Question Answering and Chatbots

1st Practical exercise – Named Entity Recognition & Linking

Aleksandr Perevalov

aleksandr.perevalov@hs-anhalt.de

October 13, 2021

Hochschule Anhalt

Anhalt University of Applied Sciences

Text preprocessing

Tokenization? Lemmatization? Stemming?

Text preprocessing

- Tokenizaiton – dividing text into tokens (words, sub-words, BPE);

Text preprocessing

- Tokenizaiton – dividing text into tokens (words, sub-words, BPE);
- Stop words removing (but, how, or, etc.);

Text preprocessing

- Tokenizaiton – dividing text into tokens (words, sub-words, BPE);
- Stop words removing (but, how, or, etc.);
- Special characters removing (!@#$%);

Text preprocessing

- Tokenizaiton – dividing text into tokens (words, sub-words, BPE);
- Stop words removing (but, how, or, etc.);
- Special characters removing (!@#$%);
- Lemmatization – “changing” $\rightarrow$ “change”;

Text preprocessing

- Tokenizaiton – dividing text into tokens (words, sub-words, BPE);
- Stop words removing (but, how, or, etc.);
- Special characters removing (!@#$%);
- Lemmatization – “changing” $\rightarrow$ “change”;
- Stemming – “changing” $\rightarrow$ “chang”;

Text preprocessing

- Tokenizaiton – dividing text into tokens (words, sub-words, BPE);
- Stop words removing (but, how, or, etc.);
- Special characters removing (!@#$%);
- Lemmatization – “changing” $\rightarrow$ “change”;
- Stemming – “changing” $\rightarrow$ “chang”;

There is a little impact if we apply this to DNNs, for example BERT. Text preprocessing doesn't affect the modern model's quality so much.

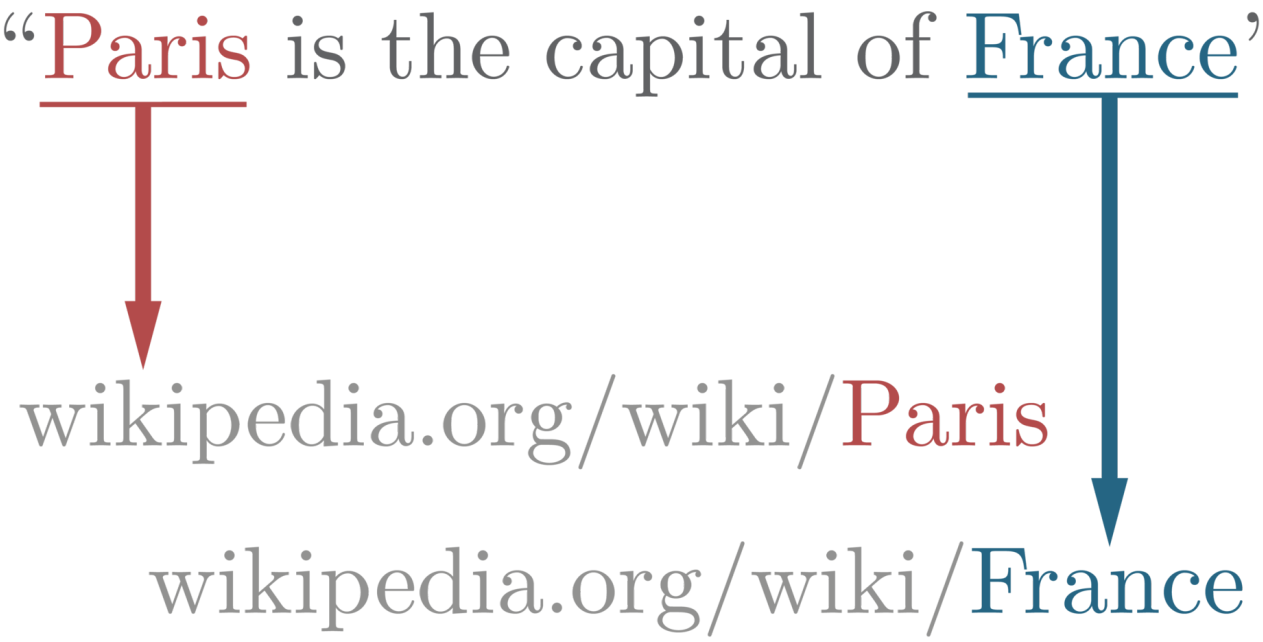
Natural Language Processing Today

Named Entity Linking (NEL)

Named Entity Recognition vs. Named Entity Linking?

Named Entity Linking (NEL)

NEL – is the task of determining unique identity to entities (people, locations, songs, etc.) mentioned in text.



Named Entity Linking (NEL)

NEL consists of several steps:

Named Entity Linking (NEL)

NEL consists of several steps:

1. Named Entity Recognition (NER) – spot text spans in that contain Named Entity label;

Named Entity Linking (NEL)

NEL consists of several steps:

1. Named Entity Recognition (NER) – spot text spans in that contain Named Entity label;
2. Named entity search – given the text form, find possible candidates in database;

Named Entity Linking (NEL)

NEL consists of several steps:

1. Named Entity Recognition (NER) – spot text spans in that contain Named Entity label;
2. Named entity search – given the text form, find possible candidates in database;
3. Candidates ranking – order candidates according to a parameter e.g. PageRank, Views per month.

Named Entity Linking (NEL)

NEL consists of several steps:

1. Named Entity Recognition (NER) – spot text spans in that contain Named Entity label;
2. Named entity search – given the text form, find possible candidates in database;
3. Candidates ranking – order candidates according to a parameter e.g. PageRank, Views per month.

Tools NER: spaCy, StanfordNLP etc.

Named Entity Linking (NEL)

NEL consists of several steps:

1. Named Entity Recognition (NER) – spot text spans in that contain Named Entity label;
2. Named entity search – given the text form, find possible candidates in database;
3. Candidates ranking – order candidates according to a parameter e.g. PageRank, Views per month.

Tools NER: spaCy, StanfordNLP etc.
Tools NEL: DBpedia Spotlight, TagMe, AGDISTIS, etc.

Any questions?

Exercise 1

Part 1 (manual) – Select any 10 questions from the dataset according to your variant and do the following:

1. Translate them from English to your mother tongue (e.g., German, Chinese, etc.); please do no use machine translation (if possible).
2. Extract named entities manually from these questions (and translations) and determine their types (e.g., Person, Politician, Entertainer, Location, City, Company, etc.);
3. Put everything together in the structured JSON format.

Exercise 1

Part 2 (programming) – Depending on your exercise variant write a script, which takes for input a list of questions and outputs for each question:

1. Preprocessed question (tokenization, stopwords and special characters removing, lemmatization);
2. A dictionary {'text': 'type'} of recognized named entities from the question (use spaCy).
3. A dictionary {'uri': 'text'} of linked named entities from the question (use DBpedia Spotlight).
4. The format of the script output is a JSON.

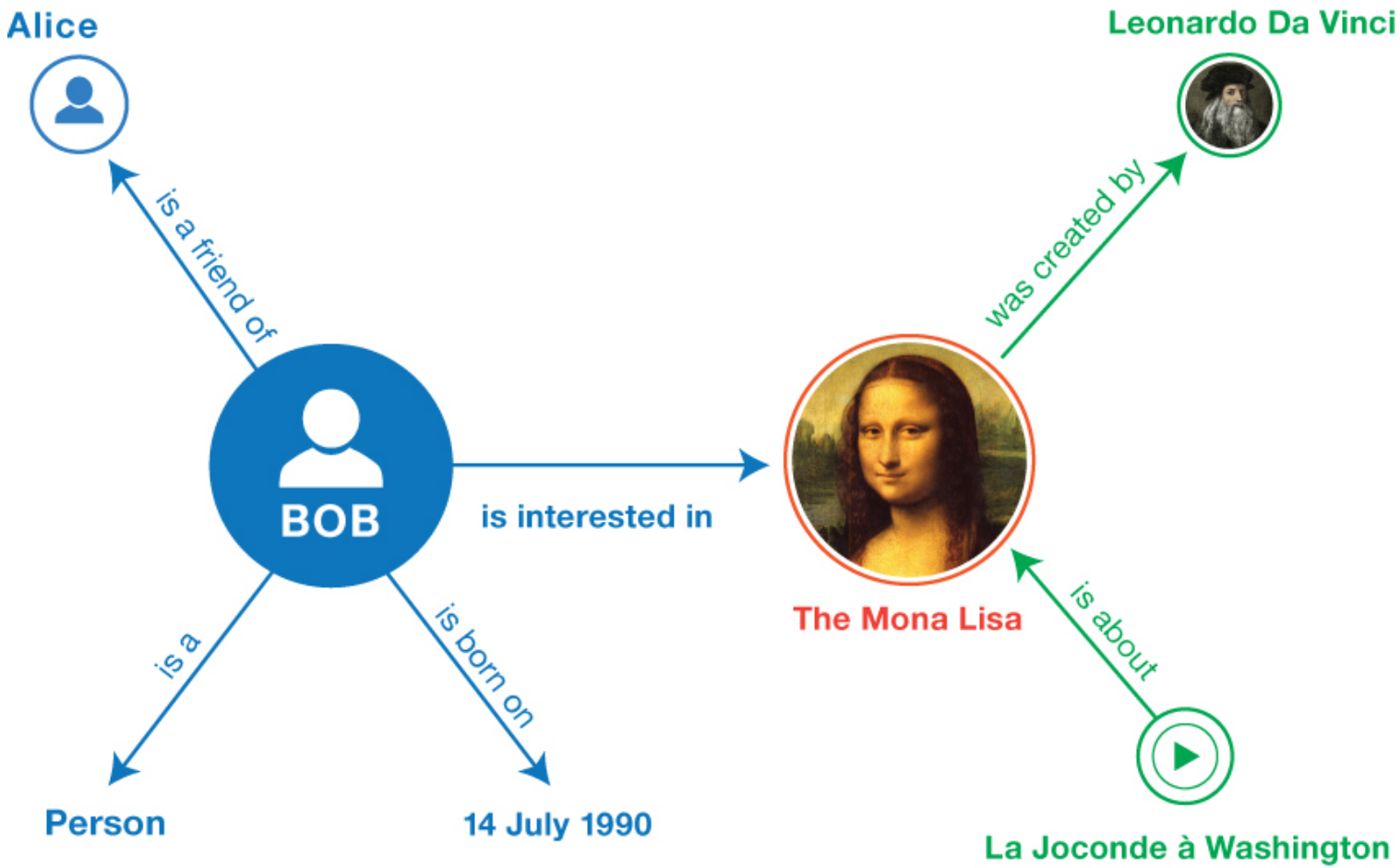
Let's do the exercise. Ask me, if you have a question.

Plan for the Exercise 2: Question classification

Task: create a relation classification algorithm for relation prediction you can use rule-based or machine learning approach.
Question: "Who created Mona Lisa?" $\rightarrow$ Relation: `dbo:author` (Author/Creator/Was created by).
Integrate implemented algorithm into a **Web Service** (RESTful API):

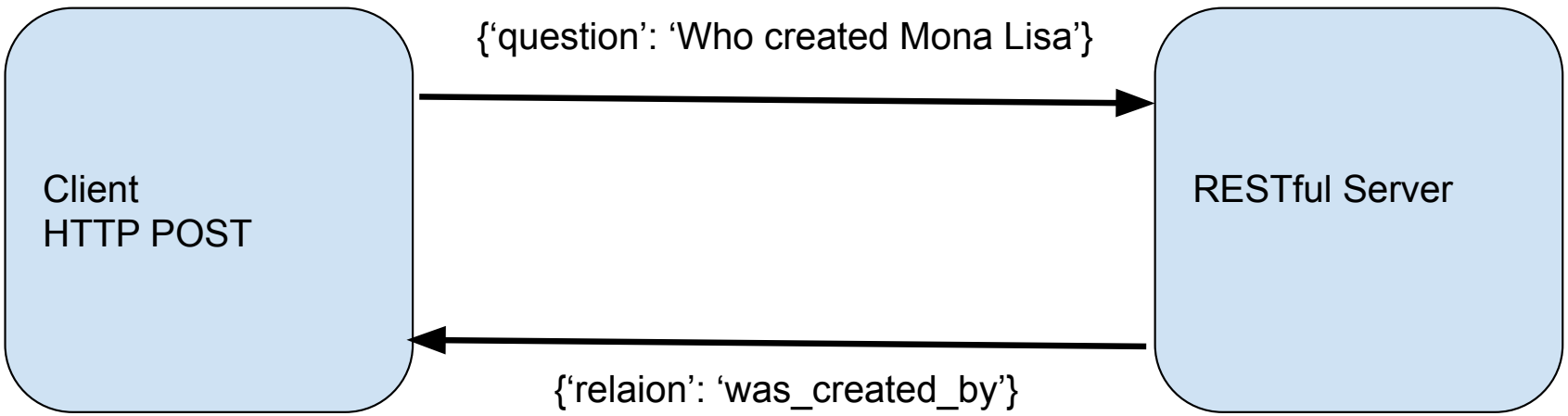
- Input: textual question;
- Output: prediction (result of classification).

Plan for the Exercise 2: Question classification



https://www.w3.org/TR/rdf11-primer/

Plan for the Exercise 2: Question classification



0 Introduction;

1 **NER & NEL**;

2 Question classification & Web service/API;

3 SPARQL queries over Knowledge Graphs;

4 Simple KGQA system – based on exercises 0, 1, 2, 3;

5 Qanary Framework – component oriented approach;

6 Simple ODQA system?;

7 Evaluation of QA systems.