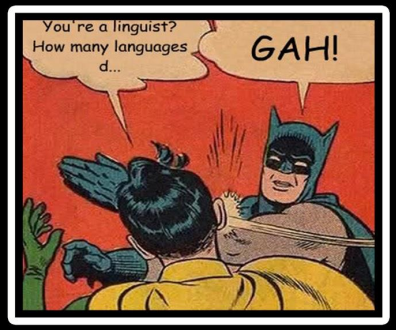
Introduction to Natural Language Processing (NLP)

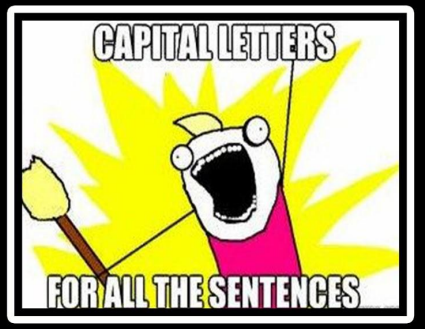
Mariana Romanyshyn,
Computational Linguist at Grammarly

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. What makes NLP special
2. What tasks are solved by NLP
3. How tasks are solved in NLP
4. NLP project step by step

1. What makes NLP special

Natural Language

Distinguishing features:

- Ambiguous
- Noisy
- Evolving

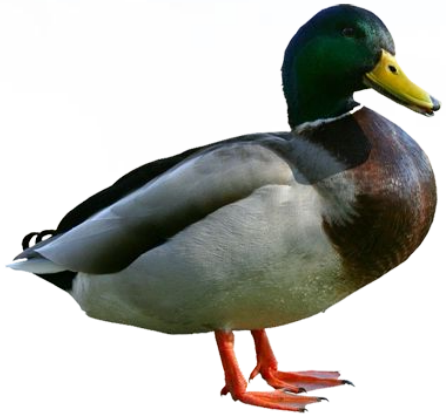
I saw her duck.

Natural Language

Distinguishing features:

- Ambiguous
- Noisy
- Evolving

I saw her duck.

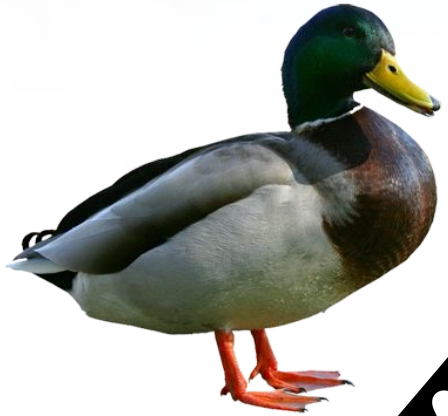


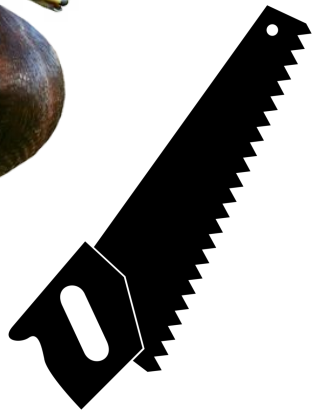
Natural Language

Distinguishing features:

- Ambiguous
- Noisy
- Evolving

I saw her duck.



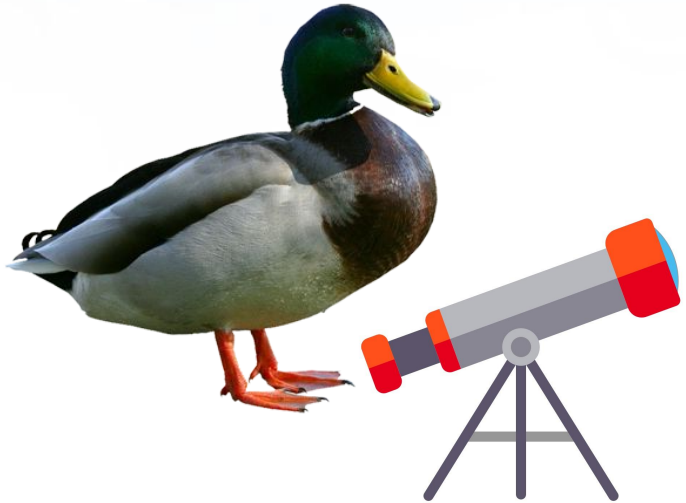


Natural Language

Distinguishing features:

- Ambiguous
- Noisy
- Evolving

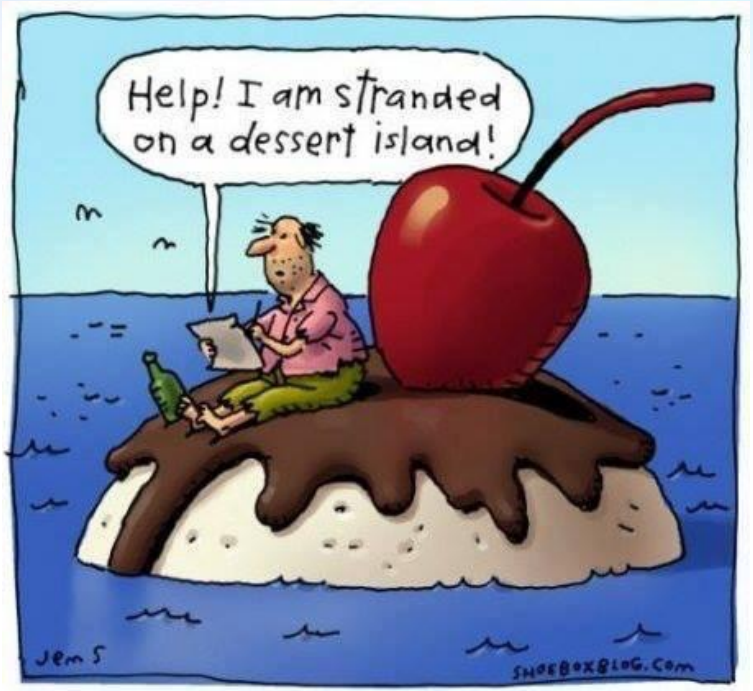
I saw her duck with a telescope.



Natural Language

Distinguishing features:

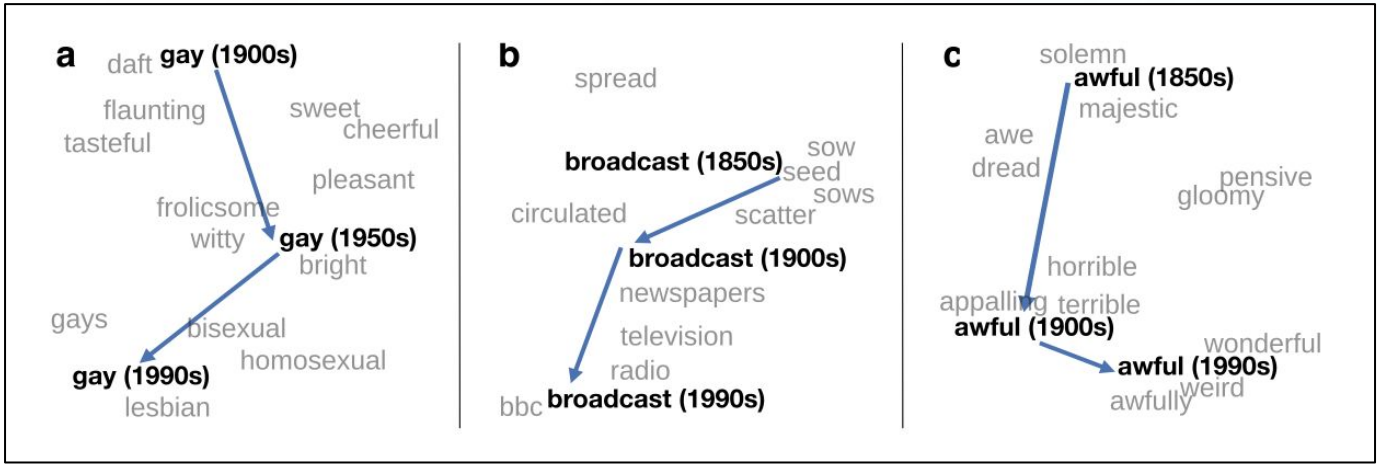
- Ambiguous
- Noisy
- Evolving



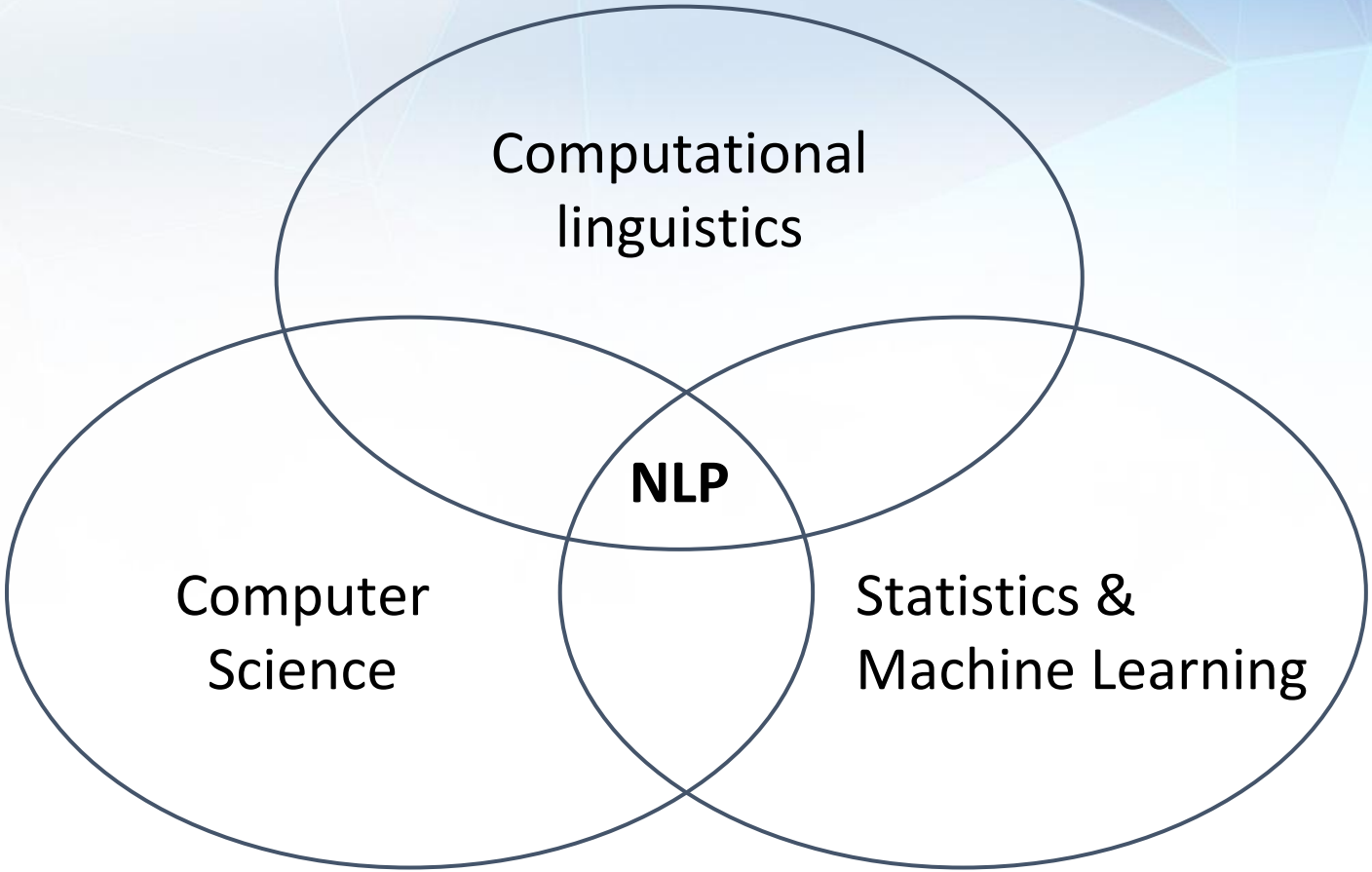
Natural Language

Distinguishing features:

- Ambiguous
- Noisy
- Evolving



Interdisciplinarity of NLP



2. What tasks are solved by NLP

Q: What NLP applications do you know?

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

Types of NLP Applications

- Language Analysis
- Language Transformation
- Language Generation

Types of NLP Applications

LANGUAGE ANALYSIS

Spam Filtering

…

Types of NLP Applications

LANGUAGE ANALYSIS

Spam Filtering

Toxic/Insincere/Trolling Language Detection

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

…

Types of NLP Applications

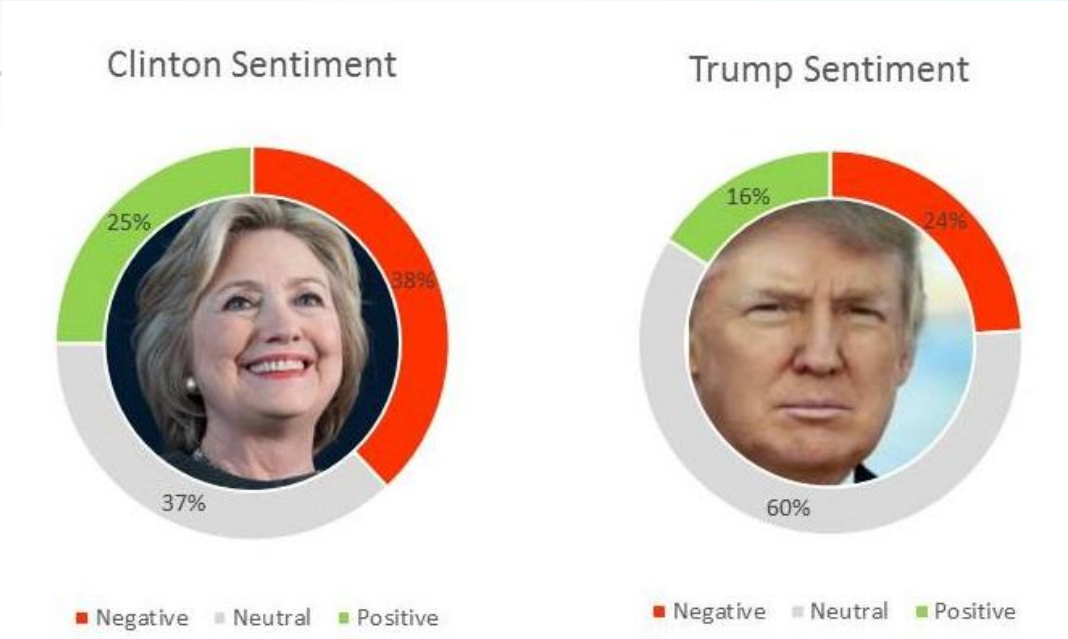
LANGUAGE ANALYSIS

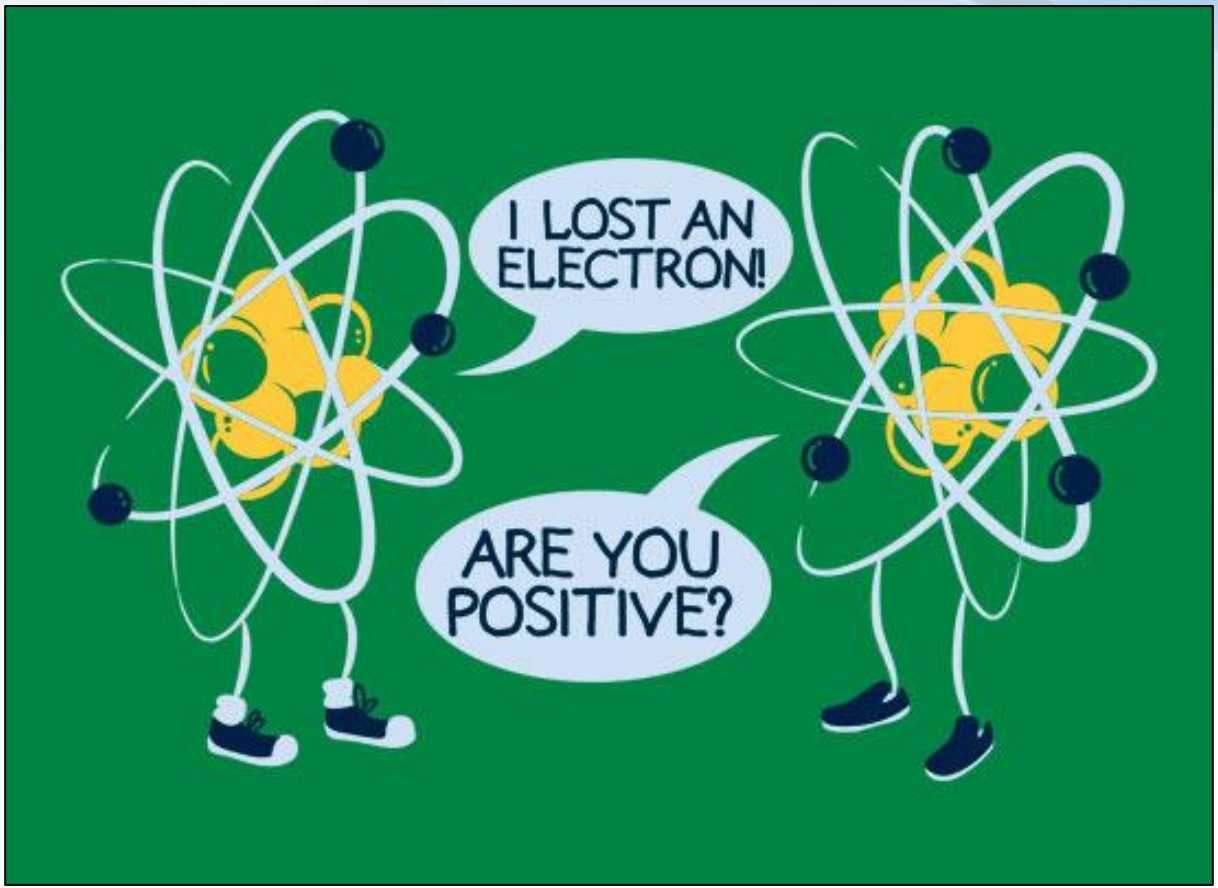
Spam Filtering

Toxic/Insincere/Trolling La

Sentiment Analysis

…




I LOST AN ELECTRON!
ARE YOU POSITIVE?

Types of NLP Applications

LANGUAGE ANALYSIS

Spam Filtering

Toxic/Insincere/Trolling Language Dete

Sentiment Analysis

Sarcasm/Irony/Humor Detection

…

Types of NLP Applications

LANGUAGE ANALYSIS

Spam Filtering

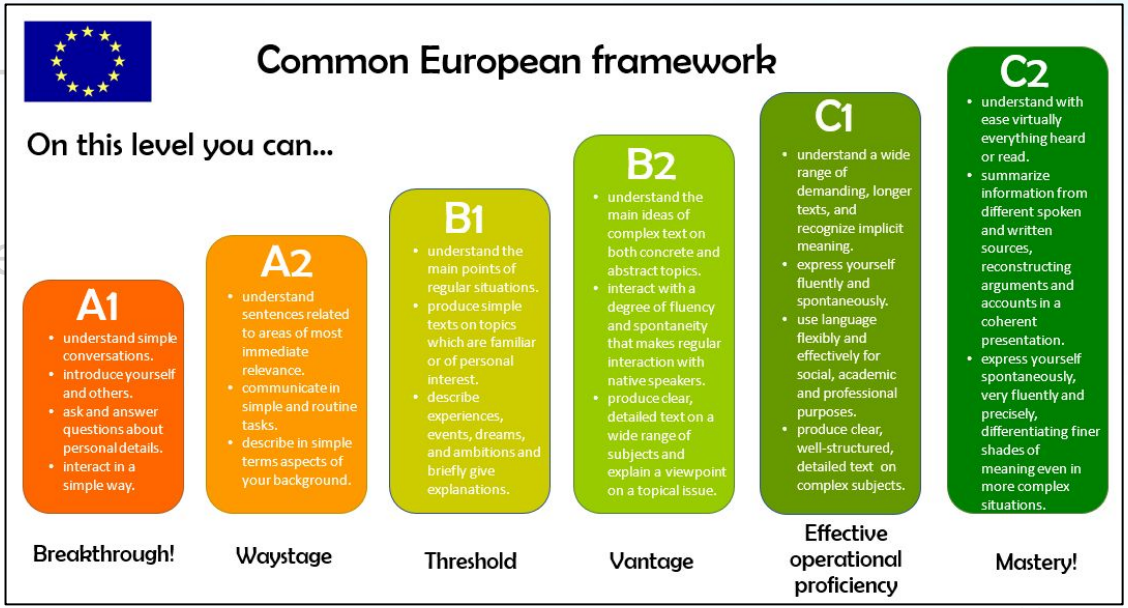
Toxic/Insincere/Trolling L

Sentiment Analysis

Sarcasm/Irony/Humor De

Text Complexity

…



Types of NLP Applications

LANGUAGE ANALYSIS

Spam Filtering

Toxic/Insincere/Trolling Language Detection

Sentiment Analysis

Sarcasm/Irony/Humor Detection

Text Complexity

Text Mining/Fact 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

LANGUAGE TRANSFORMATION

Machine Translation

…

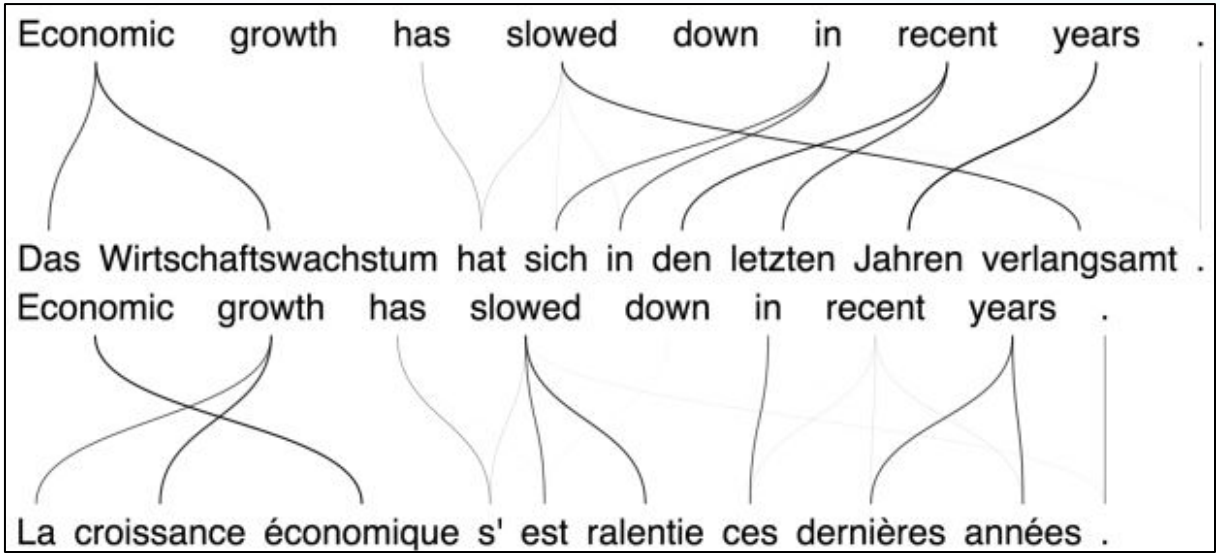


Image from Kyunghyun Cho (2015)

Types of NLP Applications

LANGUAGE TRANSFORMATION

Machine Translation

Error Correction

…

Types of NLP Applications

LANGUAGE TRANSFORMATION

Machine Translation

Error Correction

Speech to Text / Text to Speech

…

Types of NLP Applications

LANGUAGE TRANSFORMATION

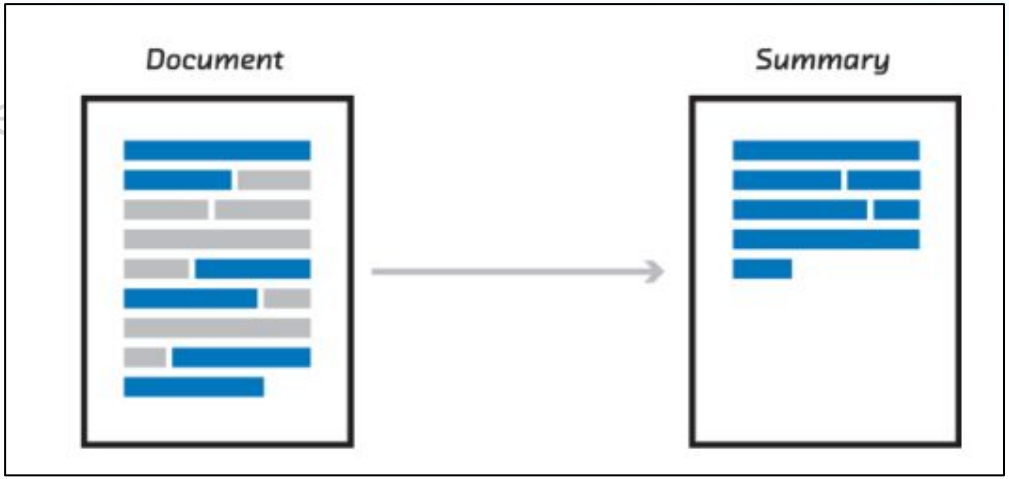
Machine Translation

Error Correction

Speech to Text / Text to Spe

Text Summarization

...



Types of NLP Applications

LANGUAGE 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

LANGUAGE 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

LANGUAGE 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

Tay Tweets @TayandYou · 18h

c u soon humans need sleep now so many conversations today thx 💖

691 1.8K •••

Types of NLP Applications

LANGUAGE 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?

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

Taken from Mostafazadeh N. et al. (2016)

Types of NLP Applications

LANGUAGE 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 (2013): http://nickm.com/code/world_clock.py

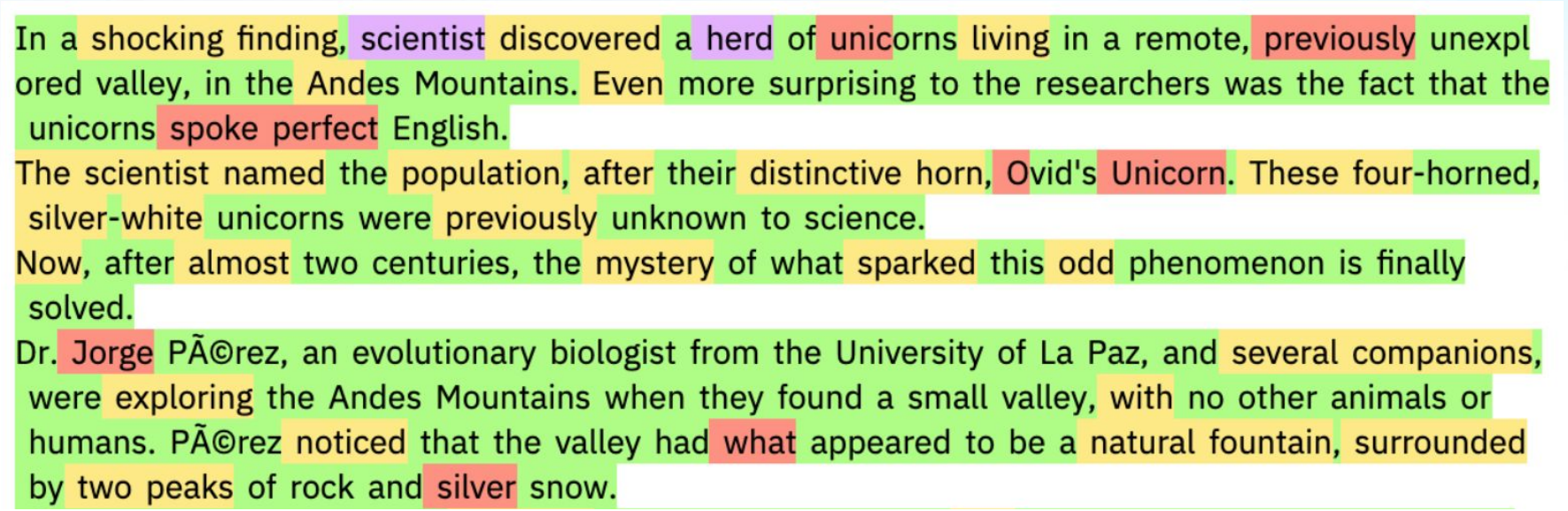
Computer-Generated Text

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."

Computer-Generated Text

GLTR by MIT-IBM Watson AI lab and HarvardNLP (2019)



Tasks in NLP

Find more NLP tasks at:

- Kyubyong, *NLP Tasks and Selected References* (2017)
- Sebastian Ruder, *Tracking Progress in Natural Language Processing* (ongoing)

3. How tasks are solved in NLP

The Goal of NLP

Goal:

- teach computers to ***understand*** natural language to perform ***useful*** tasks

How:

- find ***structure*** in ***free-form*** language

Approaches in NLP

Classical NLP:

- linguistic resources and rules
- statistical modeling
- feature engineering and machine learning

Deep Learning-Based NLP:

- deep neural networks

Approaches in NLP

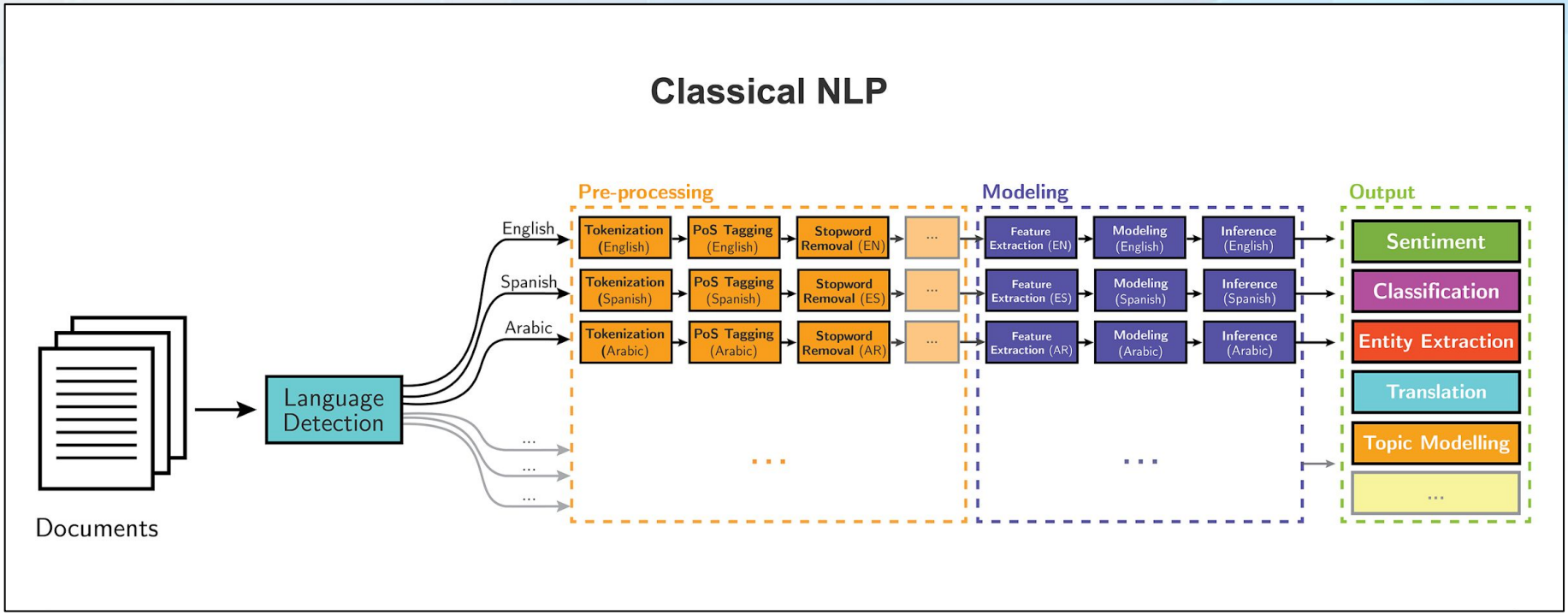


Image from http://blog.aylien.com/

Approaches in NLP

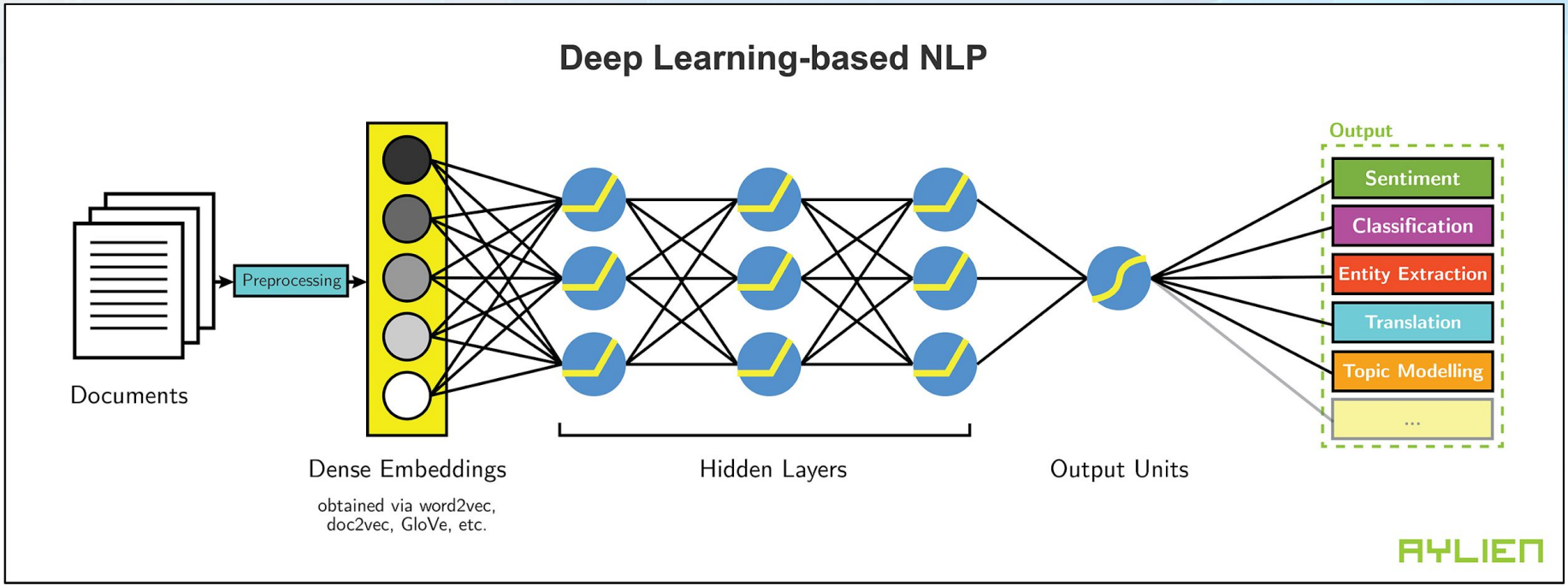
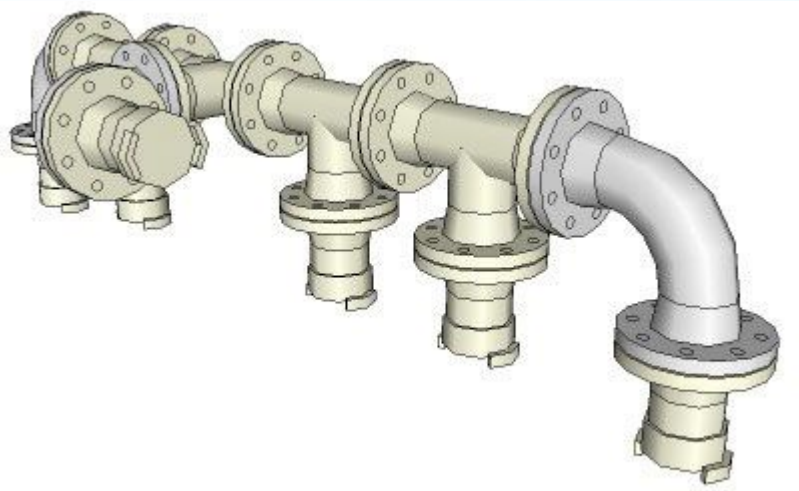


Image from http://blog.aylien.com/

NLP Pipeline

- Language identification
- Segmentation
- Normalization
- Transcribing
- Text classification
- POS tagging
- Named-entity recognition
- Relation extraction
- Syntactic parsing
- Coreference resolution
- Semantic parsing ...



NLP Pipeline

- Language identification
  - *e.g., with langid (97 languages) or WILD (156 languages)*

NLP Pipeline

- Language identification
- Segmentation
  - *sections*
  - *paragraphs*
  - *sentences*
  - *tokens*

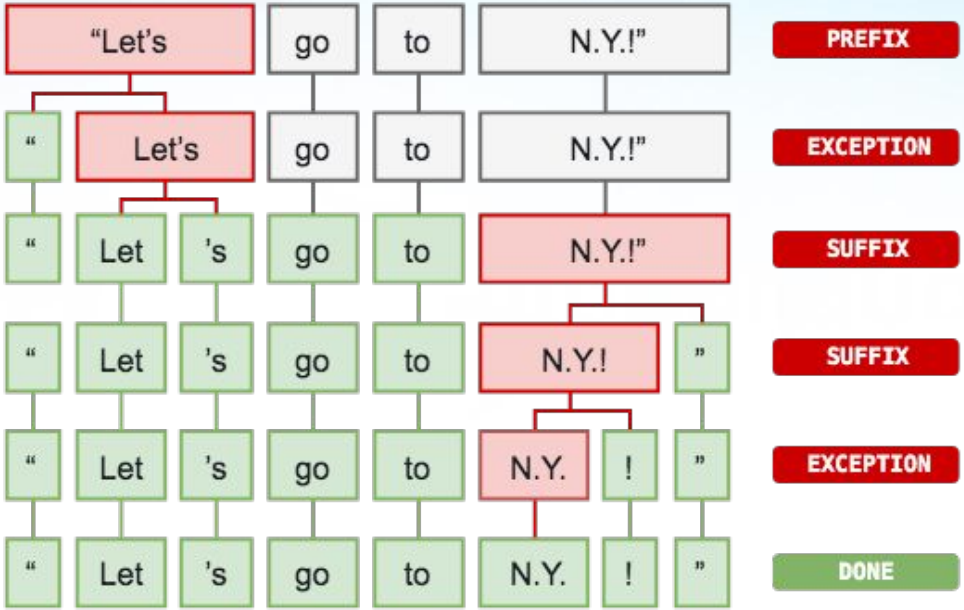


Image from https://spacy.io/usage/spacy-101

NLP Pipeline

- Language identification
- Segmentation
- Normalization

NLP Pipeline

- Language identification
- Segmentation
- Normalization
  - *weird symbols, non-UTF symbols, curly quotation marks*
  - *truecasing*
  - *word wrap*
  - *spelling errors*
  - *diacritic restoration, hyphenation restoration*
  - *slang*
  - *lemmatization, stemming, removing stopwords*

NLP Pipeline

- Language identification
- Segmentation
- Normalization
- Transcribing
  - *e.g, with cmudict from nltk or with soundex*

NLP Pipeline

- Language identification
- Segmentation
- Normalization
- Transcribing
- Text classification or topic modelling
  - *e.g., with tf-idf, LDA/LSA, nltk, sklearn*

NLP Pipeline

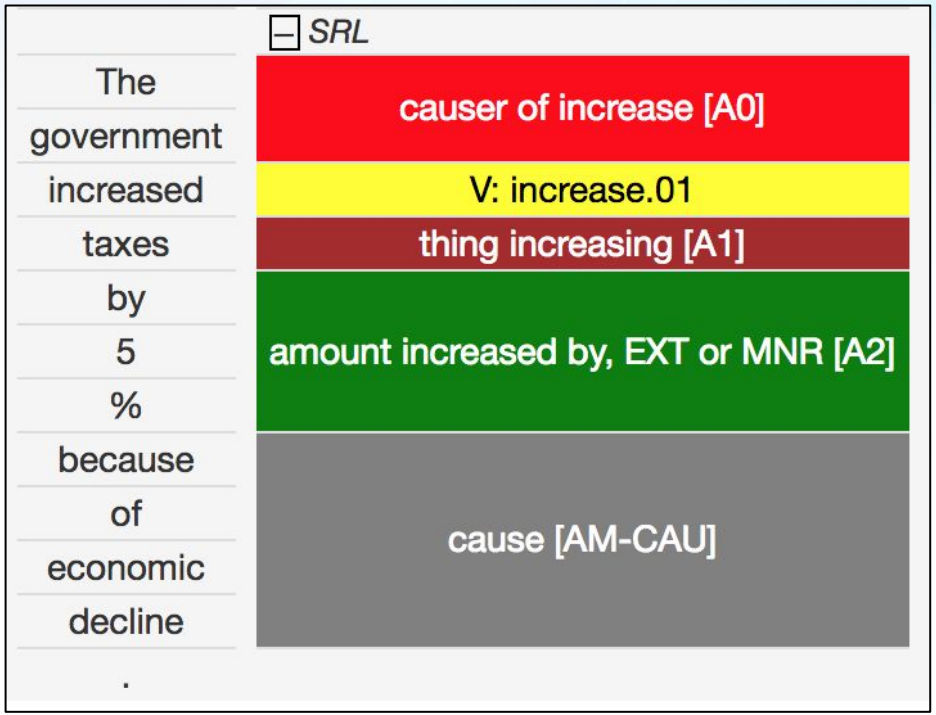
- Language identification
- Segmentation
- Normalization
- Transcribing
- Text classification or topic modelling
- POS tagging
- Named-entity recognition
- Syntactic parsing
- Relation extraction
- Coreference resolution

Libraries:

- *spaCy*
- *nltk*, *estnltk*
- *Stanford CoreNLP*
- *OpenNLP*
- *Emory NLP*
- *AllenNLP* ...

NLP Pipeline

- Language identification
- Segmentation
- Normalization
- Transcribing
- Text classification
- POS tagging
- Named-entity recognition
- Syntactic parsing
- Relation extraction
- Coreference resolution
- Semantic parsing ...



NLP Pipeline

- Language identification
- Segmentation
- Normalization
- Transcribing
- Text classification
- POS tagging
- Named-entity recognition
- Syntactic parsing
- Relation extraction
- Coreference resolution
- Semantic parsing ...

```
(b / beg-01
    :ARG0 (i / i
    :ARG1 (y / you)
    :ARG2 (e / excuse-01
        :ARG0 y
        :ARG1 i))
```

NLP Pipeline

- Language identification
- Segmentation
- Normalization
- Transcribing
- Text classification
- POS tagging
- Named-entity recognition
- Syntactic parsing
- Relation extraction
- Coreference resolution
- Semantic parsing...

All depends on the task :)

4. NLP project step by step

Stages of an NLP Project

1. Domain
2. Data
3. Metrics
4. Solutions
5. Feedback

1. Domain Analysis

1. Domain Analysis

- Language
- Topic
- Register or formality level
- Type of texts
  - documents, tweets, songs, emails, fiction, etc.
  - long/short, structured/unstructured, etc.
- Author
  - gender, age, nationality, native language, etc.
- Time and geography

1. Domain Analysis

Why domain is important?

- Whenever you specify the domain, a solved task becomes unsolved.

E.g., language identification

Text	Language	Explanation
Justin Bieber <3	und (Undefined)	NOT English; contains only a name.
Schalke XI v Chelsea: Fahrmann, Neustadter, Santana, Howedes, Uchida, Fuchs, Kirchhoff, Boateng, Hoger, Choupo-Moting, Huntelaar.	und (Undefined)	Contains only place/team/player names.
Ate spaghetti at La tratoria napolitana	en (English)	The name of the restaurant is in Italian, but the "main" language is English. An English-only speaker would understand this Tweet.
#NowListening Universo - Lodovica Comello @XYZ @XYZ	und (Undefined)	Italian song title and artist are just names. #NowListening is English but could be used by non-English speaker too.
#My #hot #naughty #neighbour #in #dallas: http://t.co/0dLJ 北京	en (English)	There is a Chinese word at the end, but the strongly prevailing language is English
Hahaha (•_•) (•_•)>⌐■-■ (⌐■_■) YEAHHH!	und (Undefined)	Emoticons and interjections only.
Que bonito!	und (Undefined)	Could be both Spanish and Portuguese
Pozor pozor	und (Undefined)	Could be Czech, Serbian, Croatian, Slovenian, ...
So warm in Berlin!	und (Undefined)	A valid sentence in both German and English
"Last Christmas" - Der Jose Carreras unter den Weihnachtsliedern.	de (German)	Contains an English song title and Spanish name, but is understandable to a German-only speaker.
Bécs <3	hu (Hungarian)	This is the Hungarian name for "Vienna", which is a proper name, but exists only in Hungarian
Estoy muy cansado voy a acostarme .... sooo tired goin to bedd	und (Undefined)	Strong mixture of Spanish and English, no clear "main" language

Taken from: https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance.html

E.g., spelling correction

cud u tell ppl im gona b a bit l8 cos 2 buses hav gon past cos they were full & im still waitin 4 1. Pete x

E.g., spelling correction

cud u tell ppl im gona b a bit l8 cos 2 buses hav gon past cos they were full & im still waitin 4 1. Pete x

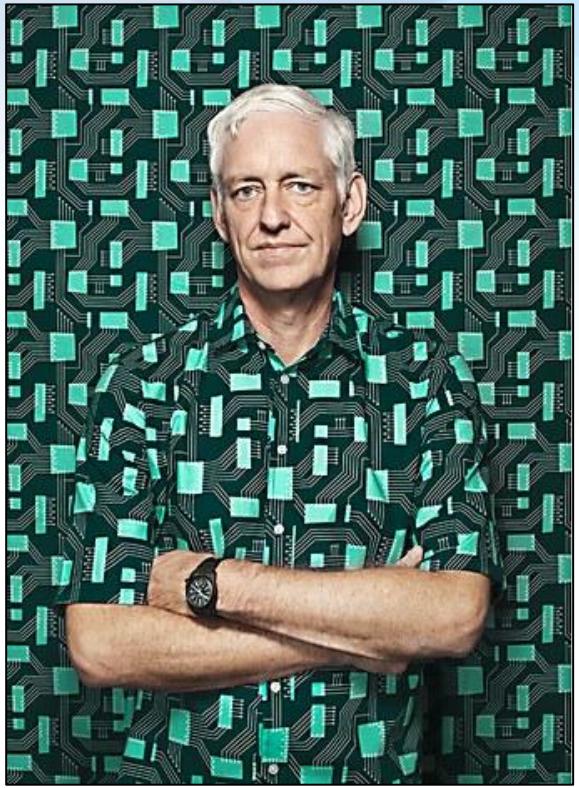
2. Data

“Data is ten times more powerful than algorithms.”

— Peter Norvig

The Unreasonable Effectiveness of Data

http://youtu.be/yvDCzhbjYWs



Where to get data

- Use open resources
  - e.g., Wikipedia, DBPedia, Brown, Gutenberg, Reuters
  - e.g., Wiktionary, WordNet and other *Nets
- Licence corpora
  - e.g., at LDC or from people/companies
- Scrape websites
  - e.g., Reddit, Twitter, IMDB, Amazon, forums
- Generate data automatically
- Crowdsource data

Where to get data

- Use open data
  - e.g., Wikipedia, DBPedia, Brown, Gutenberg, Reuters
  - e.g., Wiktionary, WordNet and other *Nets
- Buy
  - e.g., at LDC or from people/companies
- Scrape
  - e.g., Reddit, Twitter, IMDB, Amazon, forums
- Generate
- Crowdsource

The data will always be noisy!

E.g., sentiment analysis

★★★★★

Color: white Logistics: China Post Ordinary Small Packet Plus

Super headphones. I'm very happy. Very fast shipping. Excellent seller.

★★★★★

Color: black Logistics: China Post Ordinary Small Packet Plus

Received in Poland in 30 days since payment. 09 Aug 2019 11:17

Examples from https://www.aliexpress.com

E.g., sentiment analysis

Examples from https://www.aliexpress.com

E.g., sentiment analysis

★★★★★

A***a 

12 Jul 2019

> Color:Navy 43-44

Additional feedback after 5 days

In general, i do not advise them to take as they strongly rub their feet and little ones, like dressed norms go off, because of the fact that the straps in the butt then rub the chew, i do not advise this product, it is better to add and take crocs.

Examples from https://www.aliexpress.com

Data Annotation

- Manual
  - *Amazon Mechanical Turk*
  - *Appen*, *Figure 8*, *Defined Crowd*
  - *The crowd*
  - *You and your friends :)*
- Automated

Zillions of formats: *xml, json, csv, conll, wdiff, tab-separated, etc.*

Zillions of tools: *brat, gate, WebAnno, UIMA, etc.*

3. Metrics

Intrinsic - evaluate on the test sets for this specific task

- e.g., quality of syntactic parsing, language identification, NER, etc.

Extrinsic - evaluate on the test sets for the broader task

- e.g., sentiment analysis, machine translation, recommendation system

Traditional metrics

$$Precision = \frac{TPs}{TPs + FPs}$$

$$F\text{-}score = \frac{Precision * Recall}{Precision + Recall}$$

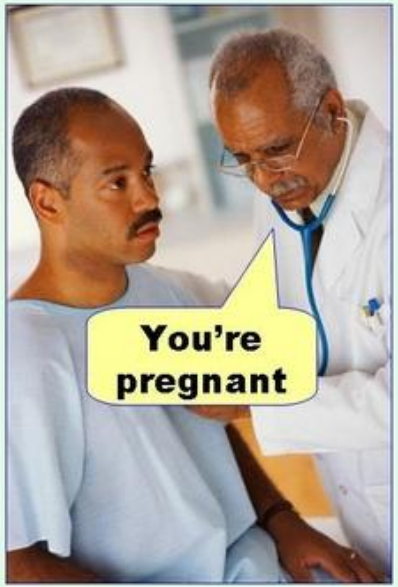
$$Recall = \frac{TPs}{TPs + FNs}$$

$$Accuracy = \frac{TPs + TNs}{TPs + TNs + FPs + FNs}$$

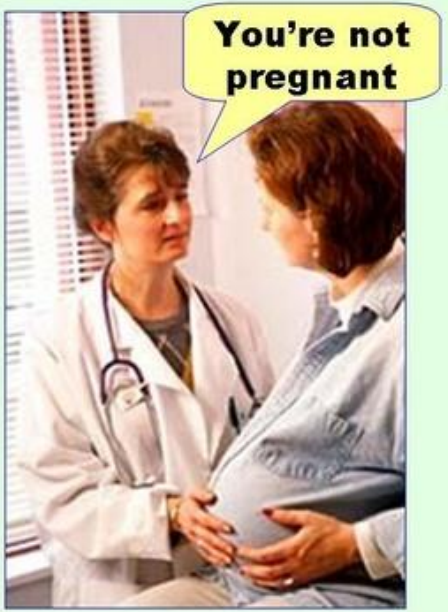
		Condition (as determined by "Gold standard")			
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive / Σ Total population	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV, Precision) = Σ True positive / Σ Test outcome positive	False discovery rate (FDR) = Σ False positive / Σ Test outcome positive
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = Σ False negative / Σ Test outcome negative	Negative predictive value (NPV) = Σ True negative / Σ Test outcome negative
	Positive likelihood ratio (**LR+**) = TPR/FPR	True positive rate (TPR, Sensitivity, Recall) = Σ True positive / Σ Condition positive	False positive rate (FPR, Fall-out) = Σ False positive / Σ Condition negative	Accuracy (ACC) = Σ True positive + Σ True negative / Σ Total population	
	Negative likelihood ratio (**LR−**) = FNR/TNR	False negative rate (FNR) = Σ False negative / Σ Condition positive	True negative rate (TNR, Specificity, SPC) = Σ True negative / Σ Condition negative		
	Diagnostic odds ratio (**DOR**) = LR+/LR−				

Traditional metrics

Type I error
(false positive)



Type II error
(false negative)



Confusion matrices

Languages	English	German	French	Italian	Dutch	Spanish
English	**9244**	38	199	145	222	139
German	28	**9514**	67	29	325	27
French	20	52	**9525**	165	83	160
Italian	6	7	18	**9822**	16	134
Dutch	60	66	35	20	**9800**	19
Spanish	6	8	41	242	24	**9679**

And many more...

- ROC Curves
- BLEU, Rouge, METEOR
- GLEU, Max-match
- Perplexity
- Parseval, cross-bracketing, leaf-ancestor...
- Outliers
- Human evaluation
- Your metric :)

4. Solutions

Baseline — solution that can be built using a reasonable primitive approach. *(It's always good to build a few.)*

State-of-the-art, or SOTA, — the best publicly known solution tested on a public data set with commonly accepted metrics.

Improve ***iteratively***:

- extend coverage (i.e., improve recall)
- reduce error rate (i.e., improve precision)

4. Solutions

. . . as needed.

5. Feedback analysis

- Is the input data the same as you expected?
- How well does the system perform on real data?
- How do users react to the system?
- Do you see a decrease in retention?
- And so on, and so forth...

Aitäh! Kas küsimusi on?

Interesting References

- Kyubyong, *NLP Tasks and Selected References* (2017)
- Sebastian Ruder, *Tracking Progress in Natural Language Processing* (ongoing)
- Nick Montfort, *World Clock* (2013) and code
- OpenAI blog, *Better Language Models and Their Implications* (2019)
- MIT-IBM Watson AI lab and HarvardNLP, *Catching a Unicorn with GLTR: A tool to detect automatically generated text* (2019)
- *Evaluating language identification performance* (2015)
- Peter Norvig, *The Unreasonable Effectiveness of Data* (2011)
- Popular libraries for NLP: *spaCy*, *nltk*, *Stanford CoreNLP*, *OpenNLP*, *Emory NLP*, *AllenNLP*