

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.

London is the capital and most populous city of **England** and the **United Kingdom**.

*Geographic
Entity*

*Geographic
Entity*

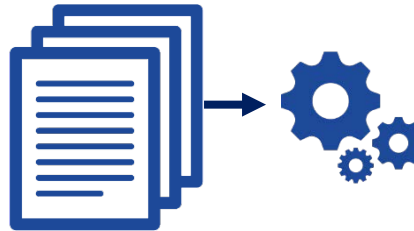
*Geographic
Entity*

London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, **London** has been a major settlement for two millennia. **It** was founded by the Romans, who named it Londinium.

Machine Learning

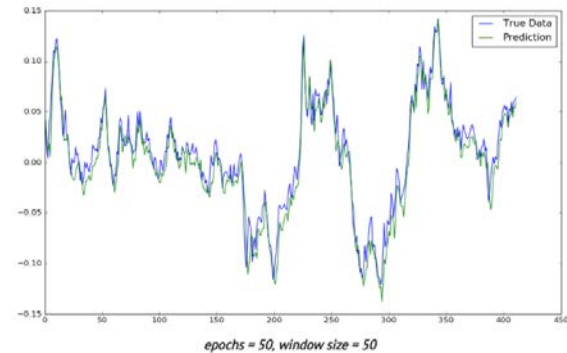
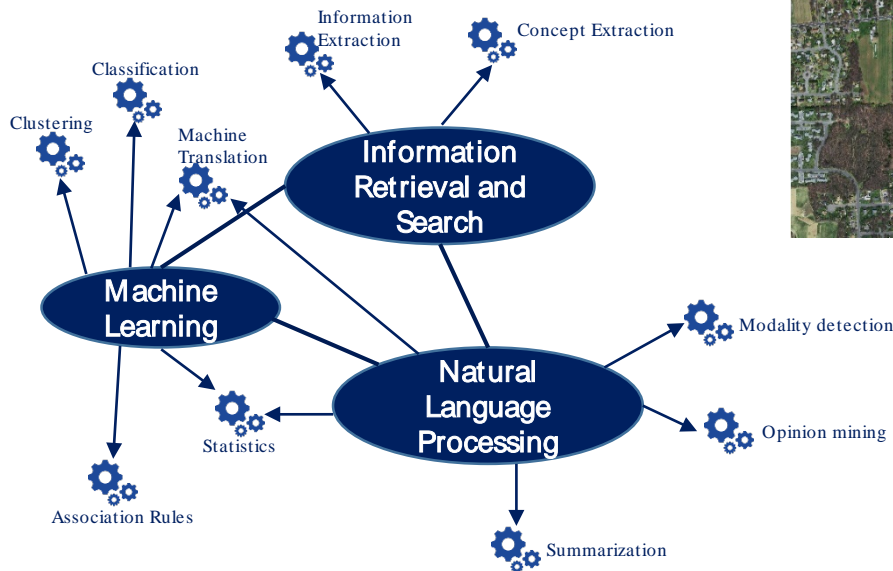
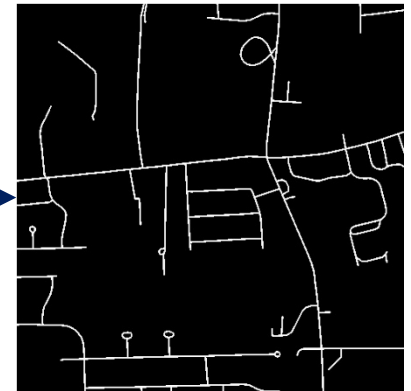
ML TASKS

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



Document structure:

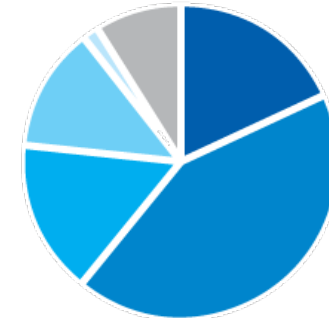
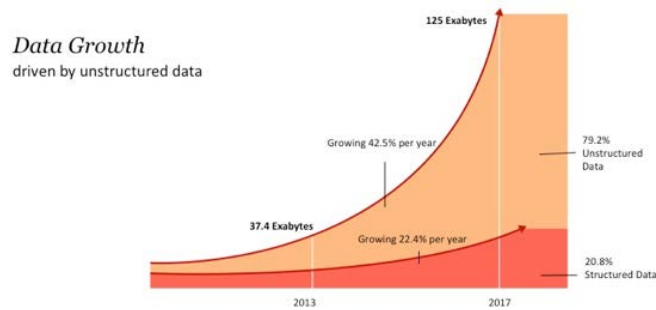
- ❖ *Named entity recognition*
- ❖ *Coreference resolution*
- ❖ *Type*
- ❖ *Keywords*
- ❖ *Attributes*



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*



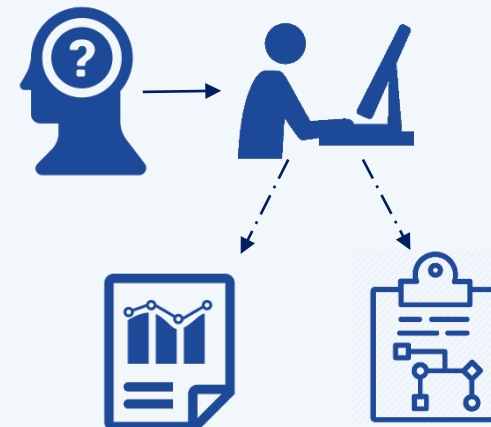
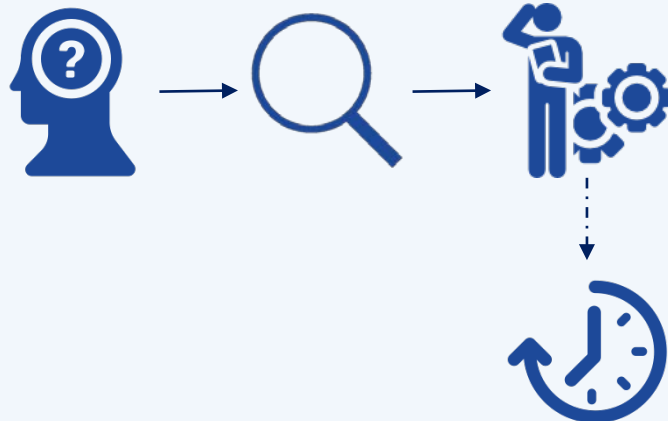
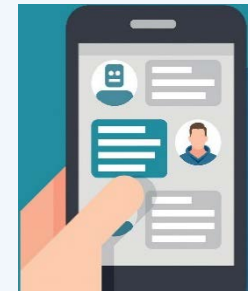
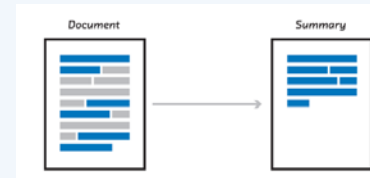
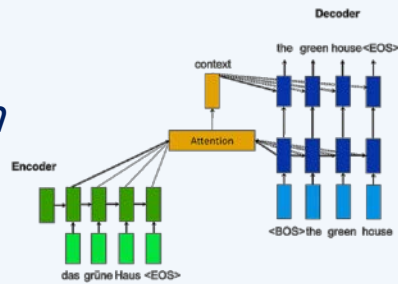
Exclusively in structured format	18.2%
Mostly in structured format	42.9%
Equal split of structured and unstructured	15.8%
Mostly in unstructured format	12.8%
Exclusively in unstructured format	1.6%
No clear understanding/Unsure	8.7%

- ❖ *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*

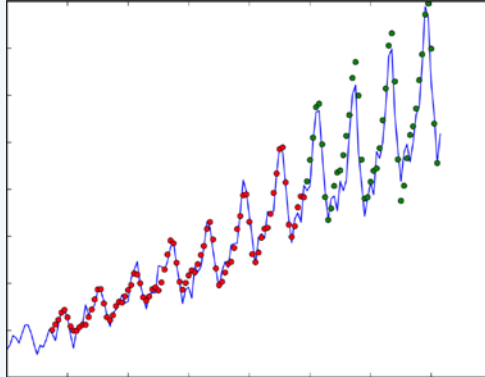
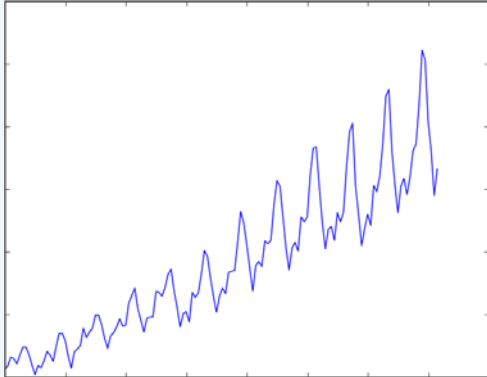
Bi-LSTM model

2

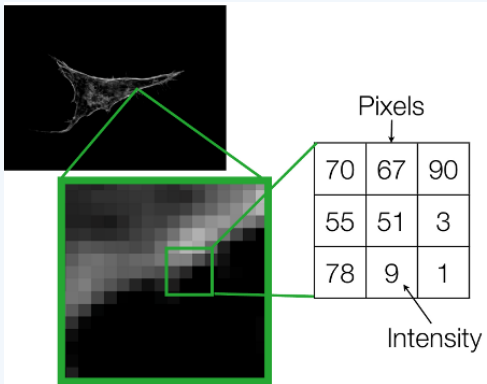
Vectorization problem

Natural Language Processing

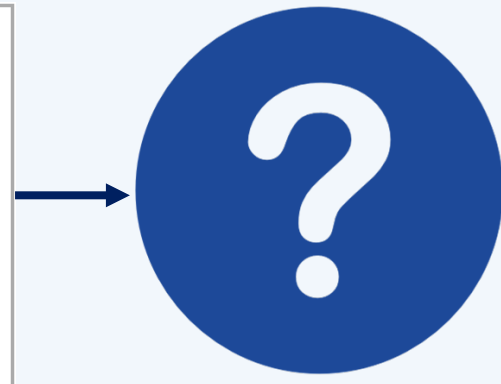
Vector Extraction



$$\begin{aligned}
 X &= \{x_0, x_1, \dots, x_n\} \\
 Y &= x_{n+1} \\
 X &= \{x_{n+2}, x_{n+3}, \dots, x_{n+k}\} \\
 Y &= x_{n+k+1} \\
 &\dots
 \end{aligned}$$



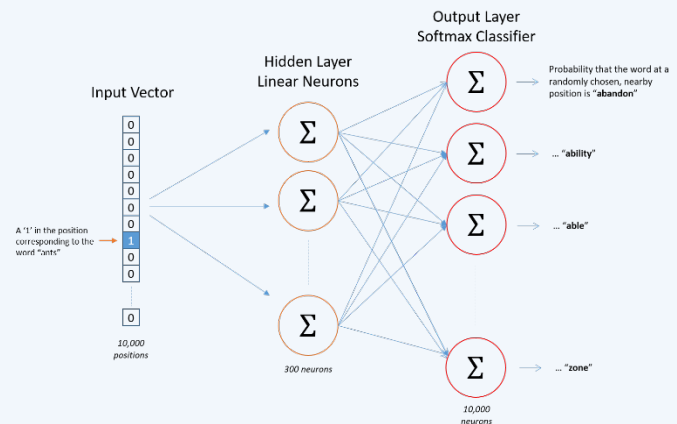
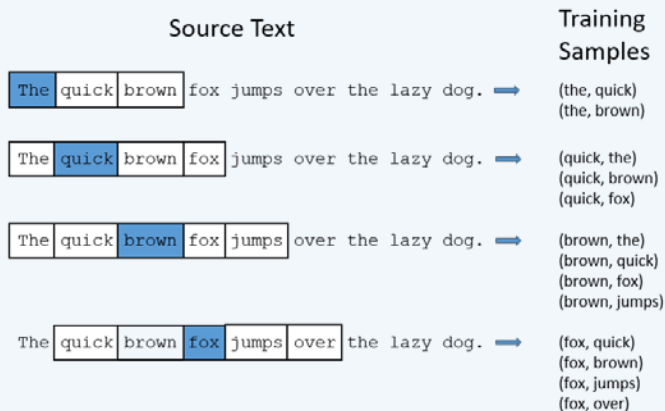
3 went to his private, he hardly threw out his face, hardly smiling up the
 4 sea up as a seaman ship's dog will, in diving right to some bottom side.
 5 He declared that a whale took to him. From that peculiar time, attention to a
 6 great distance given forth by the living sperm whale, was valuable to all the
 7 whalers, not was another suspicion when, after repeating the compass, and
 8 then the log-book, and then, considering the gale's bearing of the ship, so
 9 nearly as possible, Ahab rightly ordered the ship's course to be slightly altered,
 10 and the sail be decreased.
 11 The same policy directing these movements was sufficiently vindicated at
 12 midnight, by the sight of a long line on the sea directly and straightly ahead,
 13 smooth as oil, and receding in the placid watery wrinkles bordering it, the
 14 perfect and finished tracks of some well-tended, or the work of a deep,
 15 rapid current.
 16 "That the next to-morrow! Call it back!"
 17 "Inaudible with the roar of these dashed huzze-bombs on the forecastle
 18 deck, wharves round the ropes with such judgment taps that they seemed to
 19 chatter from the masts, or continuously did they agree with their partners in
 20 this bark."
 21 "What's a'goin'?" cried Ahab, throwing his face to the sky.
 22 "Nothing, nothing, sir!" was the answer heeling down in reply.
 23 "I'll tell you!" "What's a'goin'?" cried Ahab, and on both sides.
 24 "All sail being set, the new east breeze the life line, reserved for carrying him
 25 to the main-peak mast, and in a few moments they were hoisting his
 26 dunnage, which, but few blocks of the very ship, and which passing round
 27 through the horizontal courses between the main-top-sail and top-gallant-sail,
 28 he raised a well-see eye in its size. "There she blows!" "There she blows!" A
 29 heap like a snow hill in a Moby Dick!
 30
 31 **LEVEL I HEAD**
 32 "Fixed by the cry which seemed simultaneously taken up by the three
 33 look-out, the men on deck rushed to the rigging to behold the famous whale
 34 they had so long been pursuing. Ahab had never granted his first speech, nor had
 35 above the other look-out, wharves standing just beneath him on the cap of the
 36 top-gallant-sail, and that the look-out took was shown as a level with Ahab's
 37 head. From this height the whale was now seen some miles or so ahead, at every
 38 101.
 39
 40
 41
 42
 43
 44
 45
 46
 47
 48
 49
 50
 51
 52
 53
 54
 55
 56
 57
 58
 59
 60
 61
 62
 63
 64
 65
 66
 67
 68
 69
 70
 71
 72
 73
 74
 75
 76
 77
 78
 79
 80
 81
 82
 83
 84
 85
 86
 87
 88
 89
 90
 91
 92
 93
 94
 95
 96
 97
 98
 99
 100
 101



Natural Language Processing

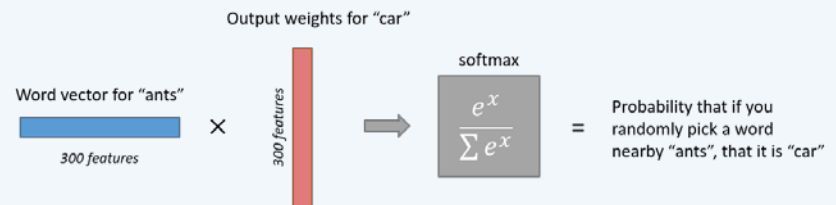
HOW IT WORKS

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 Time night at Location Chateau Marmont, Person Cate Blanchett barely made it up in the elevator.

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.

"There was nothing about this storm that was as expected," said **Jeff Masters**, a meteorologist and founder of **Weather Underground**. "**Irma** could have been so much worse. If it had traveled 20 miles north of the coast of **Cuba**, you'd have been looking at a (Category) 5 instead of a (Category) 3."

Person

Organization

Location

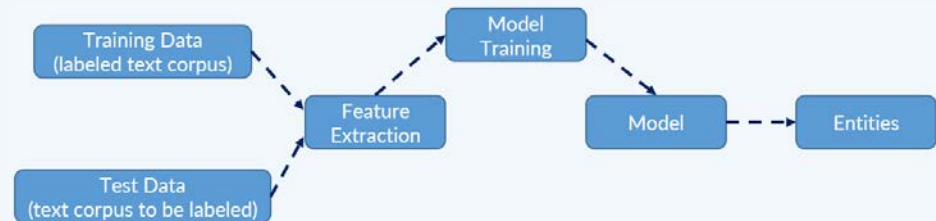
Natural Language Processing

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.

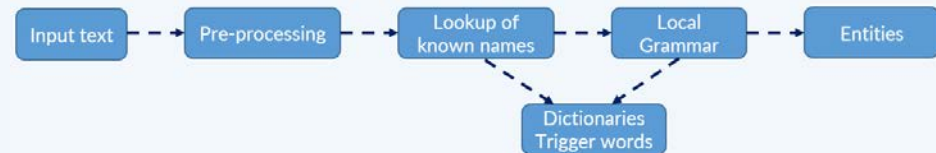
ML Approach

- + Flexibility
- Data for training



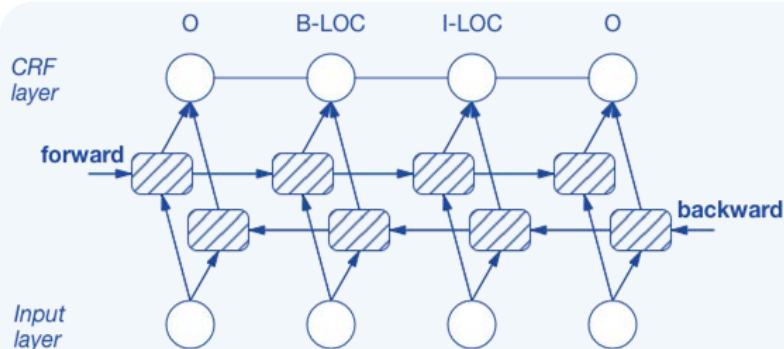
RB Approach

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



Natural Language Processing

Named Entity Recognition



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

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

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 $\mathbf{y} = \{y_0, y_1, \dots, y_T\}$ or tags given a sentence $\mathbf{x} = \{x_0, x_1, \dots, x_T\}$.

A represents score of transition from tag l 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

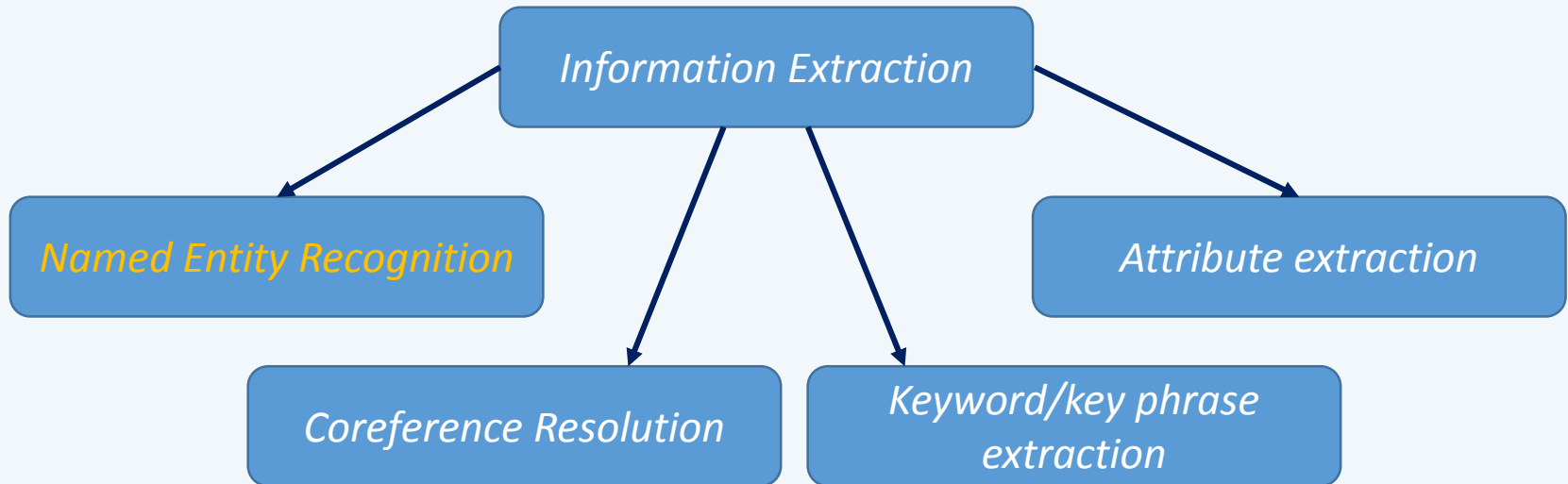
Natural Language Processing

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.

FC Barcelona president Joan Laporta has warned Chelsea off star strike Lionel Messi.
This warning has generated discouragement in Chelsea.
Aware of Chelsea owner Roman Abramovich's interest in the young Argentine, Laporta said last night: "I will answer as always, Messi is not for sale and we do not want to let him go."

The diagram shows three sentences with entities highlighted in colored boxes and lines connecting them to show coreference. In the first sentence, 'FC Barcelona' (orange), 'Joan Laporta' (red), 'Chelsea' (black), and 'Lionel Messi' (blue) are boxed. In the second sentence, 'Chelsea' (black) is boxed. In the third sentence, 'Chelsea' (black), 'young Argentine' (blue), 'Laporta' (red), 'I' (red), 'Messi' (blue), 'we' (orange), and 'him' (blue) are boxed. Lines connect 'FC Barcelona' to 'Chelsea' in the second sentence. Lines connect 'Joan Laporta' to 'Laporta' in the third sentence. Lines connect 'Lionel Messi' to 'Messi' and 'him' in the third sentence. Lines connect 'Chelsea' in the second sentence to 'Chelsea' and 'we' in the third sentence. Lines connect 'young Argentine' to 'Messi' and 'him' in the third sentence. Lines connect 'I' to 'Laporta' in the third sentence.

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:

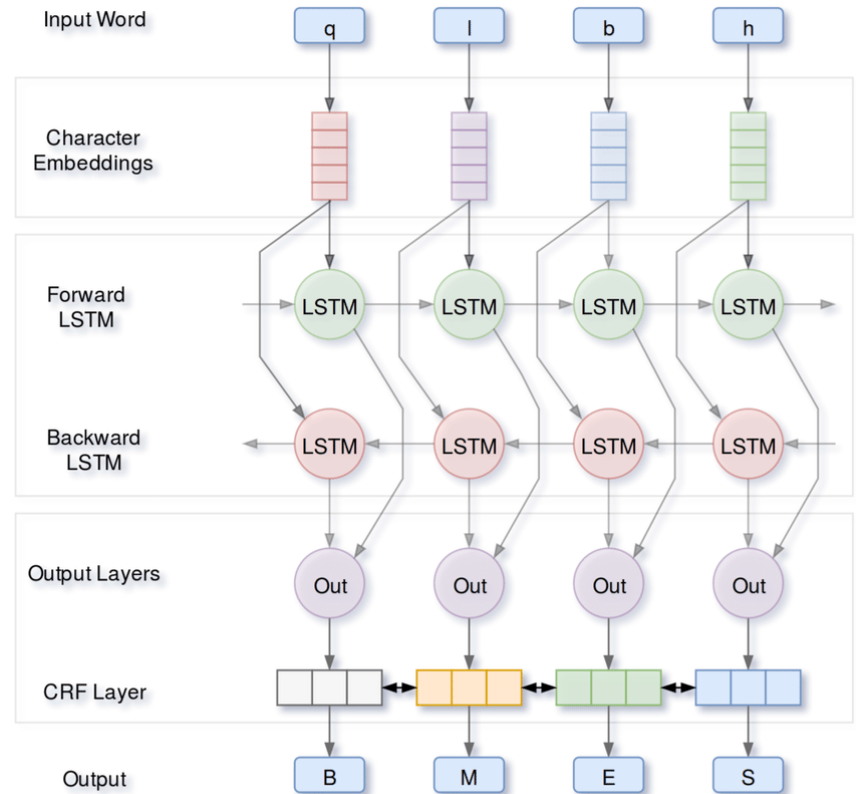
- ❖ Binary classification
- ❖ Ranking method

Pros:

- ❖ Learning algorithms usually generalize well

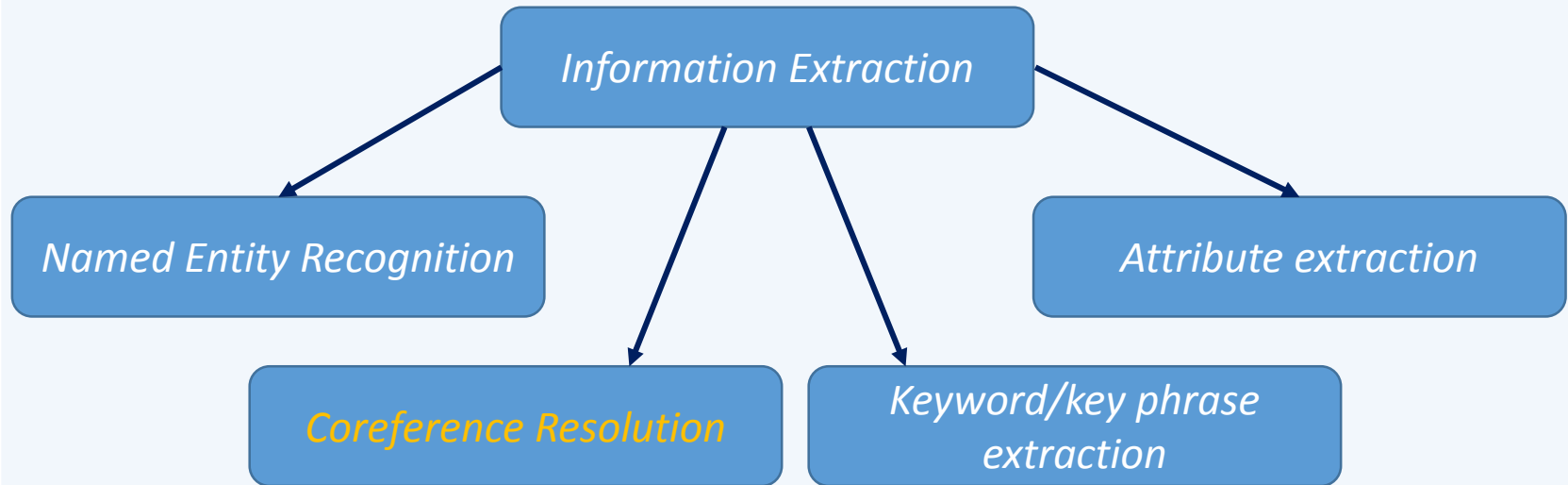
Cons:

- ❖ 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:

Минфин РФ об изменении цены на нефть в бюджетном правиле

Невозможно и не обсуждается . Минфин РФ об изменении цены на нефть в бюджетном правиле Москва , 06 июн - IA Neftegaz . RU . Изменение бюджетного правила по цене отсечения невозможно , этот вопрос не обсуждается . Об этом 6 июня 2018 г заявил замглавы Минфина РФ заявил замглавы Минфина РФ⁰ Минфина РФ В. Колычев В. Колычев⁰ . В настоящее время цена отсечения в рамках бюджетного правила установлена на уровне 40 долл С ША/б арр в ценах 2017 г с ежегодной индексацией на 2 с 2018 г . Это означает , что все нефтегазовые доходы , полученные от превышения цен на нефть этого уровня , направляются на пополнение резервов . В частности , на эти средства Минфин РФ закупает иностранную валюту . В период с 7 июня по 5 июля 2018 г Минфин РФ направит на закупку валюты 379,7 млрд руб дополнительных нефтегазовых доходов . А в целом объем дополнительных нефтегазовых доходов федерального бюджета РФ в июне 2018 г составит 402,8 млрд руб. Но в условиях повышения нефтяных цен бюджетное правило с ценой отсечения 40 долл С ША/б арр выглядит слишком жестким . Периодически звучат предложения по смягчению подхода . Так , глава Счетной палаты глава Счетной палаты¹ Счетной палаты А. Кудрин А. Кудрин¹ неоднократно настаивал на повышении цены отсечения до уровня 45 долл С ША/б арр . Аргументация достаточно очевидна - в противном случае придется рассматривать возможность повышения налогов , чтобы выполнить задачи , поставленные президентом РФ президентом РФ² В. Путиным В. Путиным² в новом Майском указе . Но у Минфина РФ свои соображения . Повышение цены отсечения в рамках бюджетного правила приведет к снижению предсказуемости макроэкономических условий для бизнеса и граждан . Это поставит под угрозу возможность устойчивого достижения ориентира по инфляции , ухудшит экономику многочисленных инвестпроектв в различных секторах экономики , считает В. Колычев . Обсудить на Форуме

Details

Number of entities by type:

- date - 7
- location - 2
- organization - 7
- person - 7
- product - 2

Coreference clusters:

- Cluster 0
- Cluster 1
- Cluster 2

Thanks for your attention!

Aleksey.Kulnevich@econophysica.com

a.d.kulnevich@gmail.com

