Intro to Natural Language Processing

Mariana Romanyshyn, *Grammarly, Inc.*
Vsevolod Dyomkin, *Franz Inc.*

Contents

1. Overview of NLP
2. NLP applications in our world
3. Real-life projects by Vsevolod Dyomkin
4. Real-life projects by Mariana Romanyshyn

WTF NLP or NLP FTW?

The Goal of NLP

Goal:

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

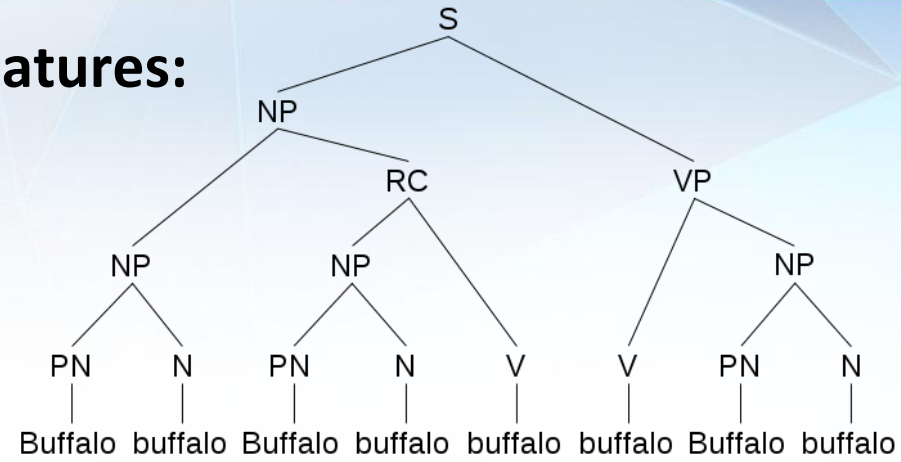
How:

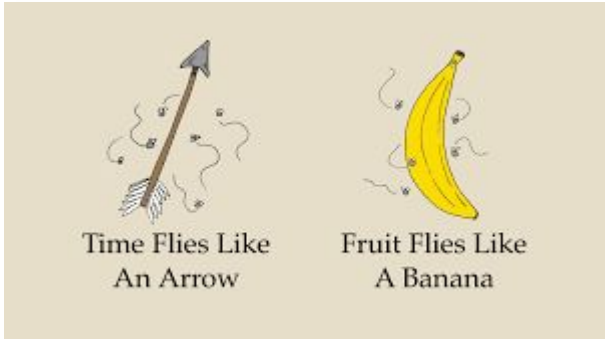
transform free-form text into structured data and back

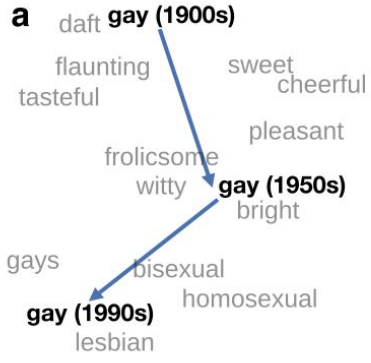
Natural Language

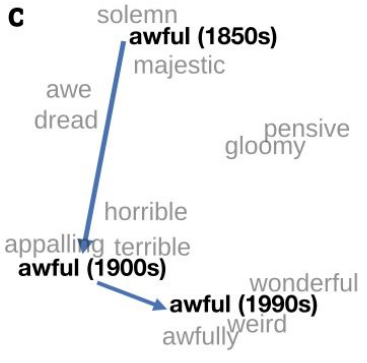
Distinguishing features:

- Ambiguous
- Noisy
- Evolving

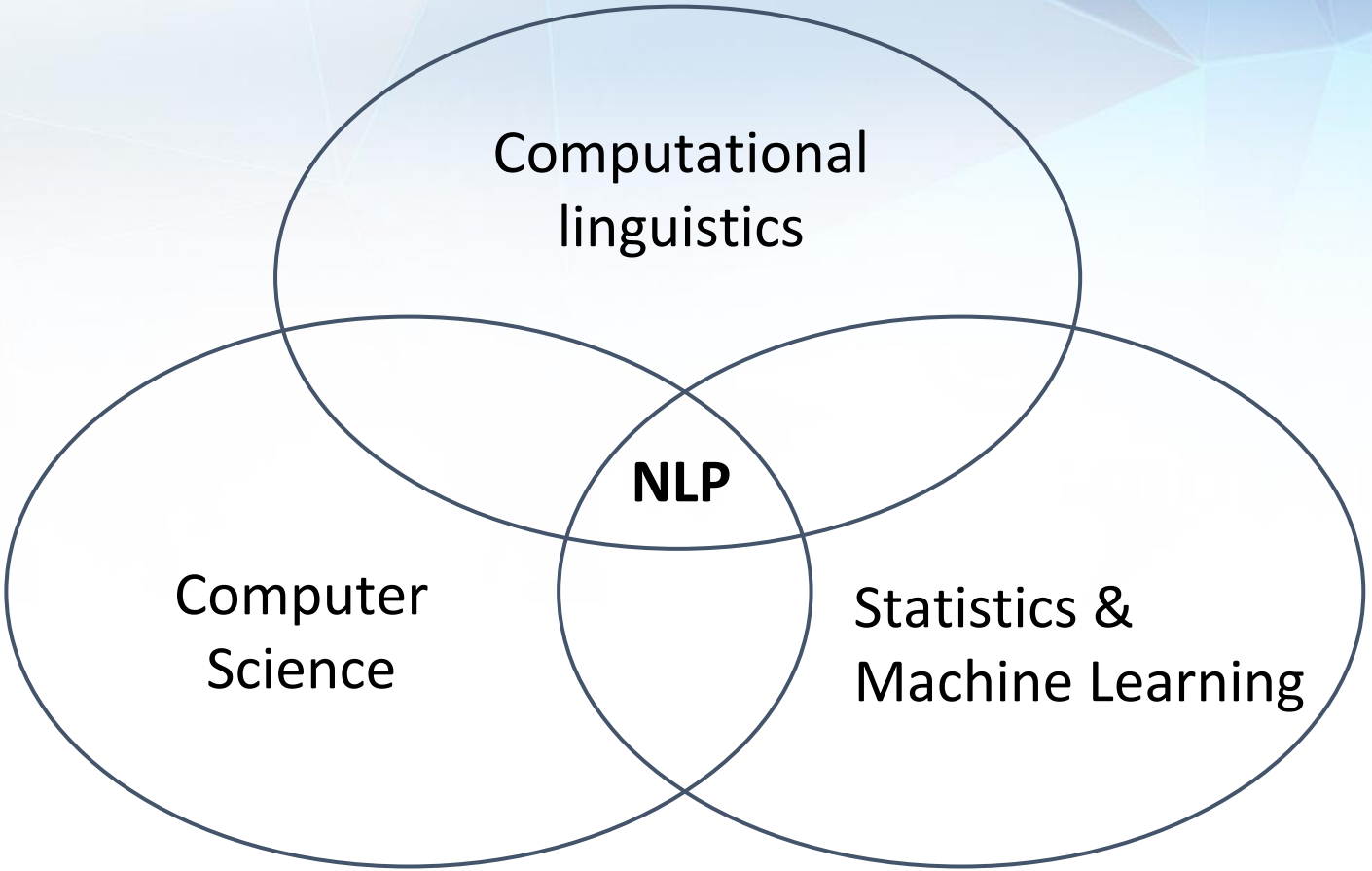




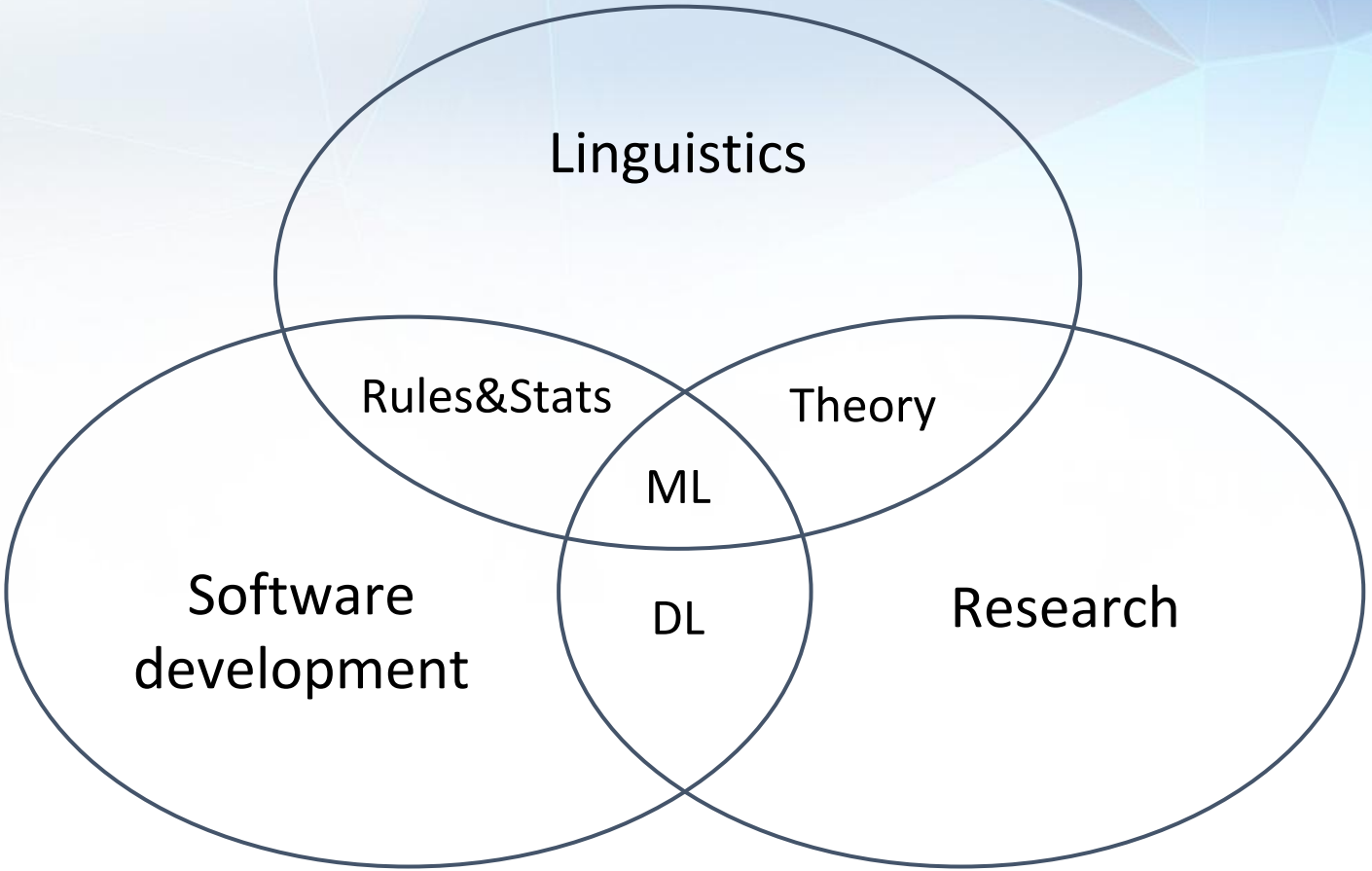




Position of NLP



Expertise in NLP



NLP & AI

CV vs NLP

NLP vs NLU

Are we there yet?

- http://nlpprogress.com/
- https://www.eff.org/ai/metrics

NLP applications in our world

Q: What NLP applications do you know?

A: https://github.com/Kyubyong/nlp_tasks

Types of NLP Applications

- Linguistic
- Analysis
- Transformation
- Generation

An NLP Pipeline

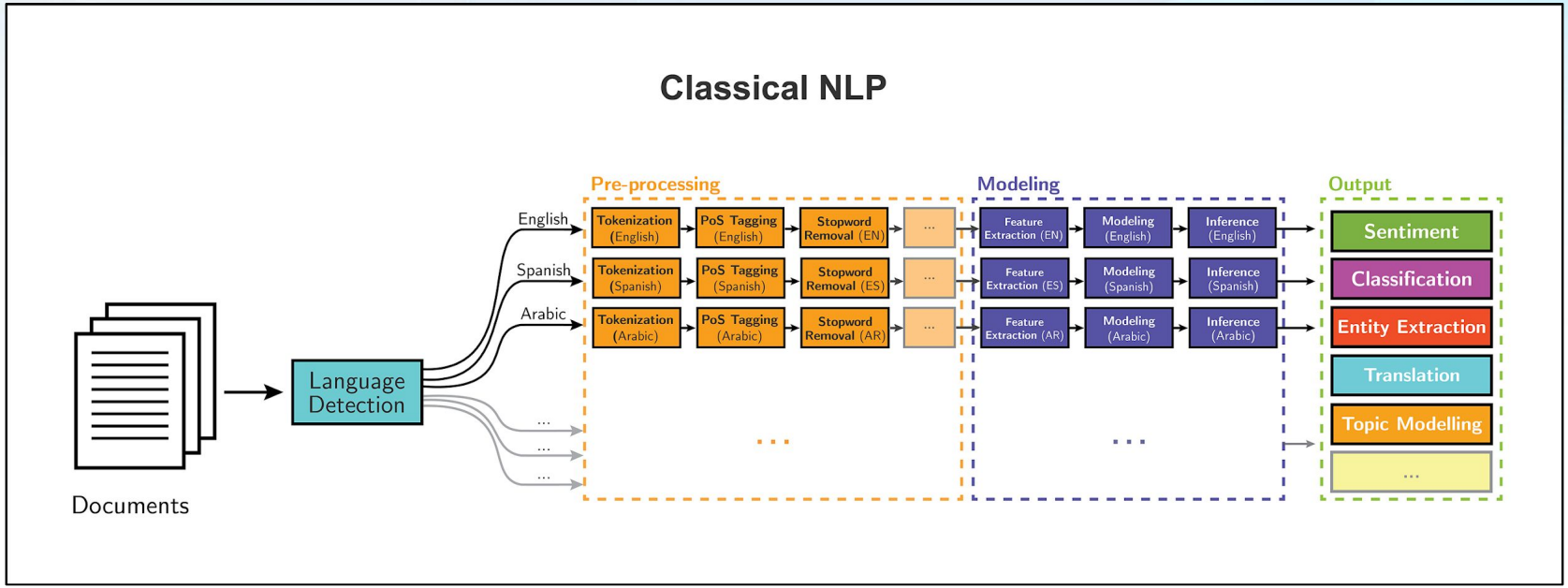


Image from http://blog.aylien.com/

An NLP Pipeline

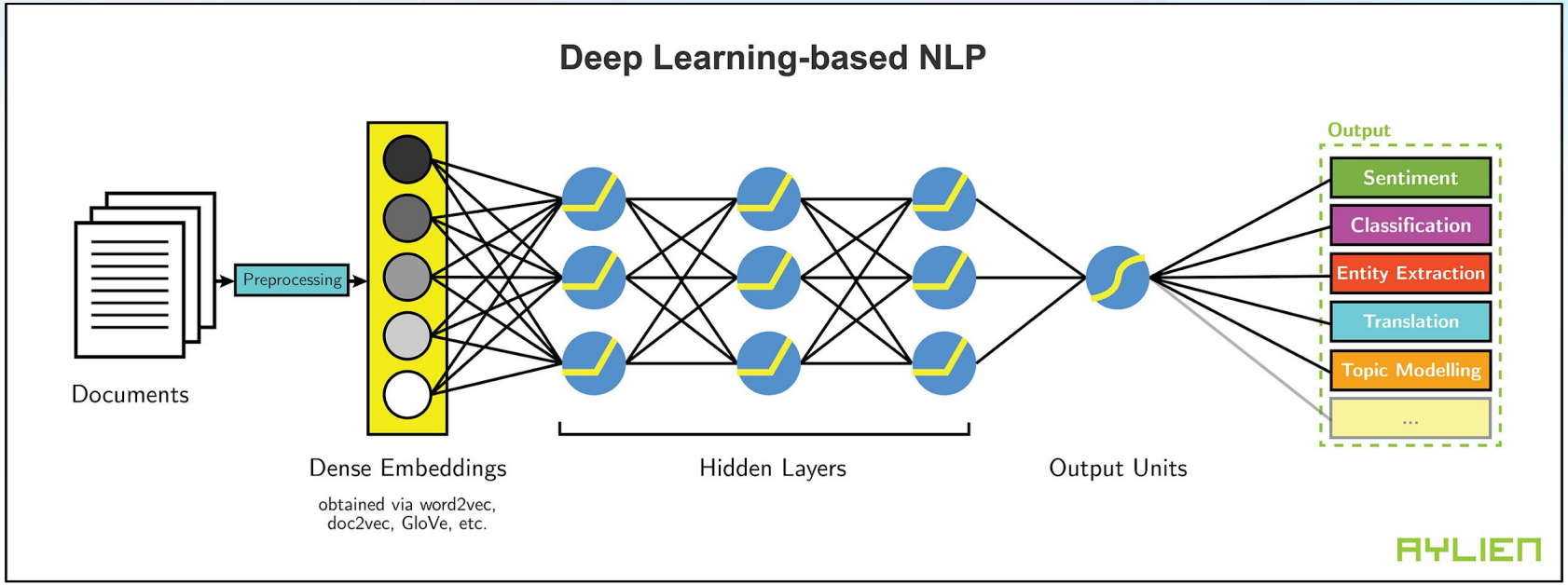


Image from http://blog.aylien.com/

Types of NLP Applications

Linguistic tasks

- Segmentation
- Part of speech tagging
- Named-entity recognition
- Relation extraction
- Syntactic parsing
- Coreference resolution
- Semantic parsing
- ...

Types of NLP Applications

ANALYSIS

Abusive/Toxic Language Detection

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

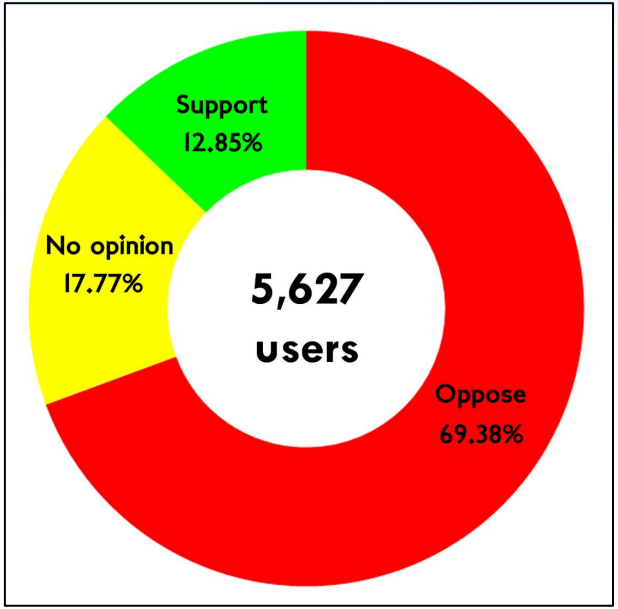
Types of NLP Applications

ANALYSIS

Abusive/Toxic Language Detection

Sentiment Analysis

…



Sentiment maps

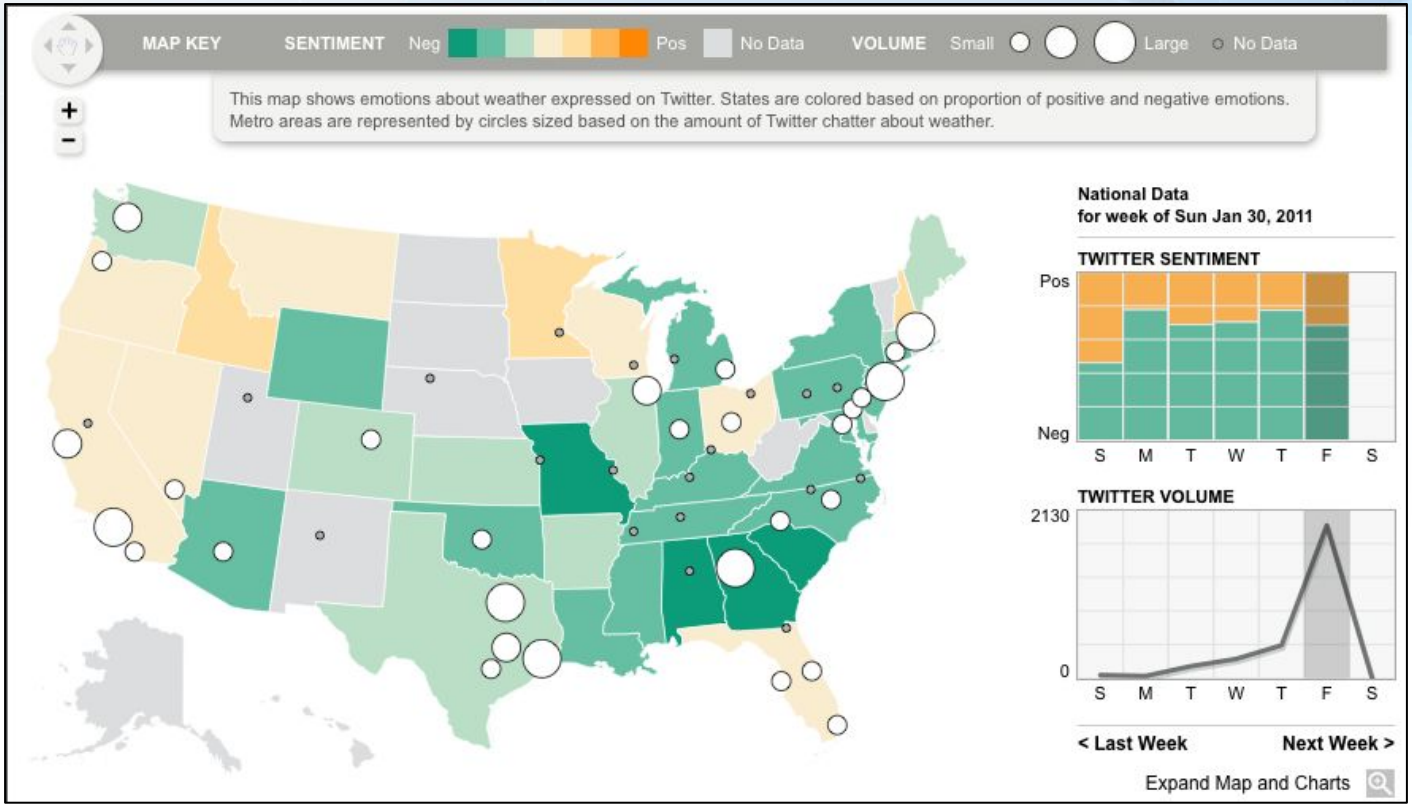


Image from http://www.dialogueearth.org/

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

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

Types of NLP Applications

ANALYSIS

Abusive/Toxic Language Detection

Sentiment Analysis

- sentiment scale
- type of emotion
- subjectivity

Types of NLP Applications

ANALYSIS

Abusive/Toxic Language Detection

Sentiment Analysis

Sarcasm Detection

Humor Detection

…

Cognitive features

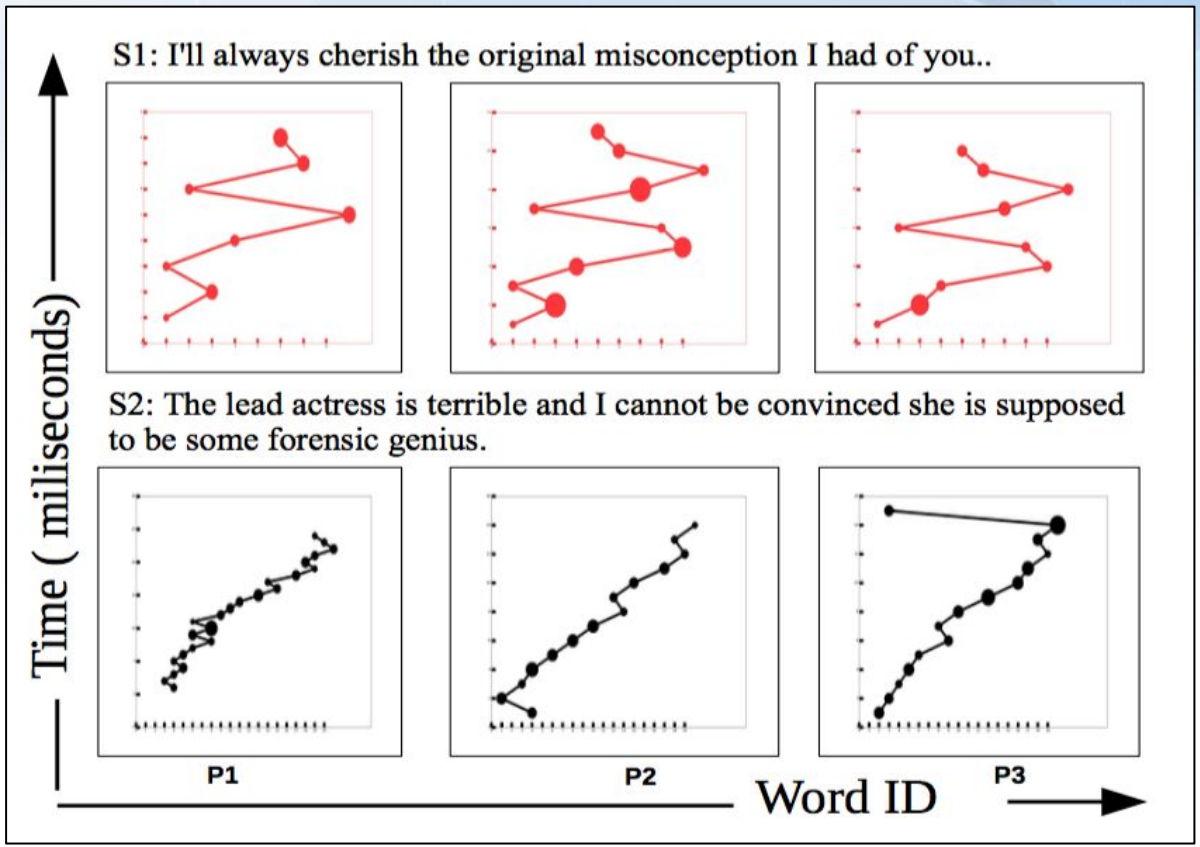


Image from Mishra A. et al (2016)

Types of NLP Applications

ANALYSIS

Abusive/Toxic Language Detection

Sentiment Analysis

Sarcasm Detection

Good/Evil Characters

…

Phonological features

Most frequent in positive characters	
Phoneme n-gram	Examples
/lɪ/	Ned Alleyn (Shakespeare in Love)
/an/	Anouk Rocher (Chocolat)
/aɪ/	Eliza Doolittle (My Fair Lady)
/nɪ/	Linguini (Ratatouille)
/ɪst/	Kevin McCallister (Home Alone)
/ɹəʊ/	Frodo (The Lord of the Rings)
/and/	Dylan Sanders (Charlie's Angels)
/stə/	C.C. Baxter (The Apartment)

Most frequent in negative characters	
Phoneme n-gram	Examples
/ən/	Tom Buchanan (The Great Gatsby)
/əʊ/	Iago (Aladdin)
/tə/	Norrington (Pirates of the Caribbean)
/ɹɪ/	Tom Ripley (The Talented Mr. Ripley)
/mən/	Norman Bates (Psycho)
/mɪs/	Mystique (X-Men)
/ktə/	Hannibal Lecter (Hannibal)

Image from Papantoniou K. and Konstantopoulos S. (2016)

Types of NLP Applications

TRANSFORMATION

Machine Translation

…

Transformations in MT

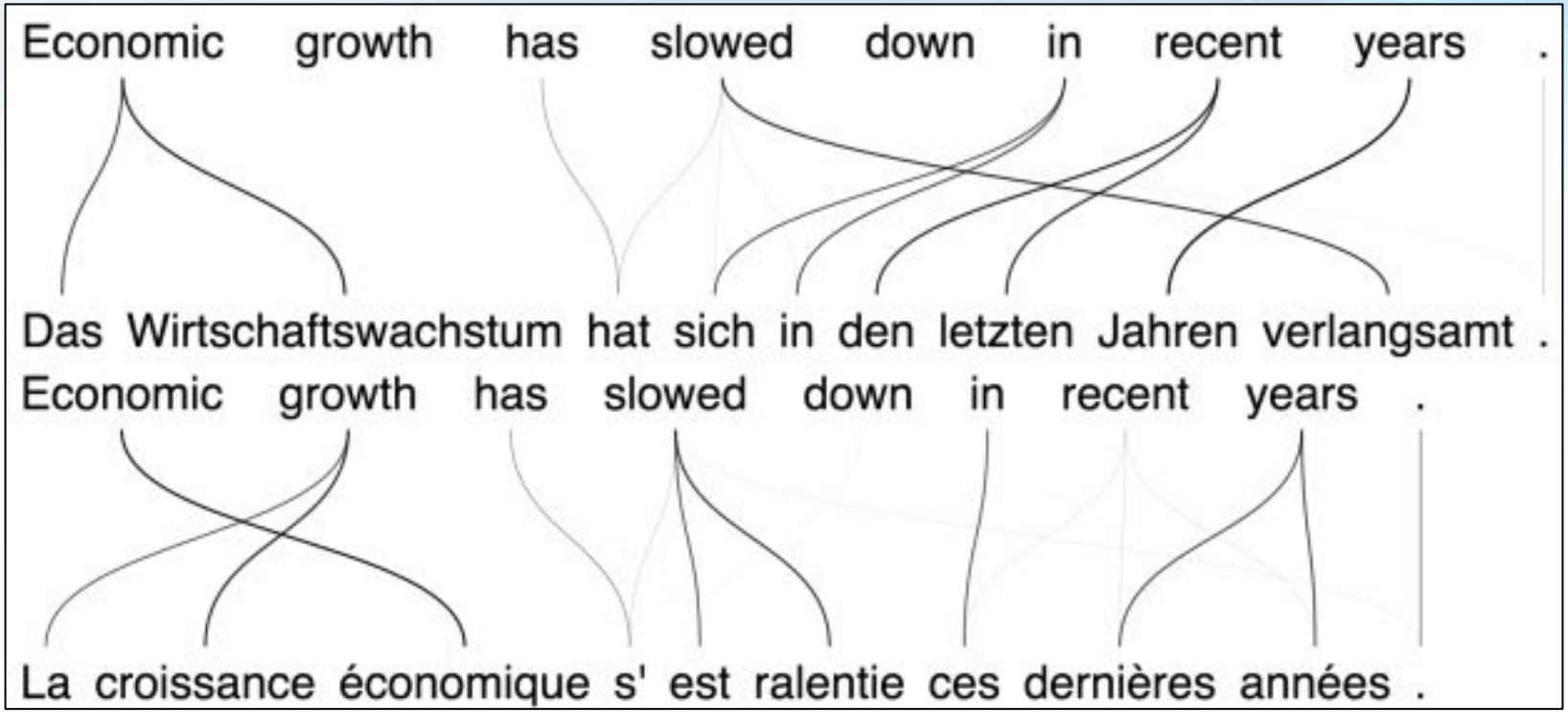


Image from Kyunghyun Cho (2015)

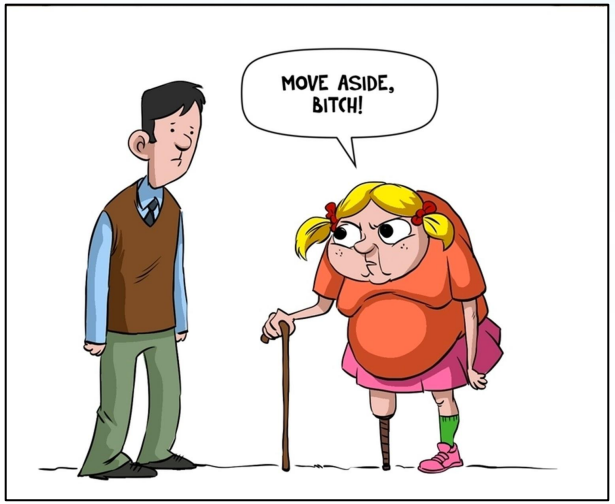
Types of NLP Applications

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

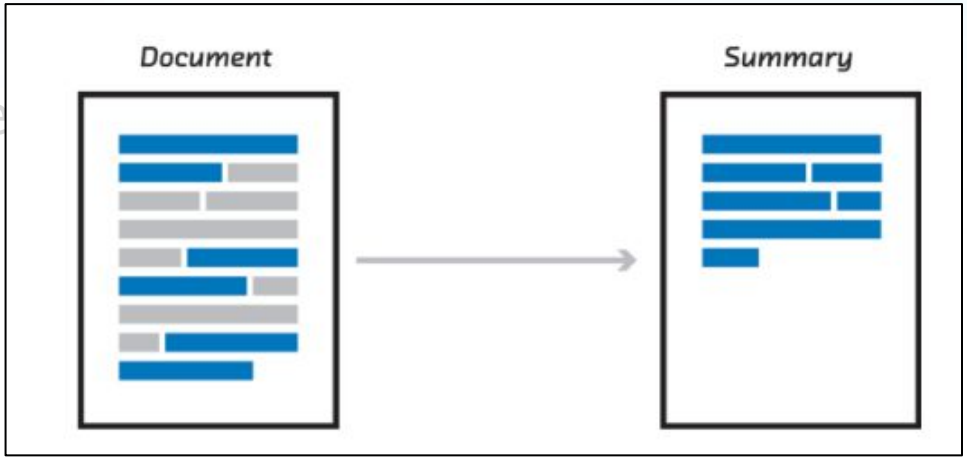
Machine Translation

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

GENERATION

Question Answering

...

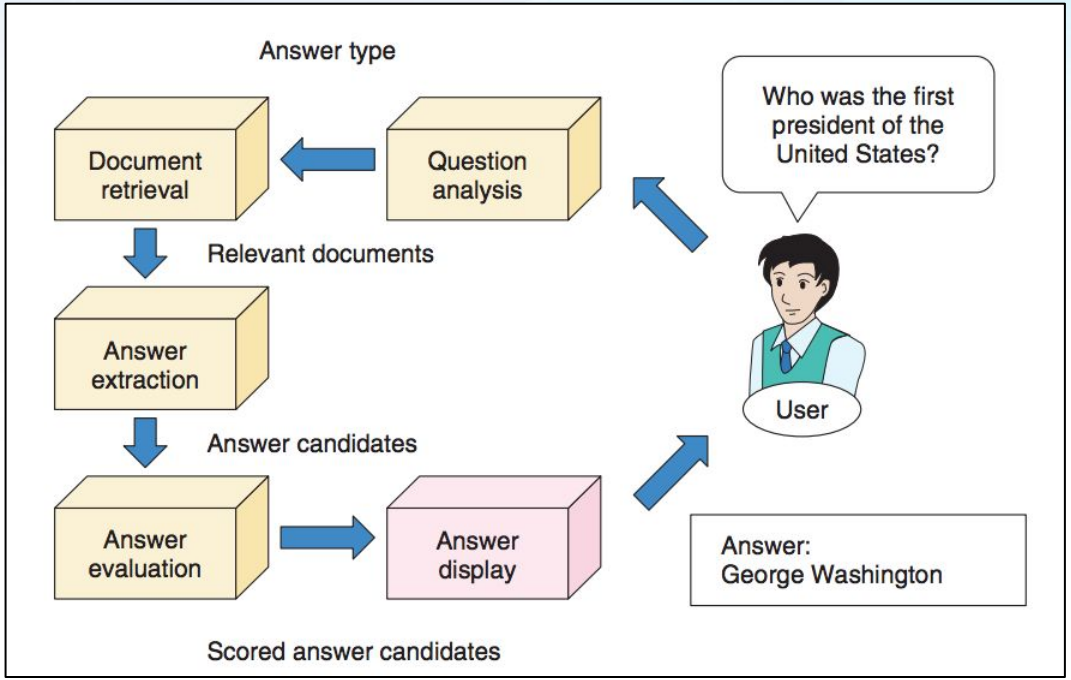


Image from www.ntt-review.jp

Types of queries

- ***Factoid:*** Who discovered America?
- ***Yes/No:*** Is Berlin the capital of Germany?
- ***Definition:*** What is leukemia?
- ***Cause/consequence:*** Why did the Iraq war start?
- ***Procedural:*** Which are the steps for getting a Master degree?
- ***Comparative:*** What is the difference between model A and model B?
- ***Queries with examples:*** What hard disks are similar to hard disk X?
- ***Queries about opinion:*** What is the opinion of the majority of Americans about the Iraq war?

Types of NLP Applications

Generation

Question Answering

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

Question Answering

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

Question Answering

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 (2013)

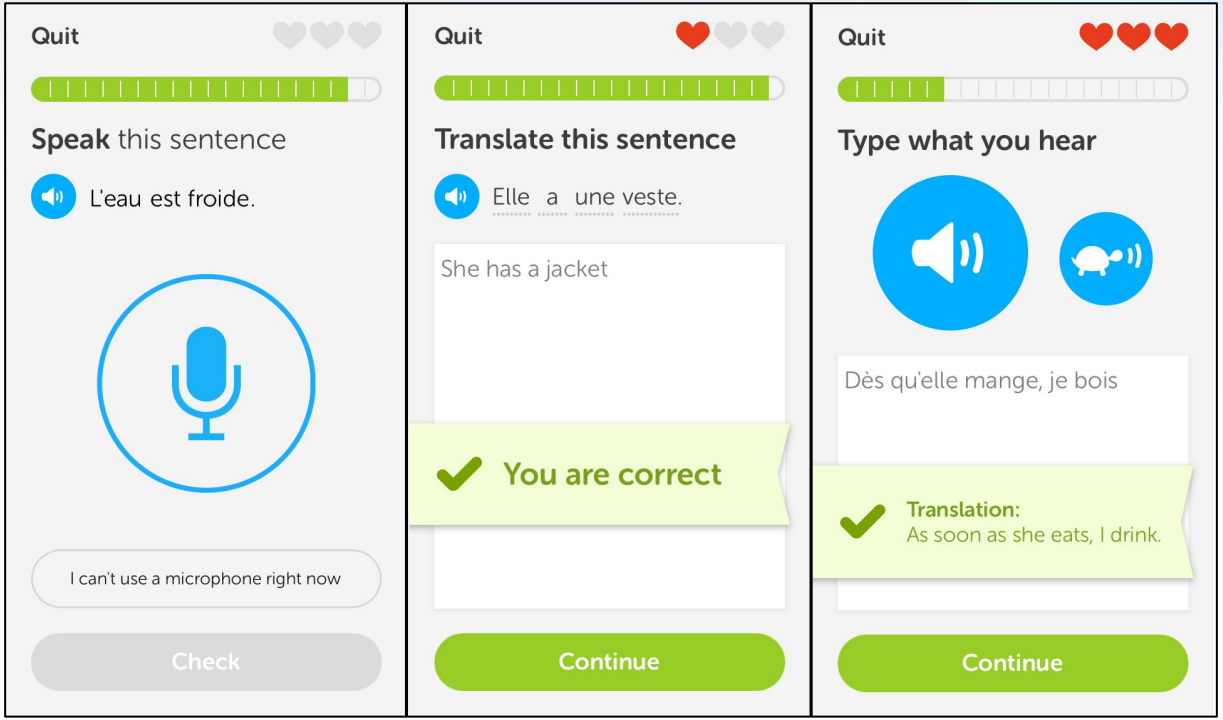
Computer-Generated Text

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."

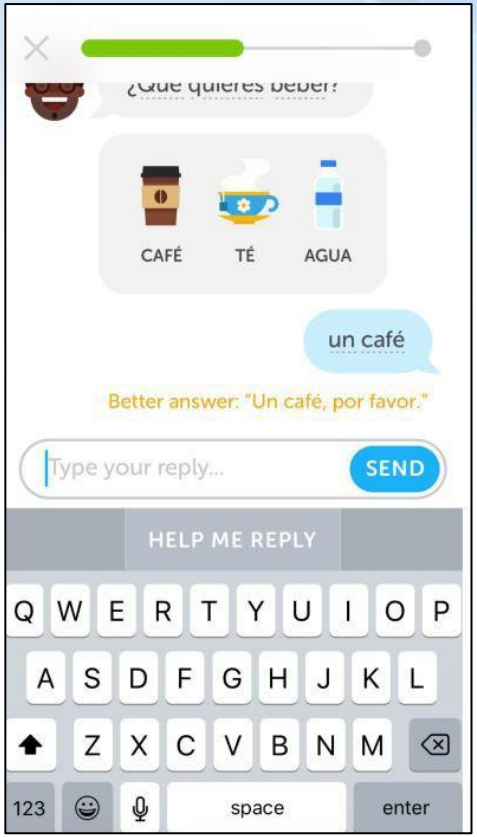
Types of NLP Applications

Language Learning



Duolingo

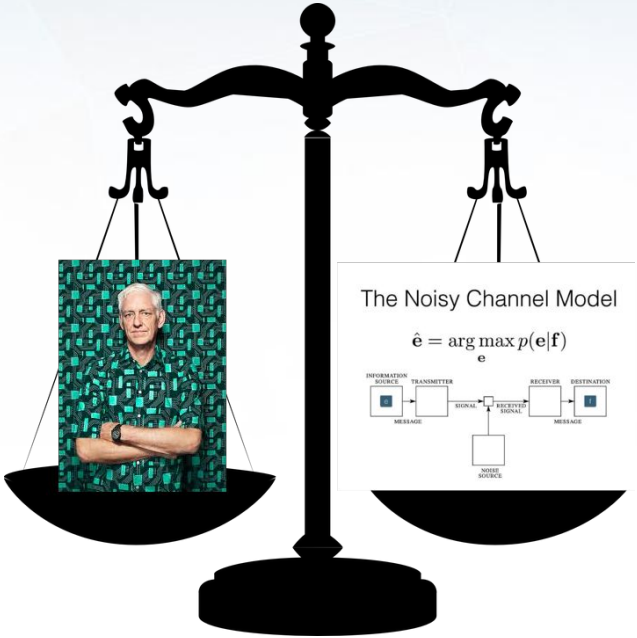

Hola
¿Cómo estás?
Bien tú?
¡Qué bien! Yo estoy bien.
Conoce a mis amigos. Primero, ¡dile "hola" a Roberto!
Type in Spanish
SEND
Somebody else is typing...
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123
space
return


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

Real-life projects by Vsevolod Dyomkin

Error correction at Sciworth

Spellchecker



Error-correction framework

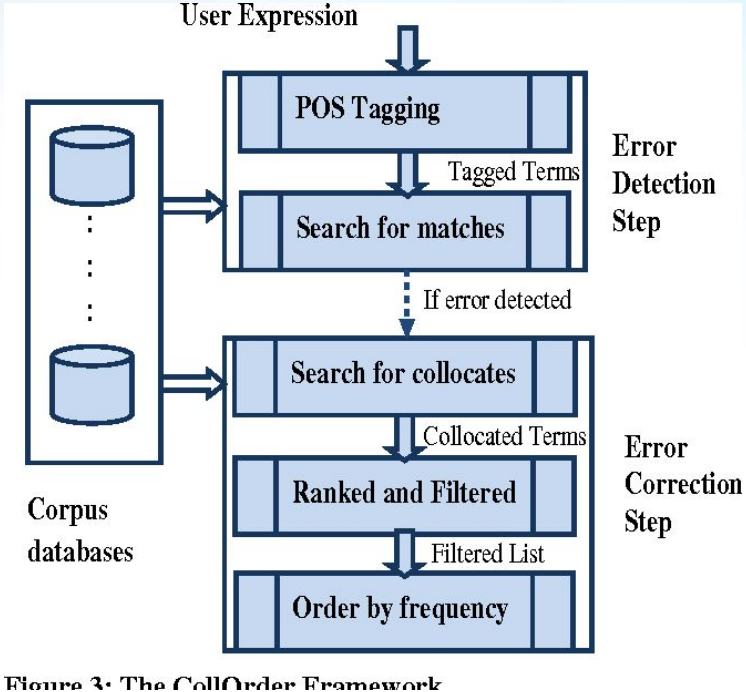


Figure 3: The CollOrder Framework

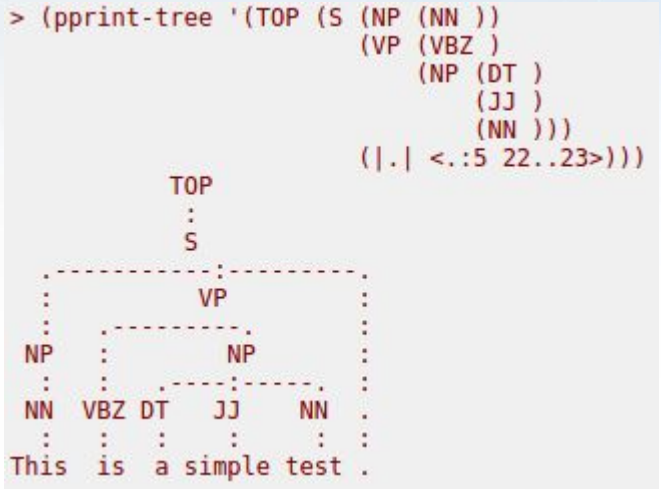
Building Classifiers

- Language identification
- Email dissection
- Product catalog with 1k categories
- Identifying subjective statements

Making Open-source Libraries

- CL-NLP

```
> (pprint-tree '(TOP (S (NP (NN ))
                        (VP (VBZ )
                            (NP (DT )
                                (JJ )
                                (NN )))
                        (|.| <.:5 22..23>)))
```

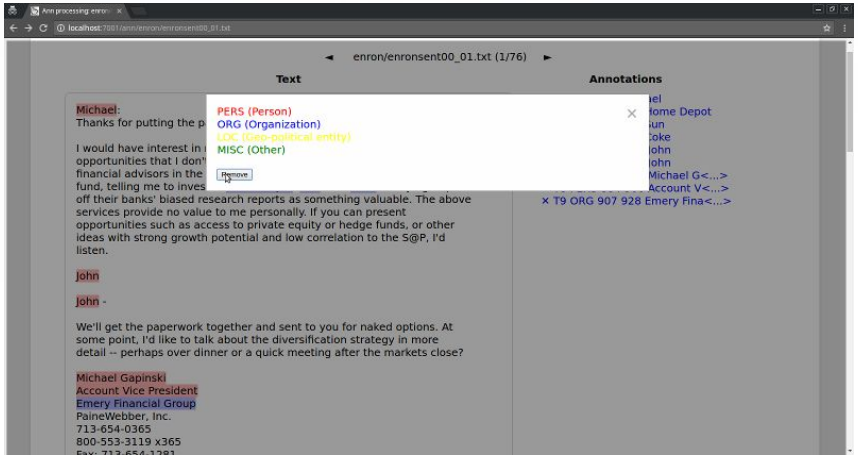


- WILD

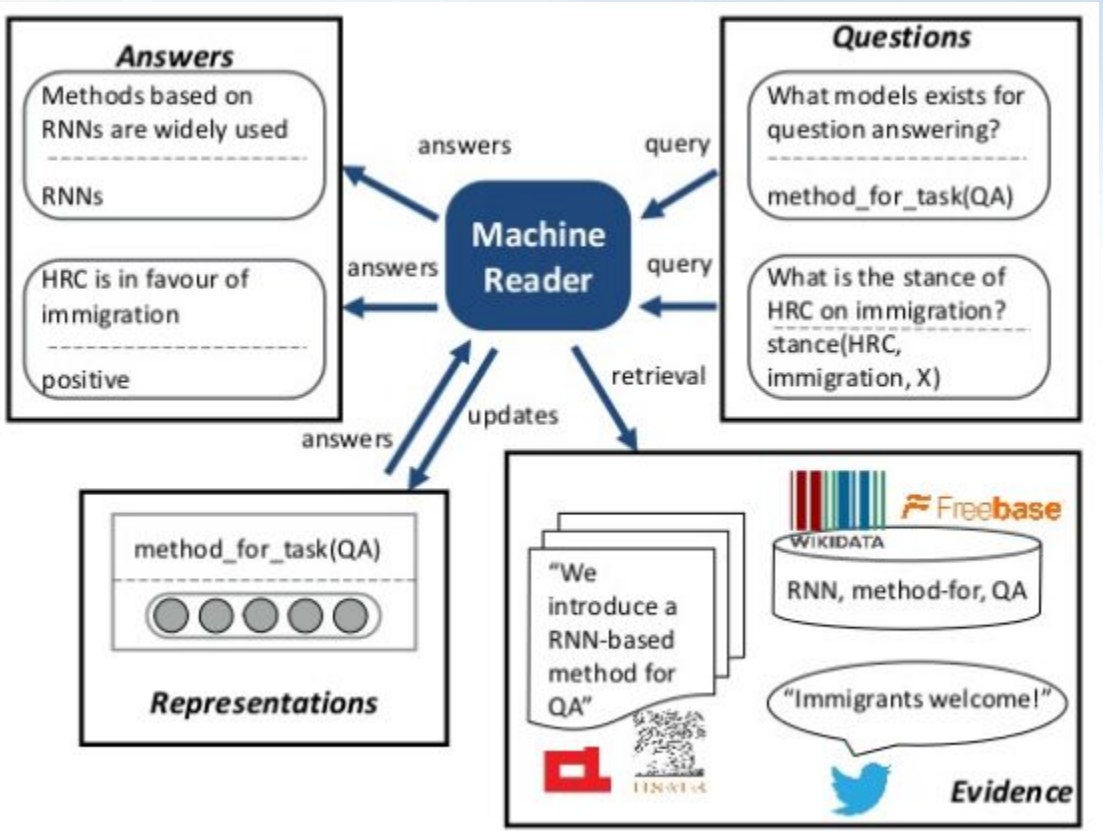
https://www.slideshare.net/vseloved/nlp-in-the-wild-or-building-a-system-for-text-language-identification

Building Corpora

- Grammarly
- lang.org.ua
- fact-checking



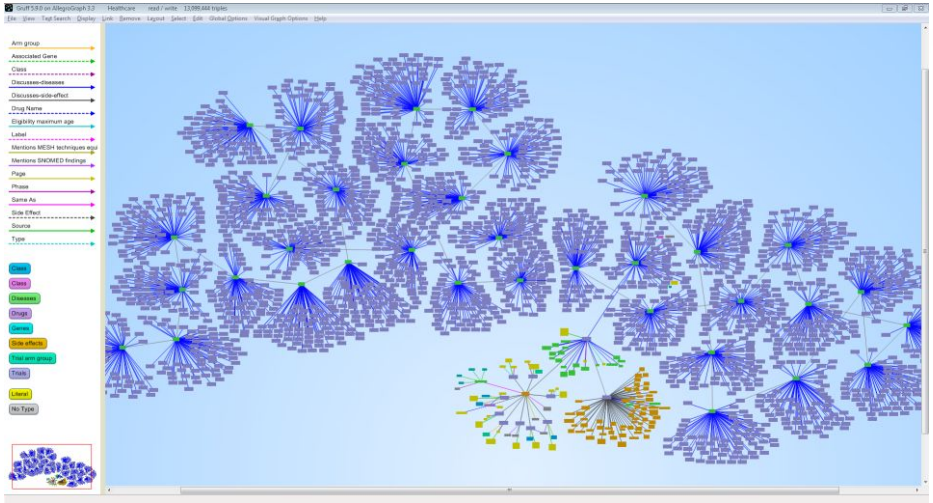
Automated Fact-checking



Picture © https://www.slideshare.net/isabelleaugenstein/learning-to-read-for-automated-fact-checking

NLP in agraph

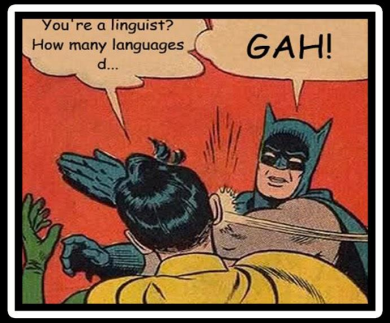
- Building an ML/NLP pipeline inside the graph DB
- Entity extraction
- Various classification projects (sales chats, company industries, customer types, tweets, etc.)



AllegroGraph
Franz Inc.

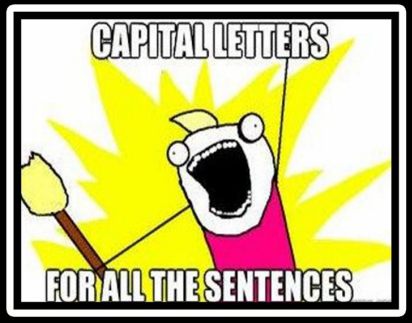
Real-life projects by Mariana Romanyshyn

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

What we do

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

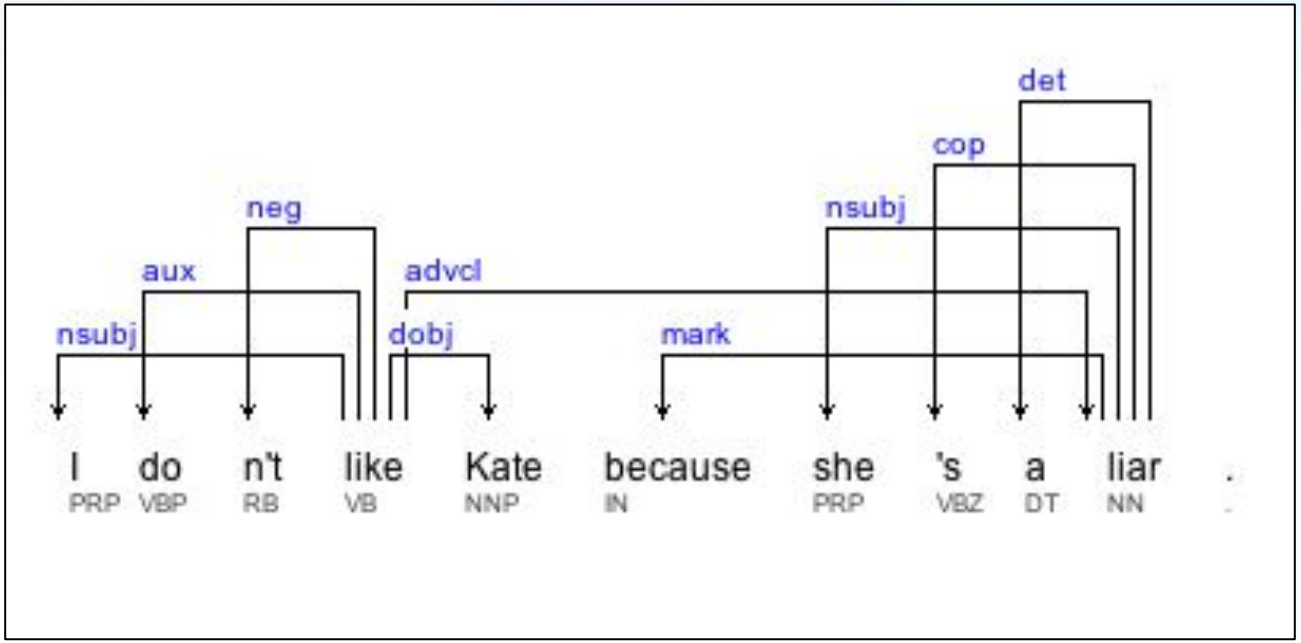
For example

You see:

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

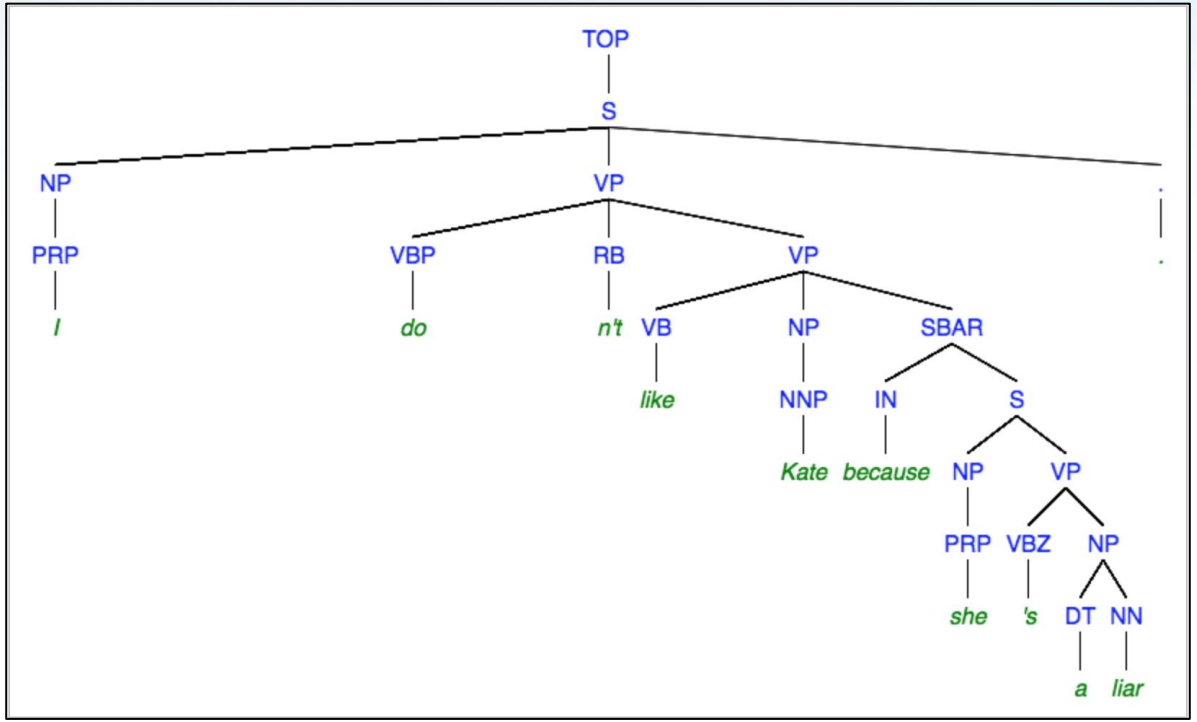
For example

Linguists see:



For example

Linguists see:



For example

Linguists see:

```
(TOP
 (S (NP (PRP I))
    (VP (VBP do) (RB n't)
        (VP (VB like) (NP (NNP Kate))
            (SBAR (IN because)
                  (S (NP (PRP she))
                     (VP (VBZ 's)
                         (NP (DT a) (NN liar)))))))
    (|.| .)))
```

For example

Linguists see:

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

For example

Linguists see:

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

For example

Linguists see:

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

I - agent

Kate - patient

And many-many more features

- number of sentences, words, words per sentence, etc.
- size and arrangement of paragraphs
- word length
- word position in a sentence
- number of syllables in a word
- ratio of vowels vs consonants
- depth of the word in the dependency tree of the sentence
- number of word senses
- ngrams
- morphemes: stems and affixes
- is the word capitalized/hyphenated/compound?
- grammatical categories of different POS...

1. Text Grading

Task: match a text with a CEFR level.

- Vocabulary
  - Extract notional parts of speech
  - Classify by weighted frequency
- Grammar
  - Extract grammatical structures
  - Map them to CEFR levels

brainglass

1. Text Grading

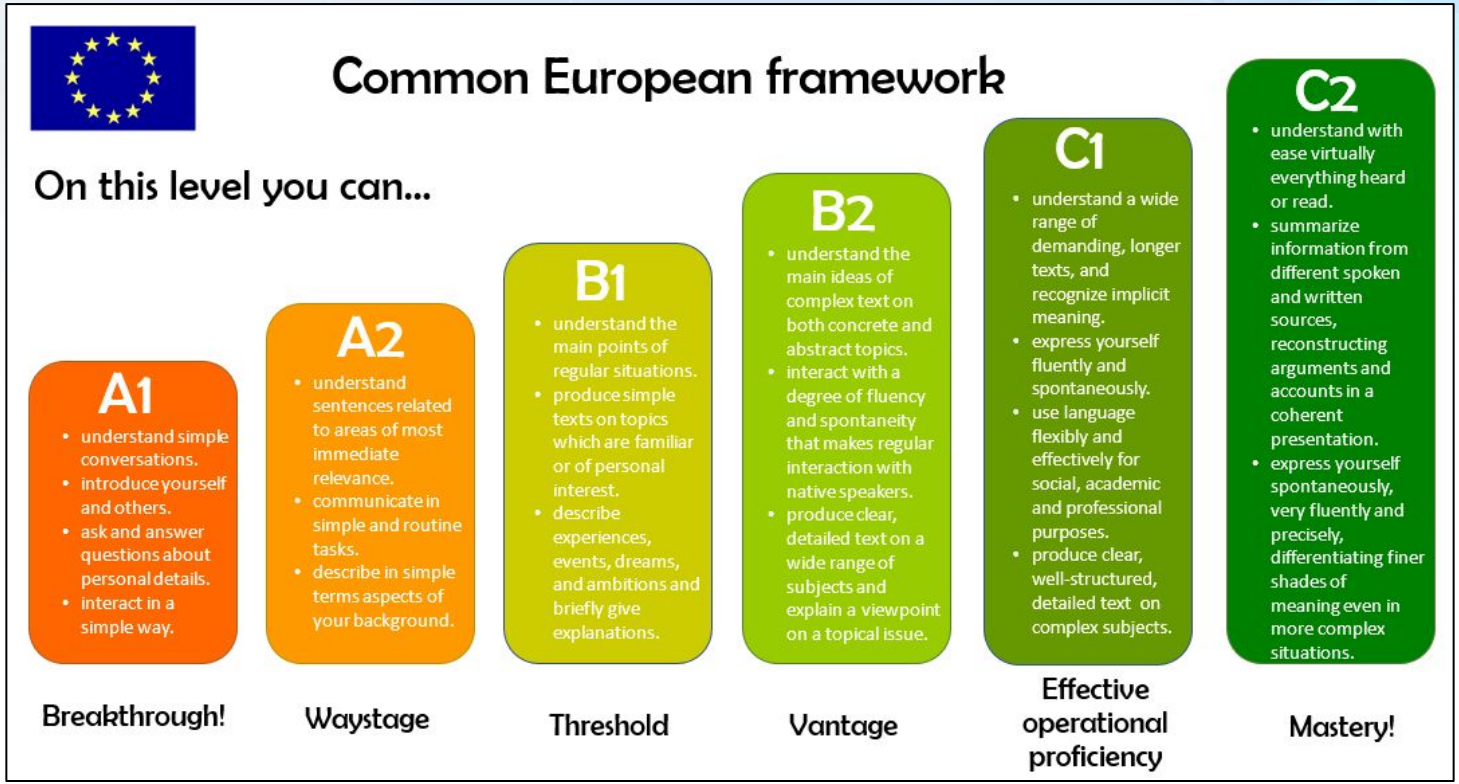


Image from http://mayraspanishschool.com/

2. Text Mining

Task: extract data about the company from its web page.

- Company name
- Phone/Fax
- Email
- Address
- Foundation date
- Working hours
- Partners and investors, etc.

2. Text Mining

Image from www.restaurantlechristine.com/

2. Text Mining

RESTAURANT LE CHRISTINE

Le Christine est membre des Maîtres Restaurateurs de France, certifiant la fraîcheur et la qualité de tous les produits de notre carte et une réalisation entièrement faite maison.

ARTICLES RÉCENTS

Le 14 février, l'amour s'installe au CHRISTINE

Nouveau site Internet

Bonne Année !!

OUVERTURE

Tous les soirs (7 jours sur 7) à partir de 18h30 & le midi, du lundi au vendredi de 12h à 14h30. Le restaurant est aussi ouvert les jours fériés

CONTACTEZ-NOUS !

1 rue Christine, 75006 Paris

+33 1 40 51 71 64

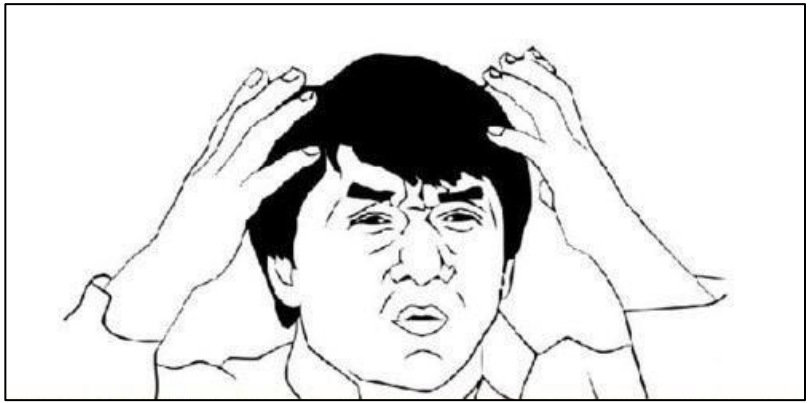
RESERVER

Image from www.restaurantlechristine.com/

2. Text Mining

OUVERTURE

Tous les soirs (7 jours sur 7) à partir de 18h30 & le midi, du lundi au vendredi de 12h à 14h30. Le restaurant est aussi ouvert les jours fériés



3. Sentiment Analysis

Task: extract objects and sentiment they acquired.

- POS tagger
- Named-entity recognition
- Syntactic parser
- Coreference resolution
- Object-oriented sentiment analysis

3. Sentiment Analysis

TomTom dealt a major blow to TimTim.

3. Sentiment Analysis

TomTom dealt a major blow to TimTim.

TomTom dealt TimTim a major blow.

TimTim was dealt a major blow by TomTom.

A major blow was dealt to TimTim by TomTom.

4. Fact Extraction

Task: extract events about operational risks in financial companies

- Semantic roles
  - Agent
  - Patient
  - Sum of money
  - Date of settlement
  - Location
  - Reason for the loss
- Classification of operational risks
- Sequence of events
- Merge duplicate articles

4. Fact Extraction

Bloomberg

Cantor Fitzgerald Sued by Partners Who Moved to Reorient

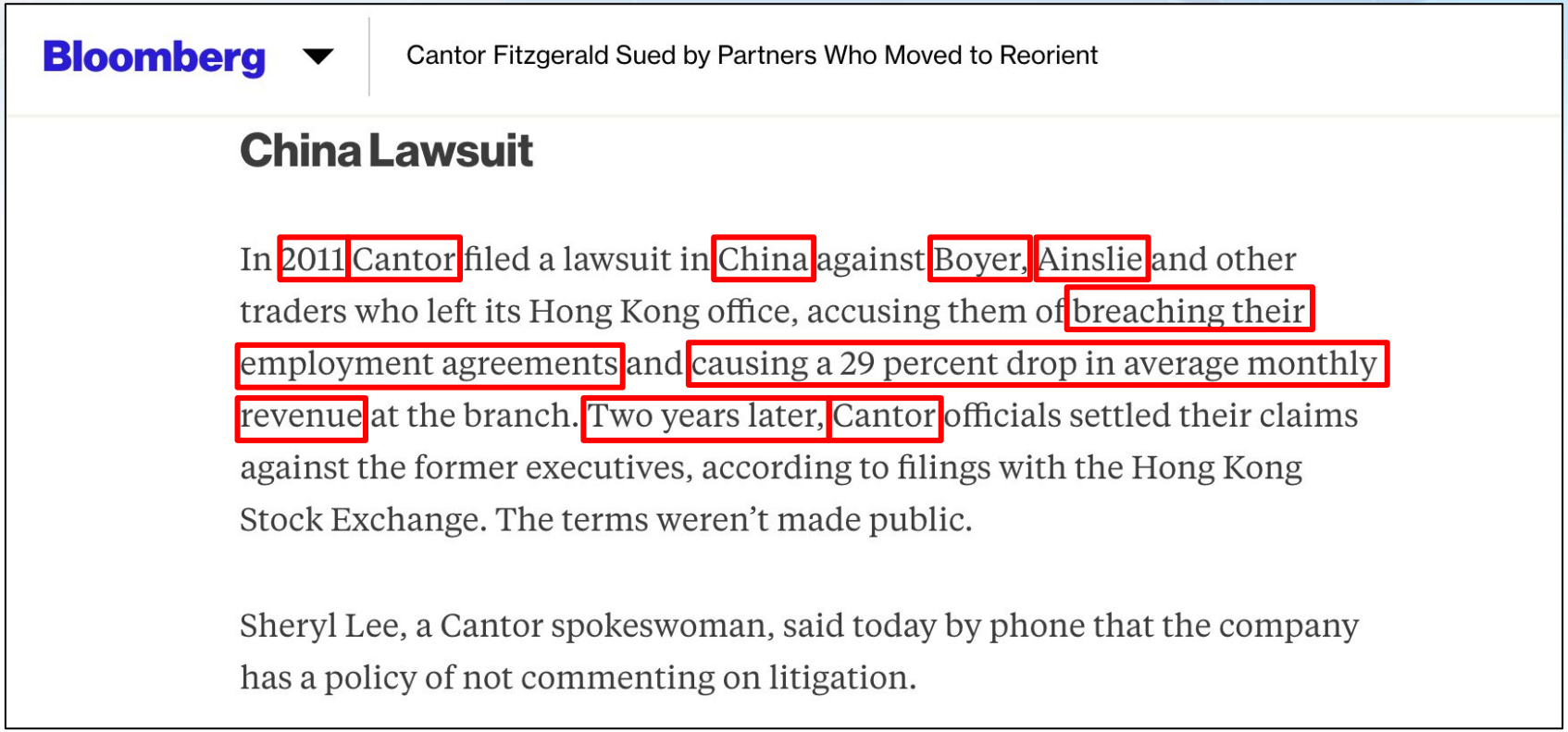
China Lawsuit

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/

4. 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/

5. Error correction

Motivation:

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

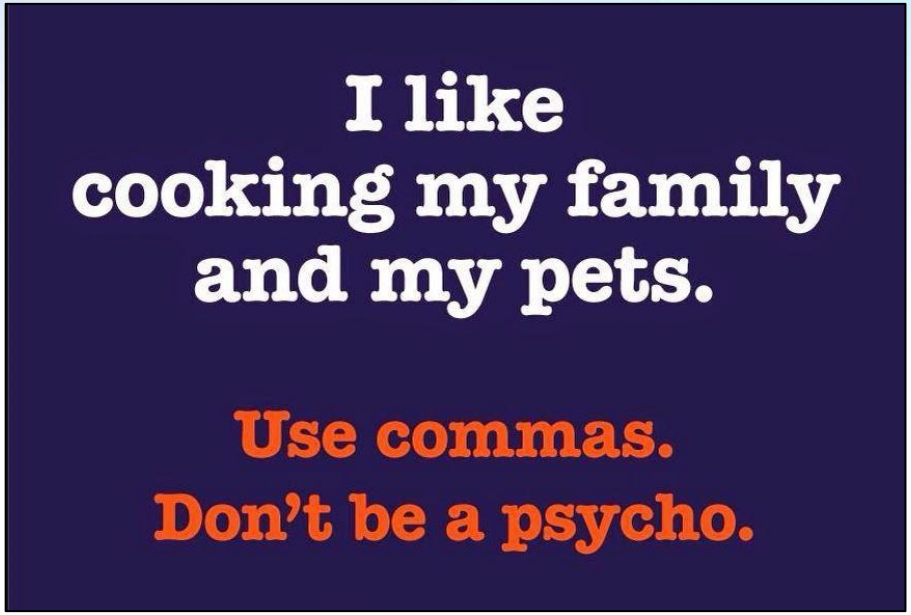


Image from www.writing.com/

5. Error correction

Task: make user texts ***correct*** and ***clear***.

Spelling, Grammar, Punctuation

- *I cutted your fnger didn't I?*
- *I cut your finger, didn't I?*

- *In daytime, he stayed in room.*
- *In the daytime, he stayed in the room.*

5. Error correction

Task: make user texts ***correct*** and ***clear***.

Style

- *There is a scarcity of food in the city.*
- *Food is scarce in the city.*

- *Could you go check what i've written???*
- *Could you check what I have written?*

5. Error correction

Task: make user texts ***correct*** and ***clear***.

Paraphrasing

- *Joey came racing at a very fast speed.*
- *Joey came racing at a breakneck speed.*

- *A very fast train runs through the city of Urumqi.*
- *A high-speed train runs through the city of Urumqi.*

5. Error correction

A documents can be found on my table.
Jim is a graduate students at the University of Kansas.
What are a dimensions in general?
If you want to learn about an arts, visit the national gallery.

~~A documents~~ → Documents
A document

The indefinite article *A* may not be required with the plural noun *documents* in this sentence. Consider removing the article, or changing the noun to singular.

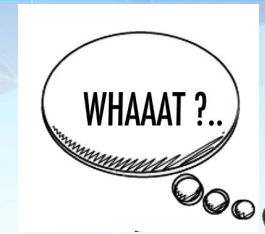
MORE IGNORE

~~students~~ → student

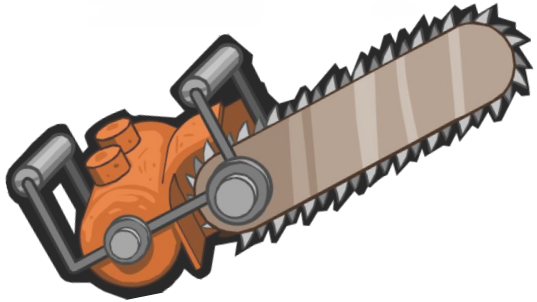
~~a dimensions~~ → dimensions

~~an arts~~ → art

5. Error correction



She sawed a black cat in the room.



6. Data anonymization

Original:

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

6. Data anonymization

Original:

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

Anonymized:

> Boris and Althea Stephanopoulos bought a car at Acme Industries for €120K on March 21st, 2001.

Etc

- Syntactic parsing
- Annotation tool
- Data annotation and data licensing
- Text analysis for Marketing (e.g., analysis of presidential debates)
- Text insights (e.g., readability, vocabulary variety)
- Spell checker for mobile language
- Grammarly cards
- Thesauri
- ...

7. brown-uk

- dict-uk - electronic dictionary of Ukrainian
  - 3,5 mln word forms
- nlp-uk - NLP API of LanguageTool for Ukrainian
  - segmentation
  - POS tagging
  - spelling, grammar, punctuation
- corpus - Brown corpus of Ukrainian
  - balanced by genres
  - balanced by regions

Questions?

Interesting References

- Peter Eckersley and Yomna Nasser, *Measuring the Progress of AI Research* (ongoing)
- Peter Norvig, *How to Write a Spelling Corrector* (2007)
- Mishra A. et al., *Harnessing Cognitive Features for Sarcasm Detection* (2016)
- Papantoniou K. and Konstantopoulos S., *Unravelling Names of Fictional Characters* (2016)
- Kyunghyun Cho, *Introduction to Neural Machine Translation with GPUs* (2015)
- Vaswani A. et al., *Attention is all you need* (2017)
- Deepmind, *WaveNet: A Generative Model for Raw Audio* (2016)
- Mostafazadeh N. et al., *A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories* (2016)
- Nick Montfort, *World Clock* (2013)

Interesting References

- Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English (2018)
- Better Language Models and Their Implications (2019)