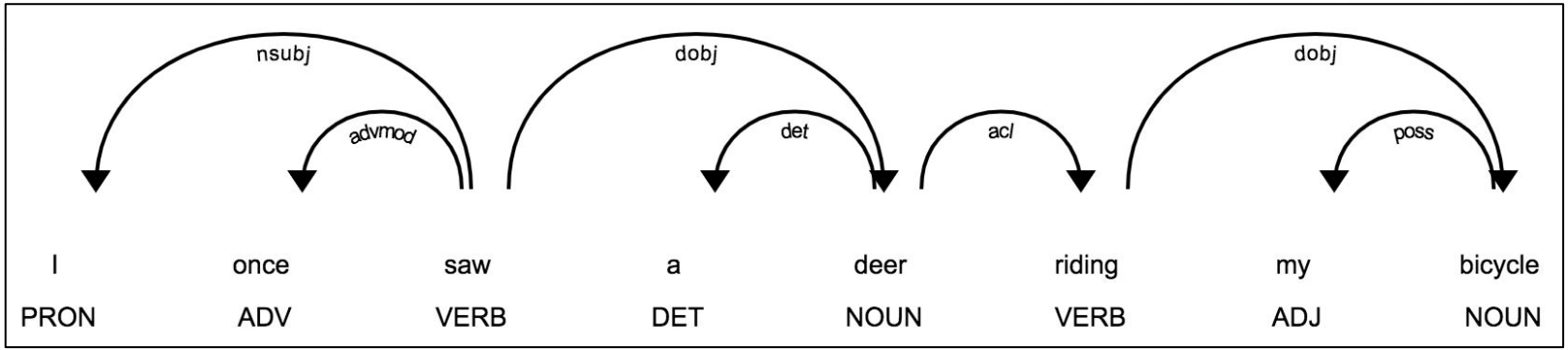
- phonetics
- morphology
- parts of speech



Linguistic ambiguities

At every level:

- phonetics
- morphology
- parts of speech
- syntax



Linguistic ambiguities

At every level:

- phonetics
- morphology
- parts of speech
- syntax
- semantics

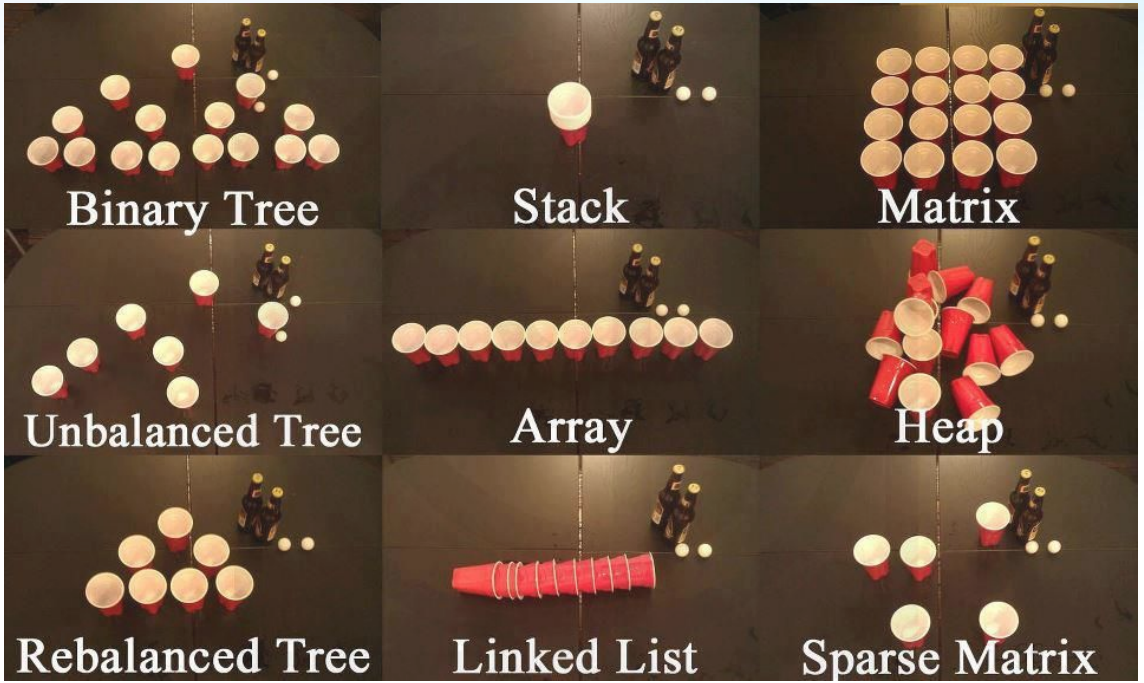
Linguistic ambiguities

At every level:

- phonetics
- morphology
- parts of speech
- syntax
- semantics
- pragmatics

Computer Science

- Mathematical statistics and theory of probability
- Scripting / OOP
- Scraping
- Algorithms
- Data structures



NLP technologies

- NLP libraries
  - nltk, spaCy, StanfordCoreNLP, OpenNLP, EmoryNLP...
- NLP algorithms
  - rule-based
  - statistical
  - machine learning
- NLP resources
  - corpora, dictionaries, ontologies, word embeddings…
- NLP methodology

A classic NLP analysis Pipeline

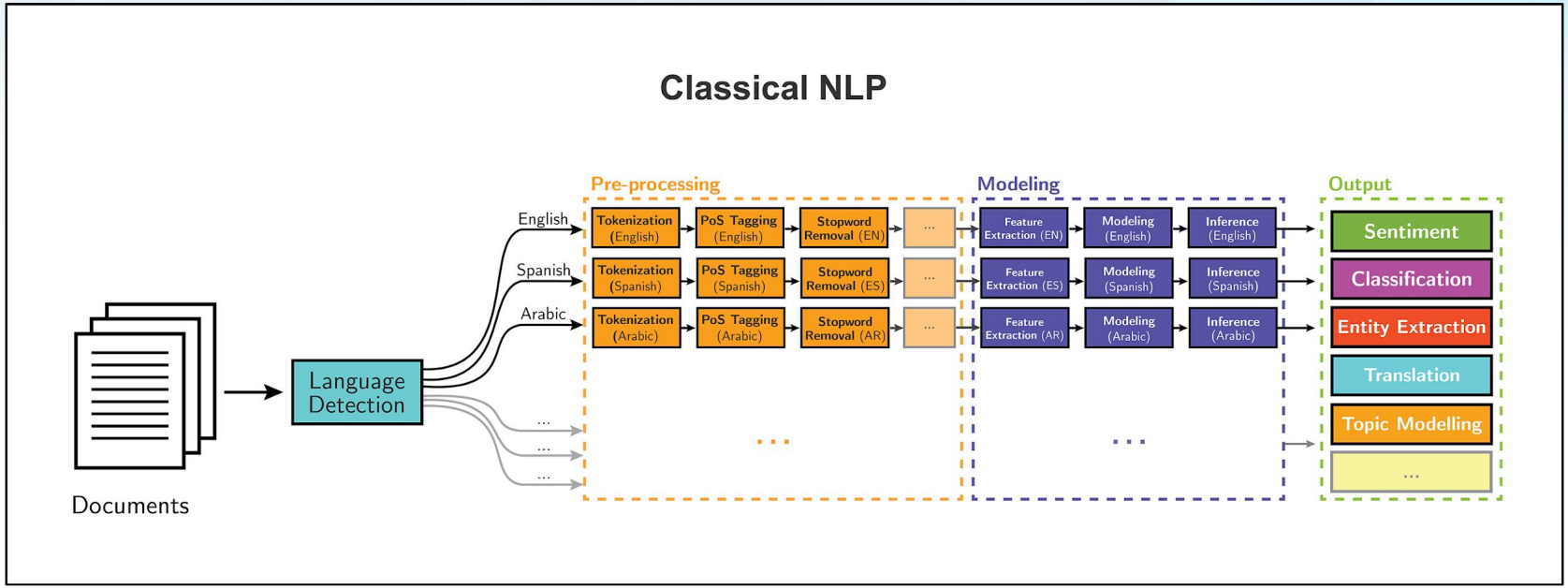


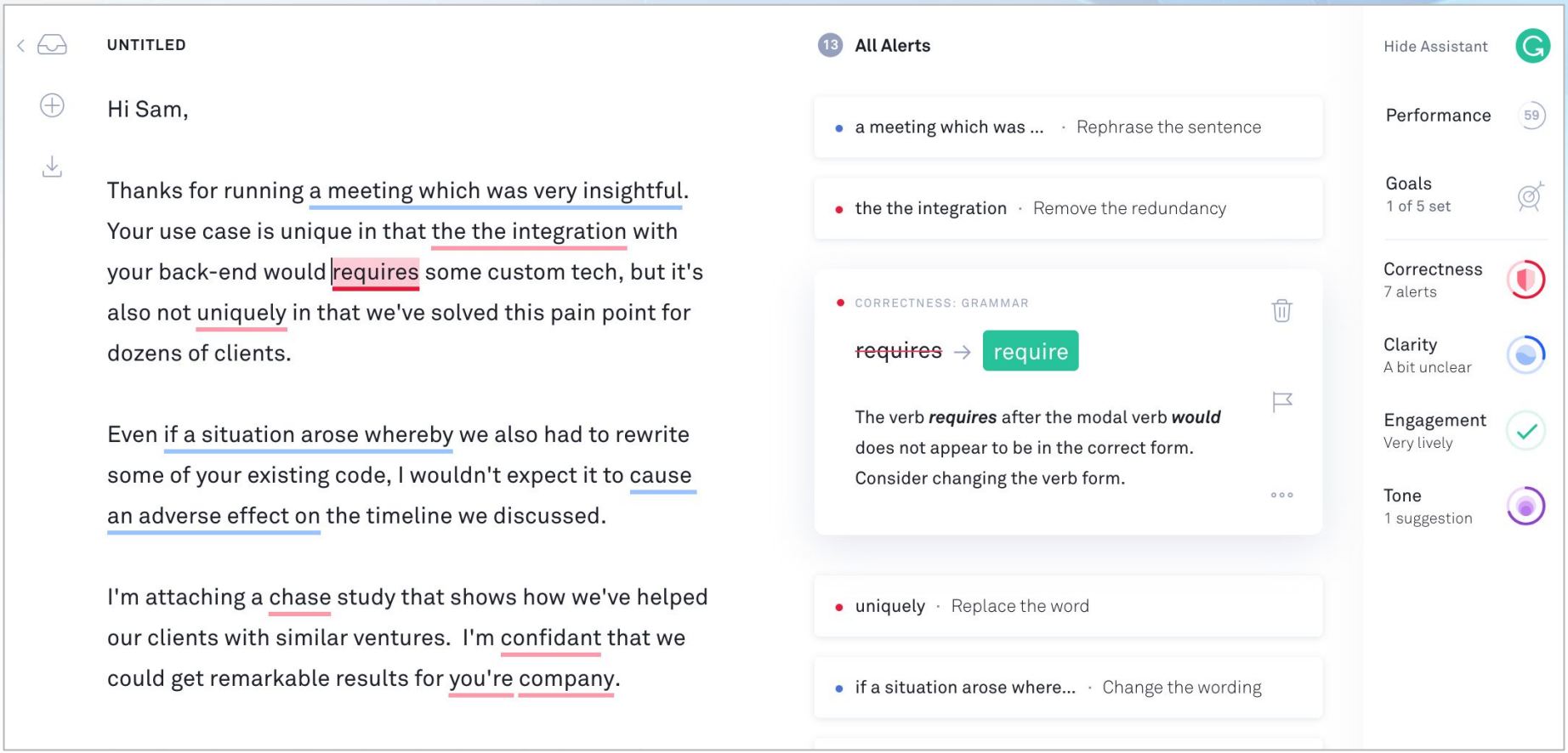
Image from http://blog.aylien.com/

3. Grammarly and CompLing Summer School

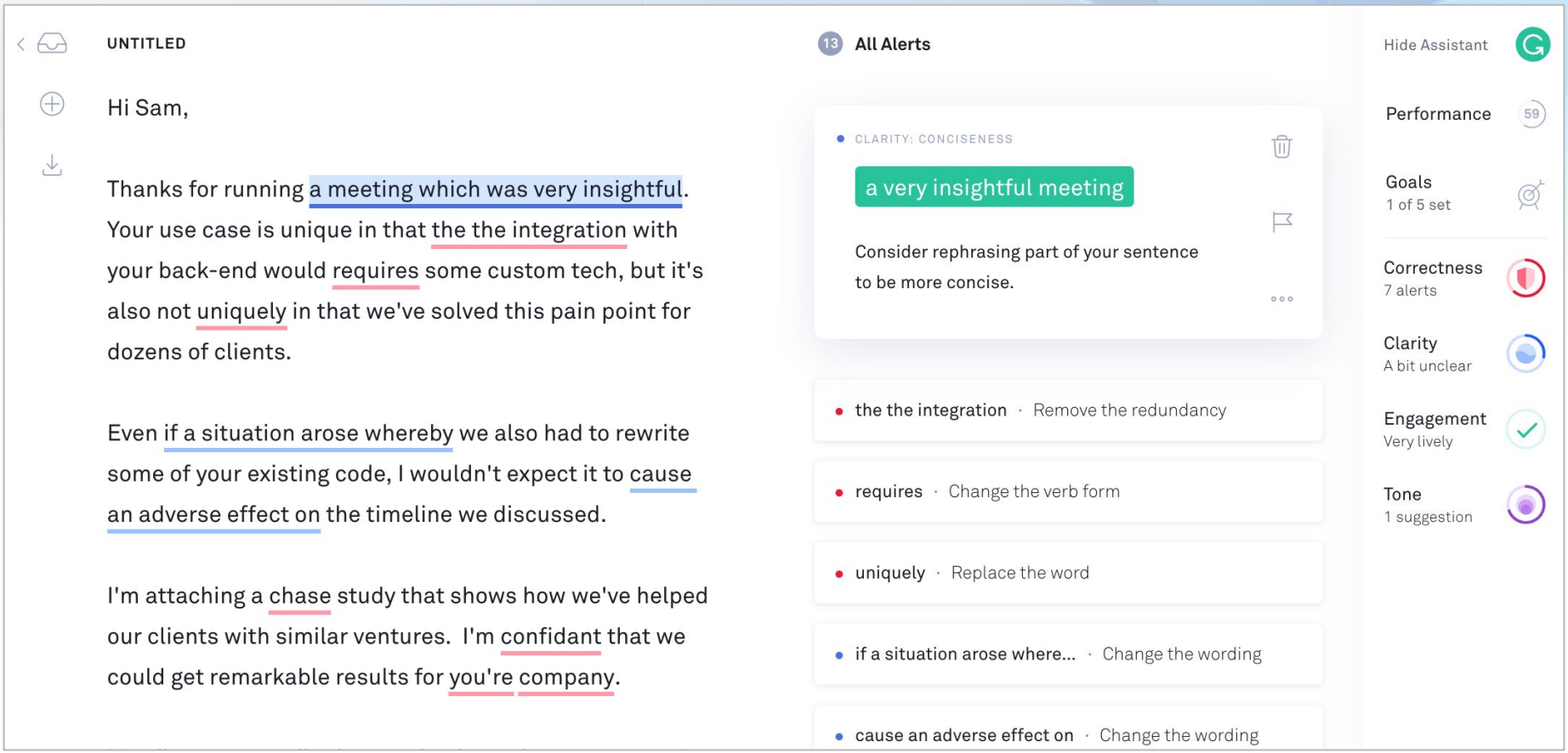
What is Grammarly?

A writing assistant that helps make your communication clear and effective, wherever you type.

What is Grammarly?



What is Grammarly?



What is Grammarly?

Available for **20M** daily users in:

- Online Editor
- Browser extension for Chrome, Safari, Firefox, Edge
- Desktop app for Windows and macOS
- Add-in for Microsoft® Office
- Mobile Keyboard for iOS and Android

What is Grammarly?

Motivation:

- an average non-native speaker makes one mistake per every ten words

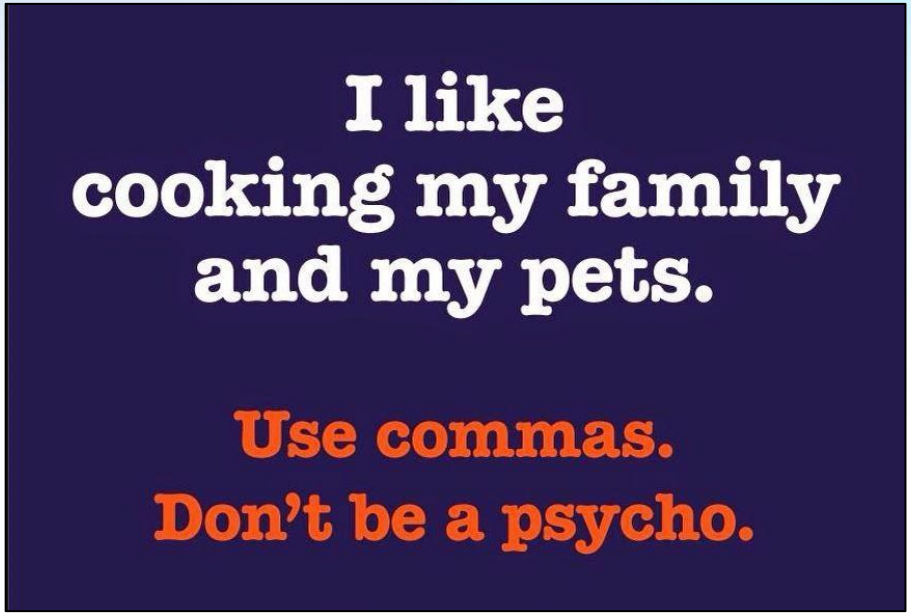
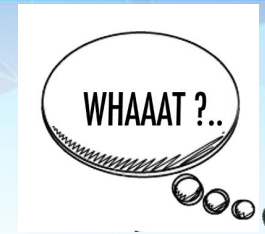
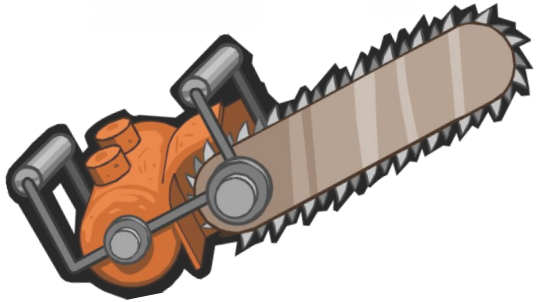


Image from www.writing.com/

Error correction



She sawed a black cat in the room.



CompLing Summer School

When: July 13-18, 2020; 9 a.m. – 7 p.m.

Where: Grammarly office in Kyiv

Who: applied linguists whose interests lie in the area of computational linguistics

The school is **free of charge**.

CompLing Summer School

Syllabus

- regular expressions
- smart text editors
- basic text processing with Python
- processing of corpora and dictionaries
- statistical text analysis
- NLP libraries, POS tagging and syntactic parsing

CompLing Summer School

You can apply if you have

- background in linguistics
- knowledge of structural linguistics
- basic level of Python programming

  (variables, if-conditions, cycles, functions)
- upper-intermediate level of English

CompLing Summer School

Timeline

- April 20 – registration deadline
- April 21 – the test task is sent
- May 24 – deadline for sending your solution
- May 25 – June 5 – the interviews
- June 8–12 – notification of acceptance
- July 13–18 – see you in Kyiv ;)

Details and registration form:
http://bit.ly/CompLing2020

grammarly.ai

CompLing Workshop: Shallow Discourse Parsing

Saturday, February 29, 10 AM - 6 PM Grammarly Office

Tatjana Scheffler

Ph.D., University of Potsdam, Germany

Registration About Event

Please write to events@grammarly.com if you want to attend. And

Questions?

mariana.romanyshyn@grammarly.com
or find me on LinkedIn