Machine learning for natural language processing tasks

Aleksey Kulnevich,

Vladislav Radishevskii

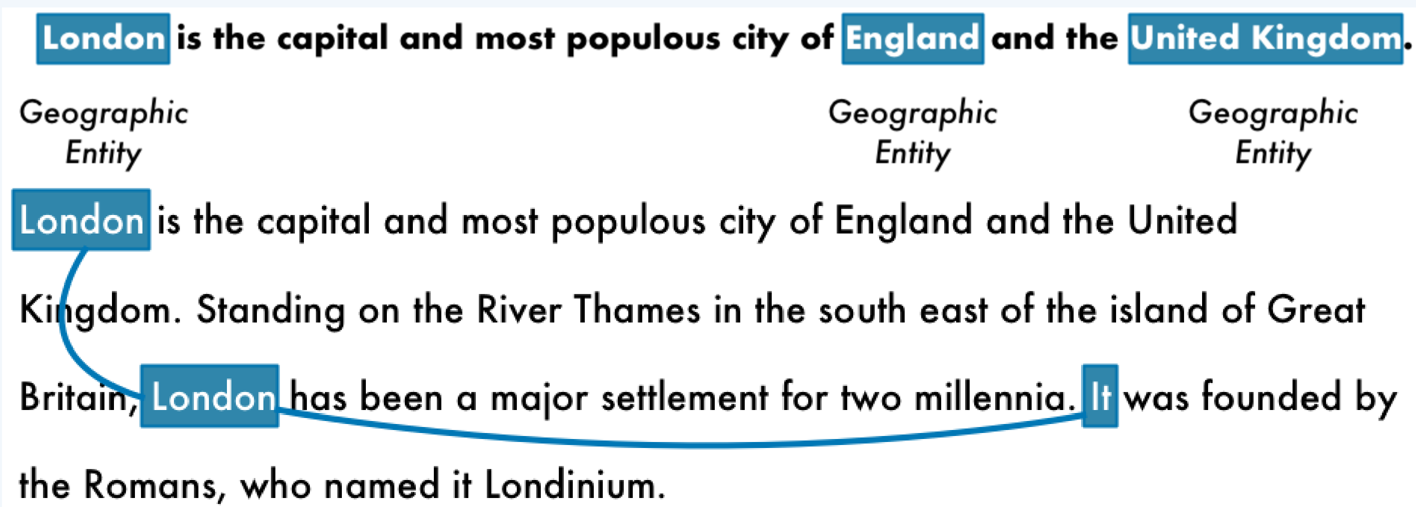
13 September 2018
Dubna, Russia

1 Introduction

Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve result from experience without being explicitly programmed.

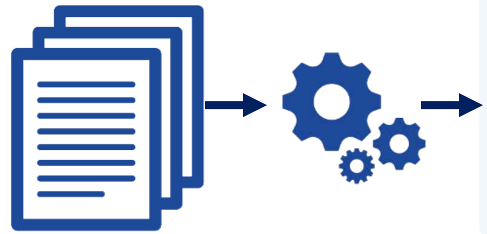
Natural Language Processing, or NLP, is the sub-field of AI that is focused on enabling computers to understand and process human languages.



Machine Learning

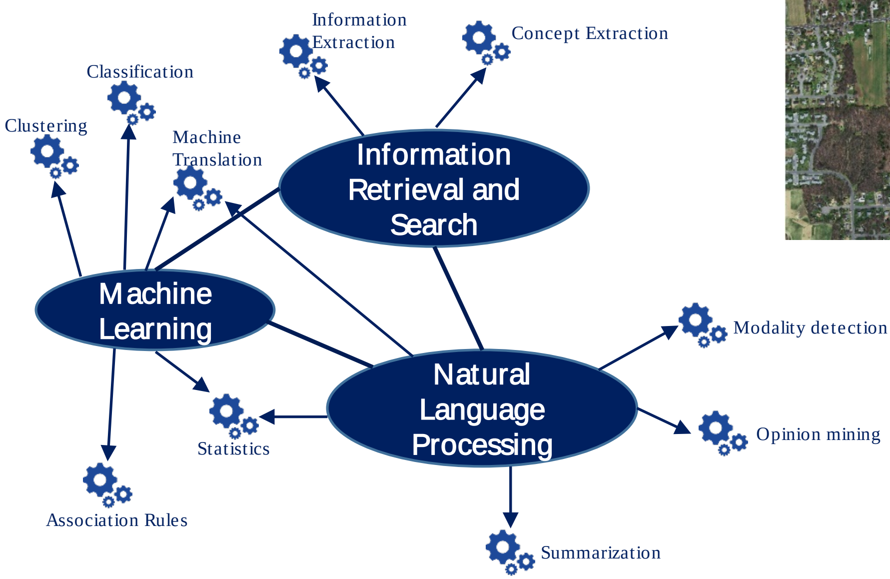
ML TASKS

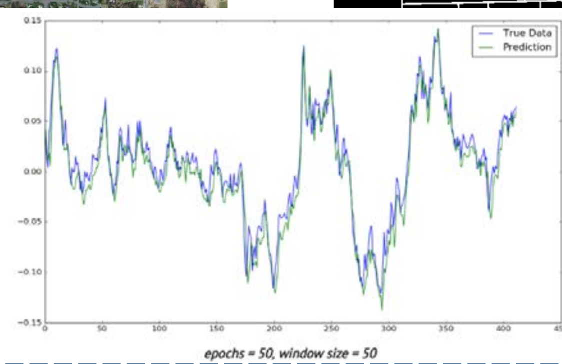
- *Information Extraction*
- *Machine Translation*
- *Image segmentation*
- *Object detection*
- *Gap filling*
- *Predictive analytics*



Document structure:

- *Named entity recognition*
- *Coreference resolution*
- *Type*
- *Keywords*
- *Attributes*

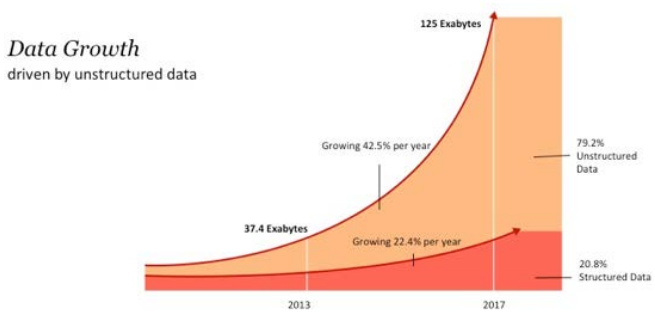


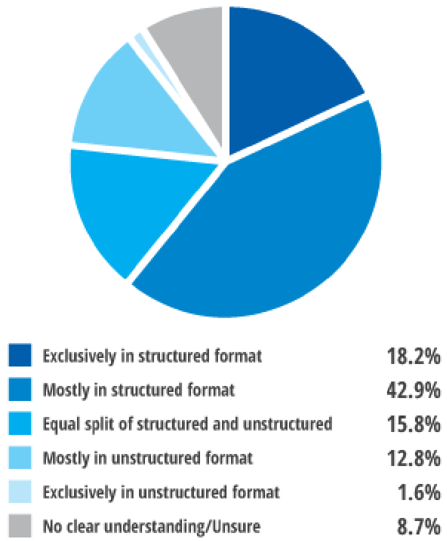


The importance of Text Analytics

Structured / Unstructured Text Data

- *Structured data represents only **20%** of the information available to an organization*
- ***80%** of all the data is in unstructured form*



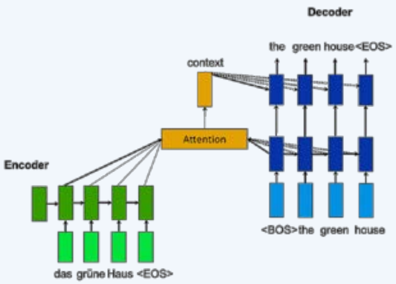


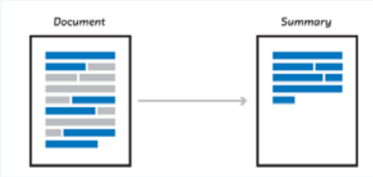
- *If structured data is big, then unstructured data is huge*
- *Text analytics is the science of turning unstructured text into structured data*

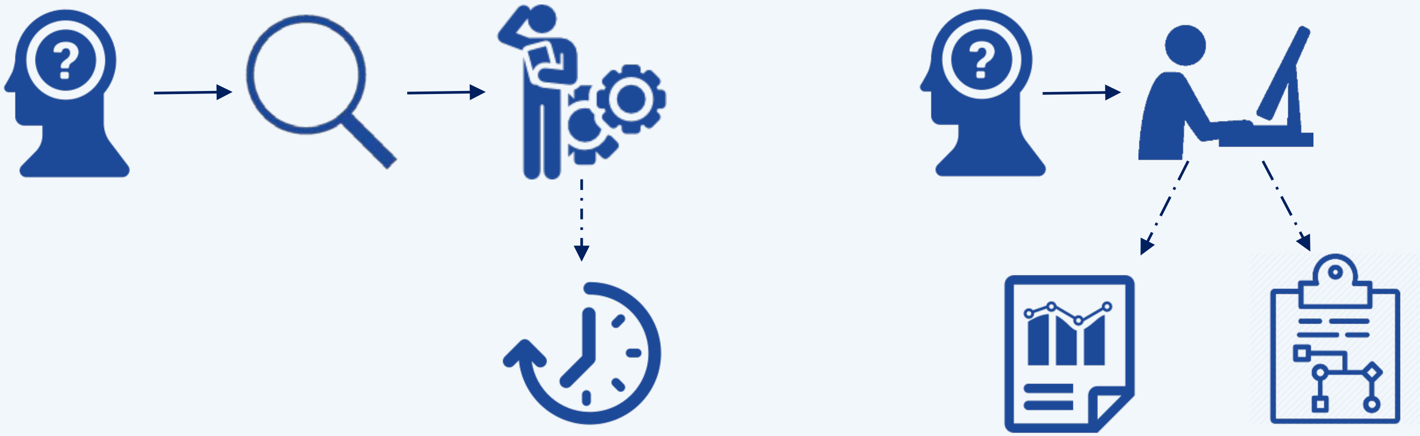
Natural Language Processing

NLP TASKS

- *Named Entity Recognition*
- *Coreference Resolution*
- *Neural Machine Translation*
- *Chatbots*
- *Summary Extraction*
- *Answering Questions*
- *Ontology building*







Objective of work

❖ ***Feature extraction***

Word embeddings, char embeddings, morphological and additional tags

❖ ***Building machine learning model for Named Entity Recognition***

Hybrid approach Bi-LSTM + CRF model

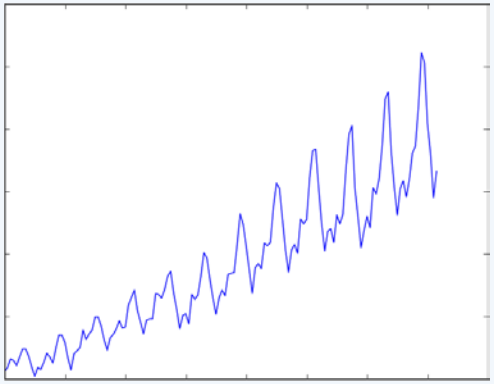
❖ ***Building machine learning model for Coreference Resolution***

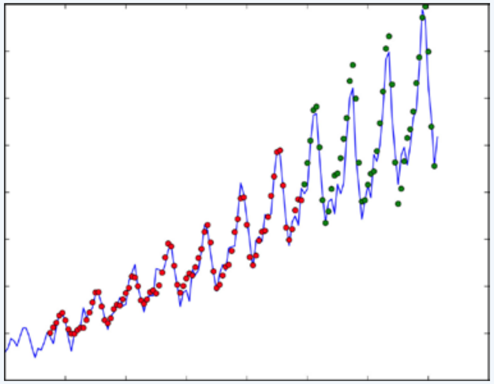
Bi-LSTM model

Vectorization problem

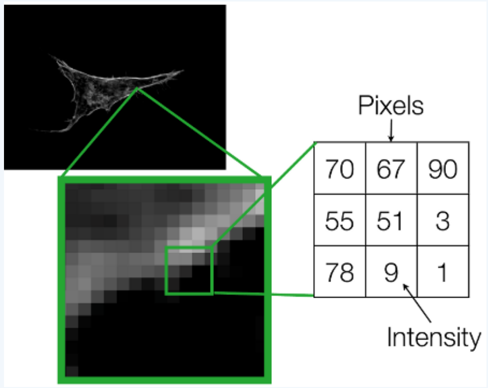
Natural Language Processing

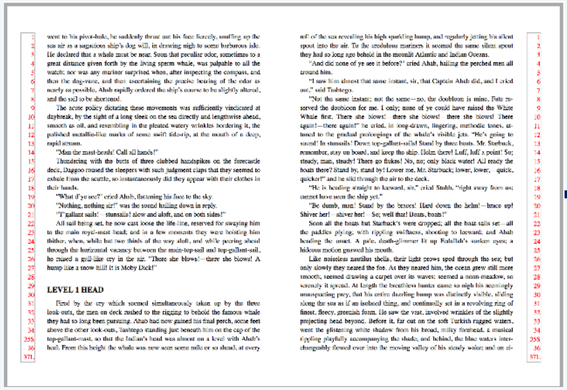
Vector Extraction





$$x = \{x_0, x_1 .., x_n\}$$
$$y = x_{n+1}$$
$$x = \{x_{n+2}, x_{n+3} .., x_{n+k}\}$$
$$y = x_{n+k+1}$$
$$\ldots$$

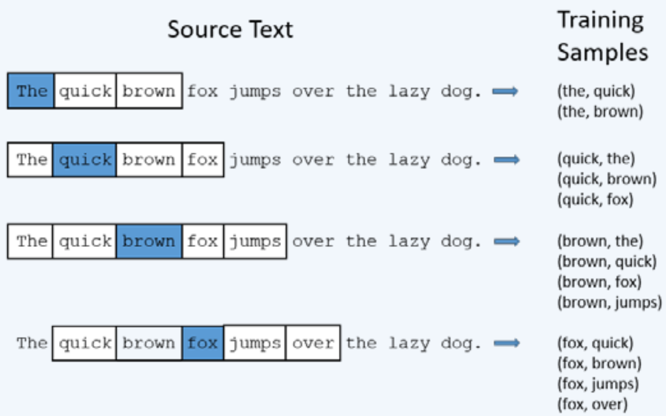


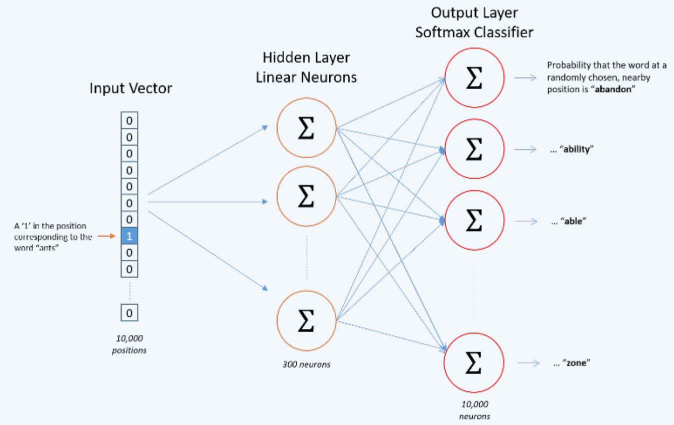


Natural Language Processing

HOW IT WORKS

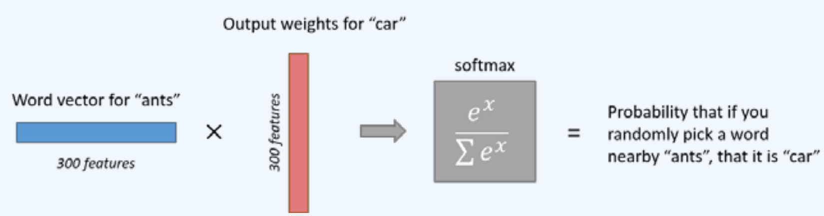
Word2Vec – The Skip-Gram Model





Intuition

If two different words have very similar "contexts" (that is, what words are likely to appear around them), then the model needs to output very similar results for these two words.



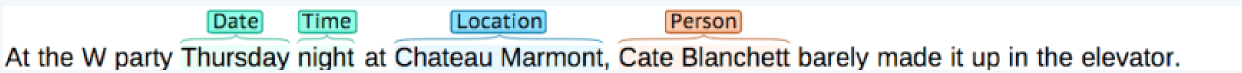
3

Named Entity Recognition

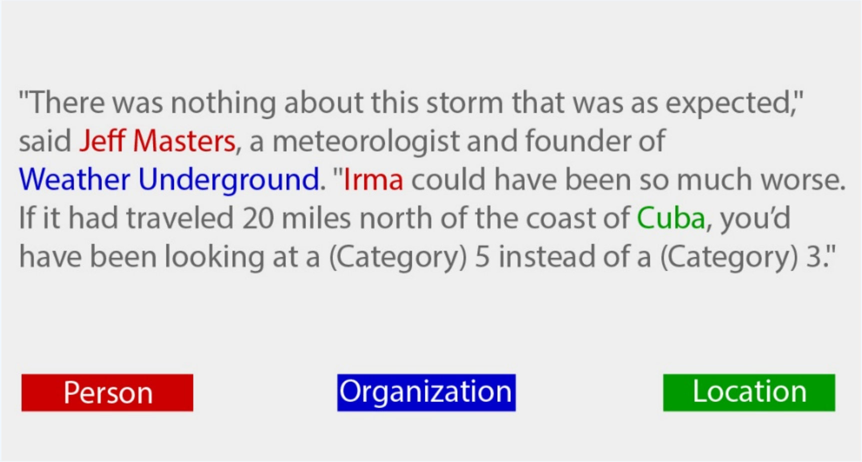
Natural Language Processing

Named Entity Recognition

Entity is concrete object of some type. For example, **Geoffrey Hinton** is an **entity** of ***type "Person"***.



At the W party **<Date>** *Thursday* **</ Date> <Time>** *night* **</ Time>** *at* **<Location>** *Chateau Marmont* **</ Location>**, **<Person>** *Cate Blanchett* **</ Person>** *barely made it up in the elevator.*



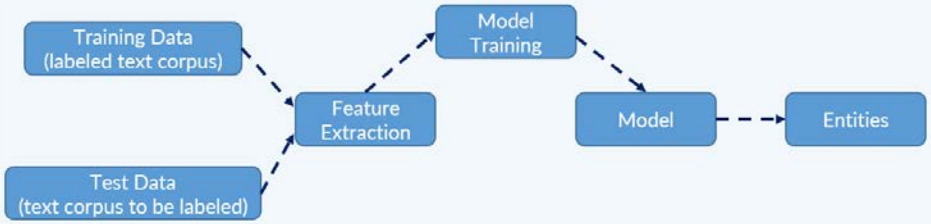
Natural Language Processing

Named Entity Recognition

Named Entity Recognition is subtask of information extraction that seeks to locate and classify *named entities* in text into pre-defined categories.

ML Approach

+ Flexibility
- Data for training

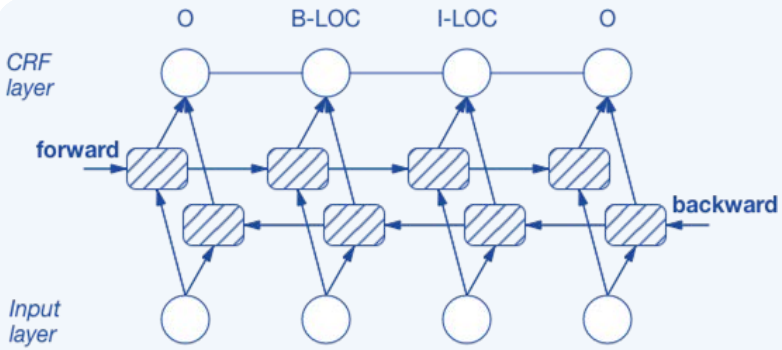


RB Approach

+ The ability to quickly find certain type of entities
- The need for specialized knowledge of linguistics
- Multiple rules

Natural Language Processing

Named Entity Recognition



$$p(\mathbf{y}|\mathbf{x}) = \frac{e^{Score(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{y}'} e^{Score(\mathbf{x},\mathbf{y}')}},$$

$$Score(\mathbf{x},\mathbf{y}) = \sum_{i=0}^{T} A_{y_i,y_{i+1}} + \sum_{i=1}^{T} P_{i,y_i},$$

***Combination** of **CRF** model with a **Bi-LSTM** neural network encoding should increase the accuracy of the tagging decisions*

The CRF model is trained to predict a vector y = {$y_0, y_1 .., y_T$} or tags given a sentence x = {$x_0, x_1 .., x_T$}.
*where **A represents** score of transition from tag I to tag j*
***P represents** score of the j^{th} tag of the word i^{th}*

Natural Language Processing

Named Entity Recognition

Features for improving accuracy:

- ❖ Word embedding
- ❖ Char embedding
- ❖ Morphological features
- ❖ Additional tags: GEO, Orgn, Trad tm

Language: Python

Frameworks: NLTK, Numpy, Keras, Tensorflow

Natural Language Analysis
with Python NLTK

Keras

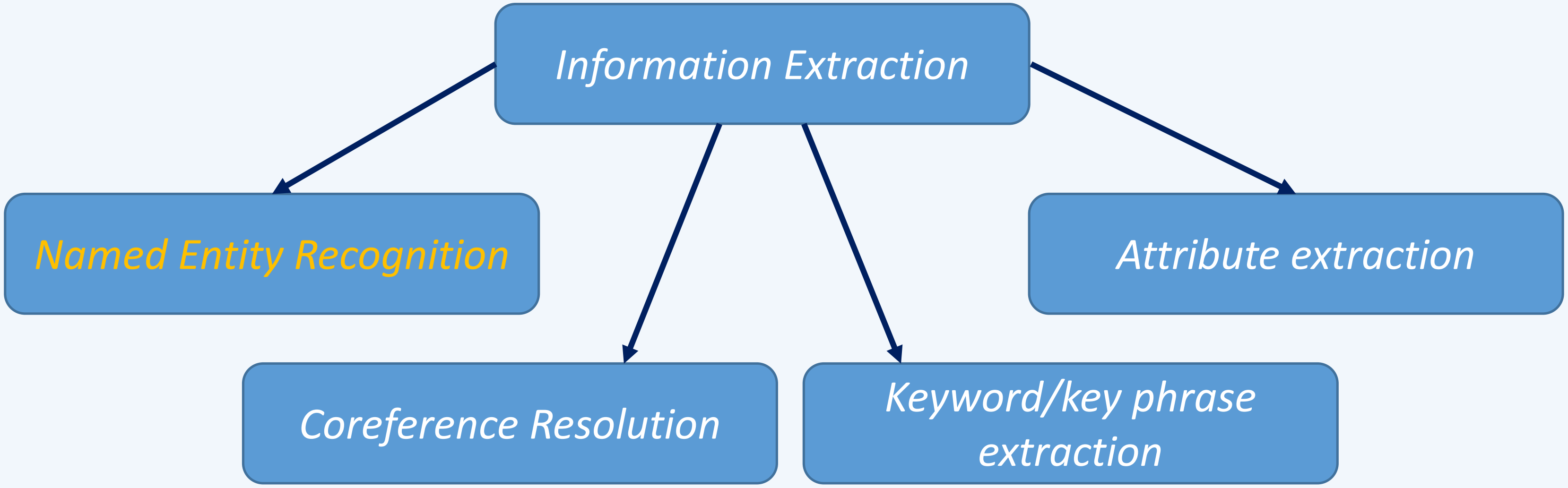
Natural Language Processing

Named Entity Recognition

Entity type	Precision	Recall	F1-Score	Support
I-ORG	0.81	0.80	0.81	1158
B-PROD	0.71	0.61	0.66	1590
B-LOC	0.82	0.85	0.83	1257
I-LOC	0.75	0.78	0.76	529
I-PER	0.89	0.87	0.88	919
B-ORG	0.78	0.73	0.76	1056
I-PROD	0.68	0.59	0.63	371
B-PER	0.85	0.86	0.86	711
I-DATE	0.97	0.98	0.97	955
B-DATE	0.91	0.92	0.91	749

Natural Language Processing

Information Extraction



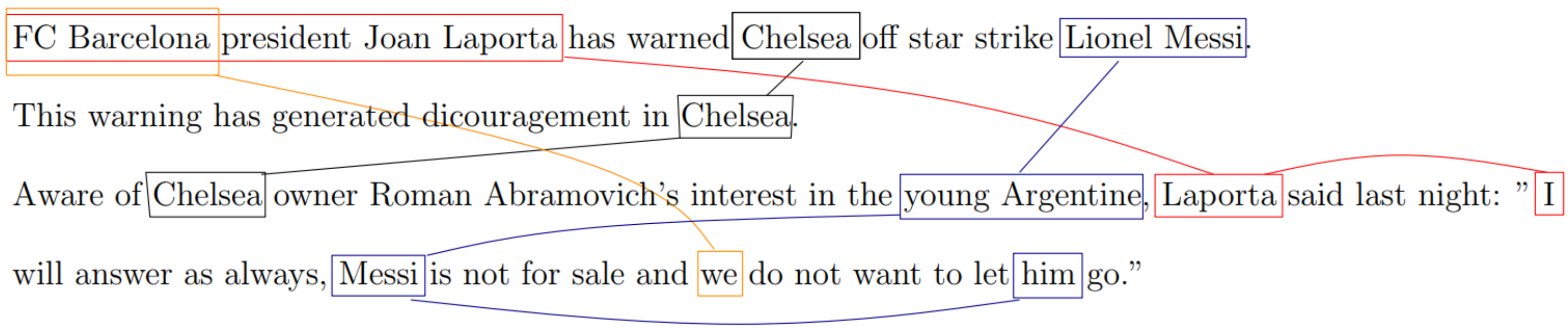
- **Named entities** – *objects of specific types*
- **Coreference** – *chain of mentioning the named entity*
- **Keywords** – *are ideas and concepts that define what content is about*
- **Attributes** – *properties of objects*

4

Coreference Resolution

Coreference Resolution

Coreference, sometimes written **co-reference**, occurs when two or more expressions in a text refer to the same person or thing; they have the same referent.



Coreference Resolution – process of determining which mentions in a discourse refer to the same entity.

Types of Coreference

Anaphora

- **The music** was so loud that **it** couldn't be enjoyed.
- **Our neighbors** dislike the music. If **they** are angry, the cops will show up soon.

Cataphora

- If **they** are angry about the music, **the neighbors** will call the cops.
- Despite **her** difficulty, **Wilma** came to understand the point.

Split antecedents

- **Carol** told **Bob** to attend the party. **They** arrived together.
- When **Carol** helps **Bob** and **Bob** helps **Carol**, **they** can accomplish any task.

Noun phrases

- Queen Elizabeth set about transforming **her husband**, **King George VI**, into a viable monarch. **Lionel Logue, a renowned speech therapist**, was summoned to help **the King** overcome his speech impediment.

Supervised Approach

Based mainly on two methods:

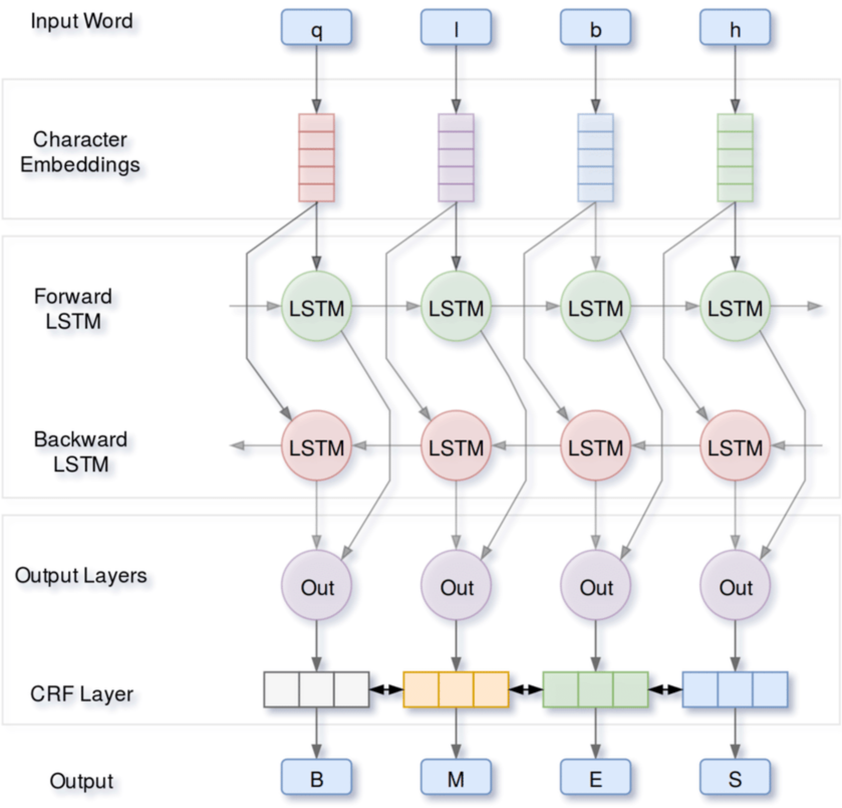
- ❖ Binary classification
- ❖ Ranking method

Pros:

- ❖ Learning algorithms usually generalize well

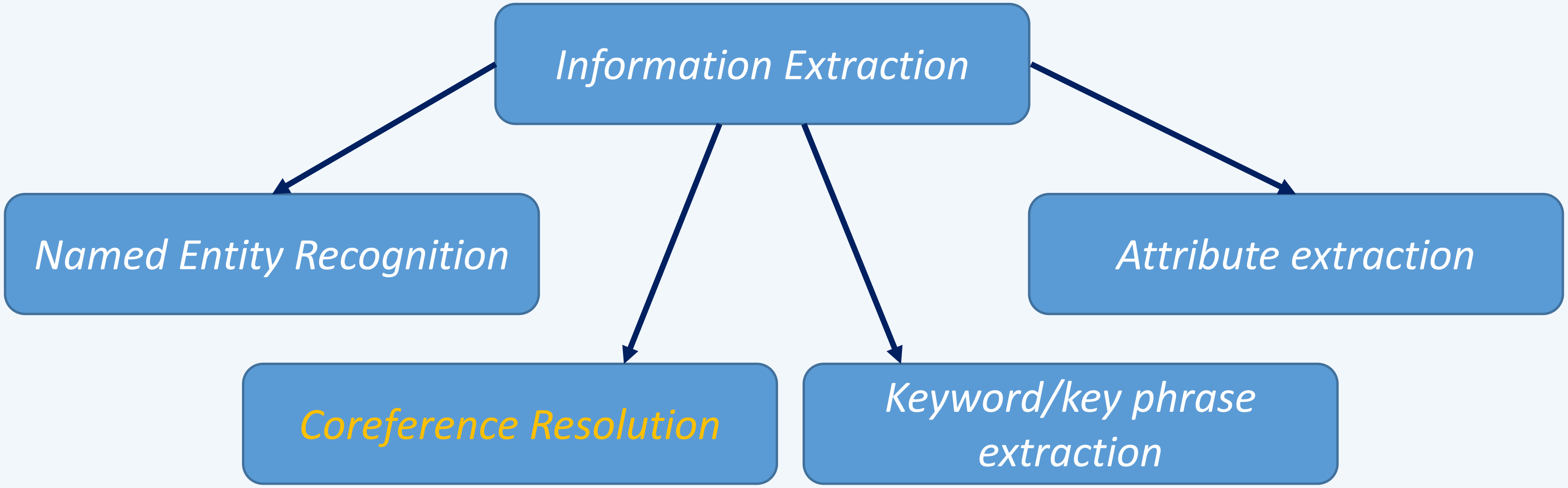
Cons:

- ❖ Quality is limited to quantity and quality of data
- ❖ Requires labeled data



Natural Language Processing

Information Extraction



- **Named entities** – *objects of specific types*
- **Coreference** – *chain of mentioning the named entity*
- **Keywords** – *are ideas and concepts that define what content is about*
- **Attributes** – *properties of objects*

Conclusion

We created neural network for solving named entity recognition and Coreference resolution problems:

- **Bi-LSTM Neural Network + CRF layer**

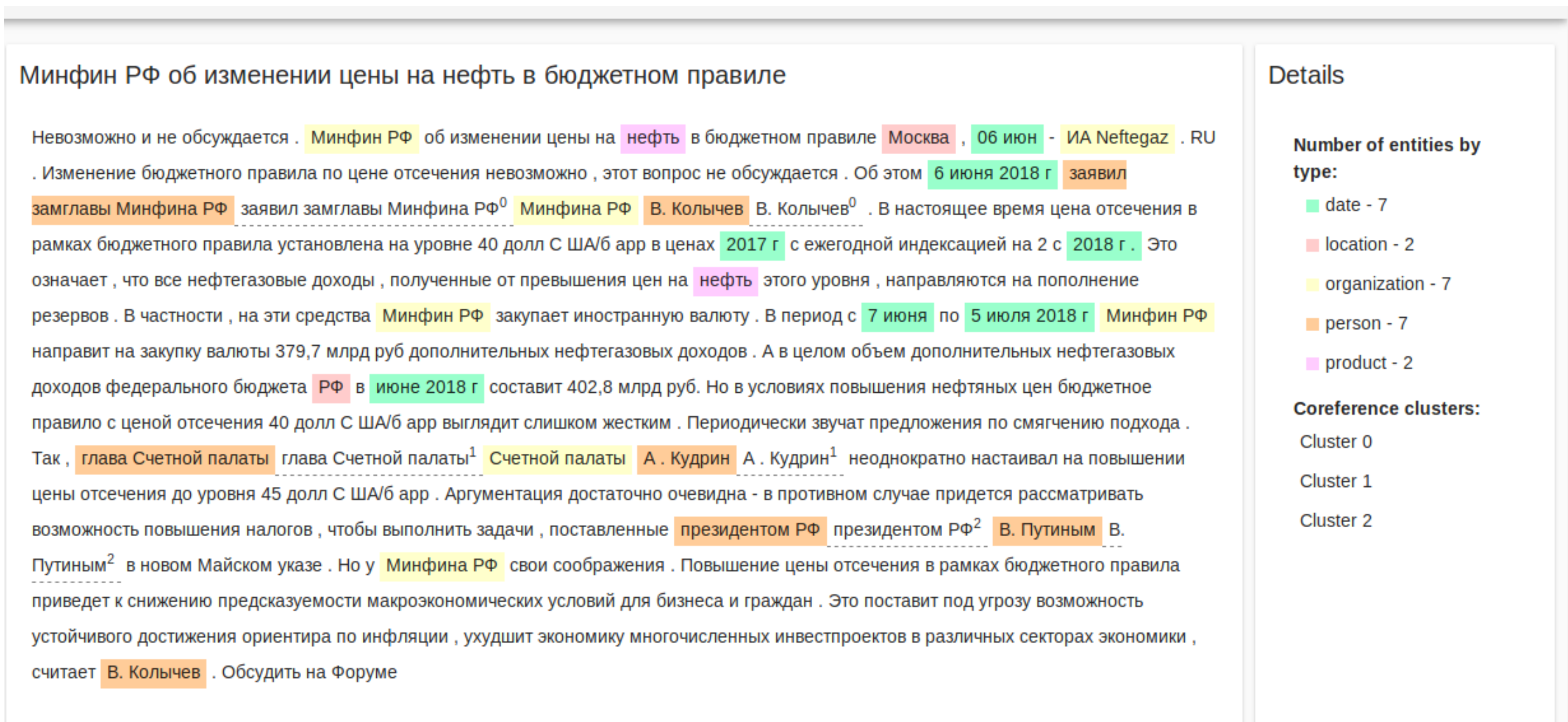
Input: word embeddings, char embeddings, morphological tags + geo tags + extra tags

- **Bi-LSTM Neural Network**

Input: word embeddings, char embeddings, morphological tags and names of entities (result of previous neural network)

Conclusion

This functionality is part of the text analytics system.
Visually it looks like this:



Thanks for your attention!

Aleksey.Kulnevich@econophysica.com

a.d.kulnevich@gmail.com