

NLP: Information Extraction

Vladimir Makharev, Danil Andreev, Mikhail Fedorov

Innopolis University, Natural Language Processing, Spring 2023, Final project

Problem

Input:

Text of document from
procurement department and its
type

Output:

Fragment with contract fulfillment
guarantee or warranty obligations
guarantee

```
'text': 'УТВЕРЖДАЮ Генеральный директор АО «САБ по уборке г. Курска» «07» '
'сентября 2022 г. _____ А.Р. Зинатулин М.П. '
'ДОКУМЕНТАЦИЯ ОБ АУКЦИОНЕ В ЭЛЕКТРОННОЙ ФОРМЕ, УЧАСТНИКАМИ КОТОРОГО '
'МОГУТ БЫТЬ ТОЛЬКО СУБЪЕКТЫ МАЛОГО 5 апреля 2013 года N 44-ФЗ "О '
'контрактной системе в сфере закупок товаров, работ, услуг для '
'обеспечения государственных и муниципальных нужд". 54. Требования к '
'участникам такой закупки и привлекаемым ими субподрядчикам, '
'соисполнителям и (или) изготовителям товара, являющегося предметом '
'закупки, и перечень документов, представляемых участниками такой '
'закупки для подтверждения их соответствия указанным требованиям, в '
'случае закупки работ по проектированию, строительству, модернизации '
'и ремонту особо опасных, технически сложных объектов капитального '
'строительства и закупки товаров, работ, услуг, связанных с '
'использованием атомной энергии Не установлено 55. Размер обеспечения '
'исполнения договора, порядок предоставления такого обеспечения, '
'требования к такому обеспечению. Обеспечение гарантийных '
'обязательств В целях обеспечения исполнения обязательств по Договору '
'участник закупки предоставляет Заказчику обеспечение исполнения '
'Договора. Размер обеспечения исполнения договора не установлен. '
'Размер обеспечения гарантийных обязательств установлен в размере 20% '
'от НМЦД: 1 644 839,76 рублей. Гарантийные обязательства '
'обеспечиваются внесением денежных средств участником закупки на '
'указанный заказчиком счет, на котором в соответствии с '
'Федерации учитываются операции со '
'Заказчику. Договор заключается после '
'закупки, с которым заключается договор, '
'обеспечения, предоставляемые вместе с '
'договора. При непредоставлении обеспечения '
'договор не заключается. В ходе исполнения '
'договора подрядчик не вправе изменить способ обеспечения исполнения '
'гарантийных обязательств. Внесение денежных средств осуществляется '
'по следующим банковским реквизитам для внесения обеспечения '
'договора: Получатель: Акционерное общество «Спецавтобаза по уборке '
'города Курска», сокращенное наименование: АО «САБ по уборке г. '
'Курска» Название Банка: Курское отделение № 8596 ПАО Сбербанк БИК: '
'043807606 р/с: 40702810433000002123 к/с: 30101810300000000606 В '
'назначении платежа указать: обеспечение гарантийных обязательств по '
'закупке «Выполнение земляных 21'
```

Data Exploration



```
{  
  id: int - document id  
  text: str - document text  
  label: str - type of the guarantee  
  extracted_part:  
  {  
    'text': [target fragmet],  
    'answer_start': [first char idx],  
    'answer_end': [last char idx]  
  }  
}
```

- Can one document have multiple labels?
- Are there any distinct features in the extracted parts?
- How well do different models perform sentence segmentation on given documents?
- Is there any difference in different tokenizers?

Data Exploration

| | min | avg | max |
|-----------------------------------|-----|--------|------|
| Tokens per sample (natasha) | 140 | 383.92 | 1087 |
| Tokens per sample (spaCy) | 134 | 375.35 | 918 |
| Sentences per sample (natasha) | 1 | 8.20 | 25 |
| Sentences per sample (spaCy) | 1 | 10.34 | 29 |
| Tokens per extracted part (spaCy) | 5 | 16.84 | 71 |

Data Exploration



```
'text': 'УТВЕРЖДАЮ Генеральный директор АО «САБ по уборке г. Курска» «07» '
        'сентября 2022 г. _____ А.Р. Зинатулин М.П. '
        'ДОКУМЕНТАЦИЯ ОБ АУКЦИОНЕ В ЭЛЕКТРОННОЙ ФОРМЕ, УЧАСТНИКАМИ КОТОРОГО '
        'МОГУТ БЫТЬ ТОЛЬКО СУБЪЕКТЫ МАЛОГО 5 апреля 2013 года N 44-ФЗ "О '
        'контрактной системе в сфере закупок товаров, работ, услуг для '
        'обеспечения государственных и муниципальных нужд". 54. Требования к '
        'участникам такой закупки и привлекаемым ими субподрядчикам, '
        'соисполнителям и (или) изготовителям товара, являющегося предметом '
        'закупки, и перечень документов, представляемых участниками такой '
        'закупки для подтверждения их соответствия указанным требованиям, в '
        'случае закупки работ по проектированию, строительству, модернизации '
        'и ремонту особо опасных, технически сложных объектов капитального '
        'строительства и закупки товаров, работ, услуг, связанных с '
        'использованием атомной энергии Не установлено 55. Размер обеспечения '
        'исполнения договора, порядок предоставления такого обеспечения, '
        'требования к такому обеспечению. Обеспечение гарантийных '
```

обязательств В целях обеспечения исполнения обязательств по Договору
'участник закупки предоставляет Заказчику обеспечение исполнения '
'Договора. Размер обеспечения исполнения договора не установлен. '
'Размер обеспечения гарантийных обязательств установлен в размере 20% '
'от НМЦД: 1 644 839,76 рублей. Гyarантийные обязательства '
'обеспечиваются внесением денежных средств участником закупки на '
'указанный заказчиком счет, на котором в соответствии с '
'законодательством Российской Федерации учитываются операции со '
'средствами, поступающими заказчику. Договор заключается после '
'предоставления участником закупки, с которым заключается договор, '
'обеспечения гарантийных обязательств, предоставляемые вместе с '
'обеспечением исполнения договора. При непредоставлении обеспечения '
'гарантийных обязательств договор не заключается. В ходе исполнения '
'договора подрядчик не вправе изменить способ обеспечения исполнения '
'гарантийных обязательств. Внесение денежных средств осуществляется '
'по следующим банковским реквизитам для внесения обеспечения '
'договора: Получатель: Акционерное общество «Спецавтобаза по уборке '
'города Курска», сокращенное наименование: АО «САБ по уборке г. '
'Курска» Название Банка: Курское отделение № 8596 ПАО Сбербанк БИК: '
'043807606 р/с: 407028104330000002123 к/с: 30101810300000000606 В '
'назначении платежа указать: обеспечение гарантийных обязательств по '
'закупке «Выполнение земляных 21'

Data Exploration



```
'text': ['Размер обеспечения гарантийных обязательств '  
        'установлен в размере 20% от НМЦД: 1 644 839,76 '  
        'рублей. Гарантийные обязательства обеспечиваются '  
        'внесением денежных средств участником закупки']
```



```
'text': ['Поставщик предоставляет обеспечение исполнения '  
        'Договора в размере 5% от начальной (максимальной) '  
        'цены договора, что составляет _____ рублей, '  
        'в форме_____']
```


Data Preprocessing

Price placeholders & Trailing underscores

Text core extraction:

1. Sentence segmentation & Embeddings
2. Fixed length segments & Embeddings
 - a. Overlapping
 - b. Non-overlapping



Размер обеспечения исполнения Контракта
составляет 5 процентов цены Контракта в
размере _____ (_____) рублей ____



Document intro. Document intro.
Target. Target. Document outro.
Document outro.

Existing methods

NER

- The most common approach for token classification

Why not NER:

- Large entities are hard to interpret
- Large entities require more context
- Large entities are made up of multiple smaller entities

Spancat

- Variety of suggester functions
- New and shiny 🌟

Why not Spancat:

- Extracting long spans required high computational resources
- High number of suggested candidates lead to slow learning and poor performance

QA

- Fact-based QA perfectly suits for our problem

What can decrease performance:

- Poor annotations
- Unstructured documents
- Small amount of training samples

Model Comparison

| Model | QA | NER |
|---|---------------|--------|
| tok2vec | | 64.15% |
| cointegrated/rubert-tiny2 | 55.35% | 67.60% |
| M-CLIP/M-BERT-Distil-40 | 72.33% | |
| bert-base-multilingual-cased | | 73.33% |
| DeepPavlov/rubert-base-cased (baseline) | 75.80% | |
| distilbert-base-multilingual-cased | 78.62% | |
| ai-forever/sbert_large_mt_nlu_ru | 83.65% | |
| ai-forever/ruBert-large | 84.91% | |

What we have learned

- Information extraction task can be approached differently
- New architectures
 - Spancat
 - Distilled models
 - Different transformers
 - Tok2Vec
- spaCy v3 tools (ner, spancat learning pipelines via config)
- 🙌 HuggingFace tools (datasets, tokenizers, models)

Future work

- Improve Question Answering pipeline
 - Formulate a proper question instead of raw label
 - Try larger transformers
- Clean data
 - Revise annotations for typos
 - Restore punctuation in documents
- Apply deeper postprocessing

Conclusion

- Experiment a lot with NER, Spancat, and QA pipelines based on different transformers
- **Fact-based QA shows best performance** and can be improved further
- Better performance requires a lot of memory to store larger transformers
 - NVIDIA GeForce RTX 3080 12GB can be not enough
- Problems with dataset annotations should be handled wisely

Thank You!