Генерация сложных примеров для нормализации денежных сумм

Московский физико-технический институт

Кафедра компьютерной лингвистики

26.06.2023

Студент: Малхасян Арсен Жирайрович
Научный руководитель: Денис Афанасьев Евгеньевич

Постановка задачи

Целю работы является создание нейронной модели, которая будет детектировать сложные примеры и порождать нормализованную форму денежных сумм.

Набор данных:

- Датасет текстовых документов.
- список аннотации денежных сумм для каждого документа.
- Свойства аннотации

На выход:

- Выделение всех свойств для каждой аннотаций.

План

- Парсинг набора данных
- Named Entity Recognition
- Char BiLstm модель
- Добавление синтетических данных
- Промежуточные результаты
- Обновление эмбеддинг аннотации
- Добавление Attentions и CRF слой
- Добавление новых синтетических данных
- Сравнение результатов и заключение

Парсинг набора данных

Типы аннотации

- объекты с полными аннотациями

Например:

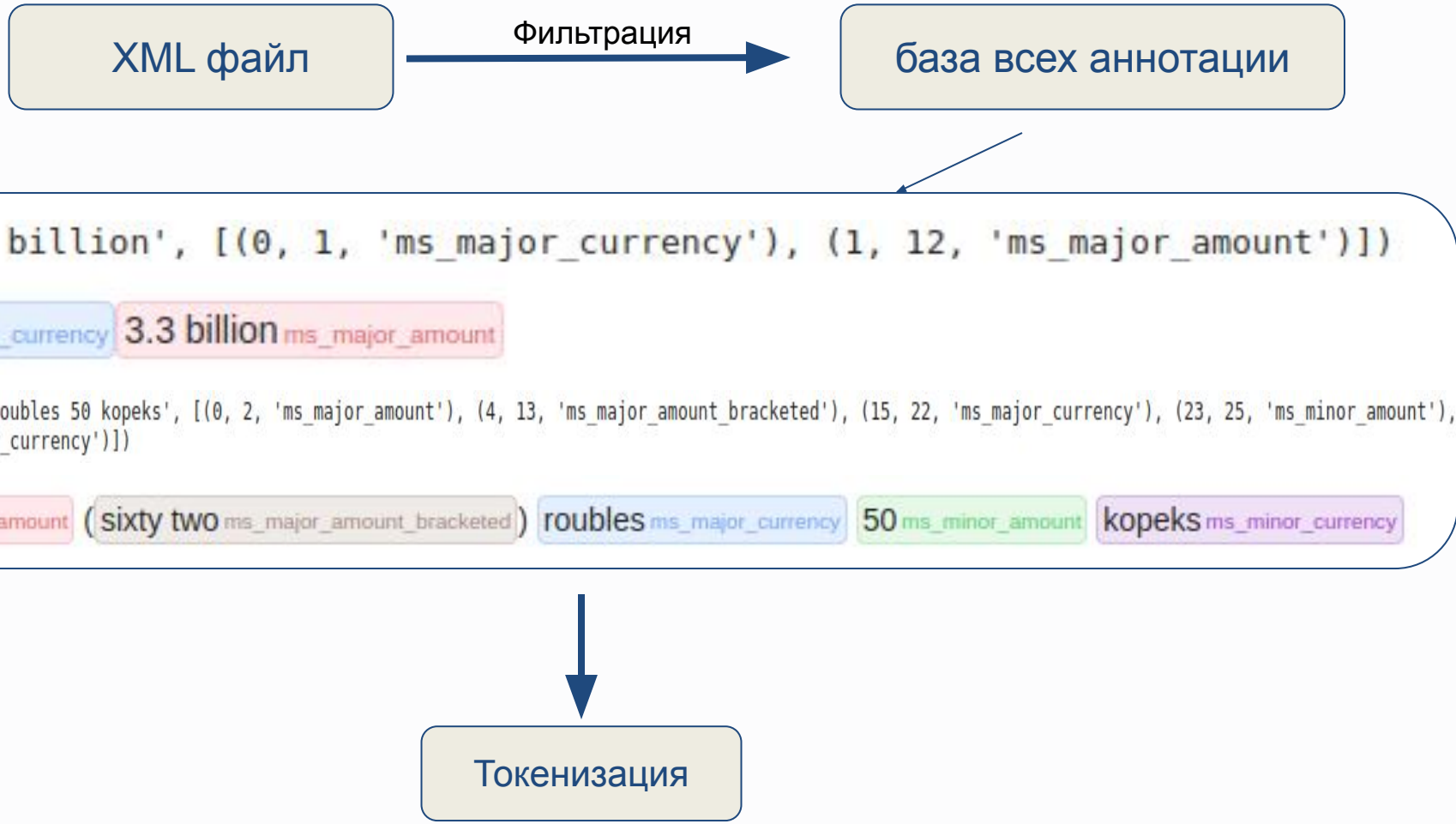
“1500 Colombian Pesos” → “1500” : amount
“Colombian Pesos” : currency

- объекты с разрывными аннотациями

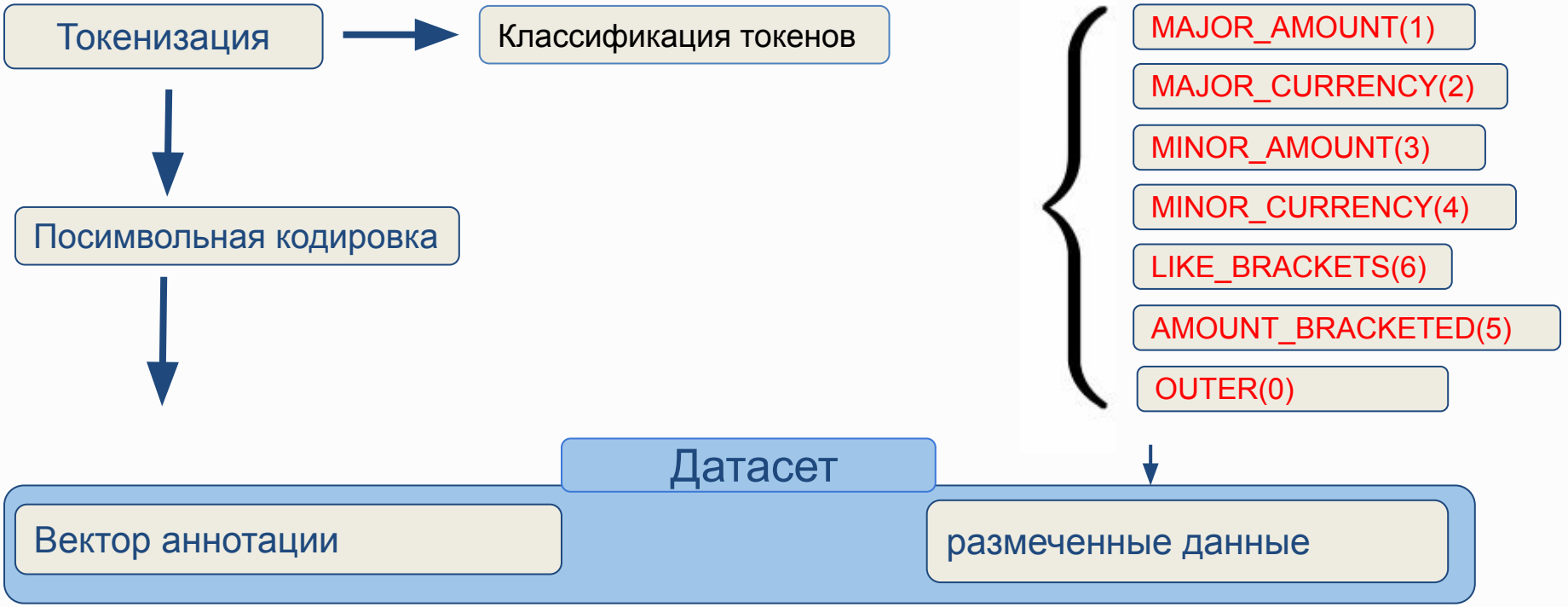
Например:

“50 to 80 Pesos” → “80 Pesos”: “80” : amount
“Pesos” : currency
“50 Pesos”: “50” : amount
“Pesos” : currency

Парсинг набора данных

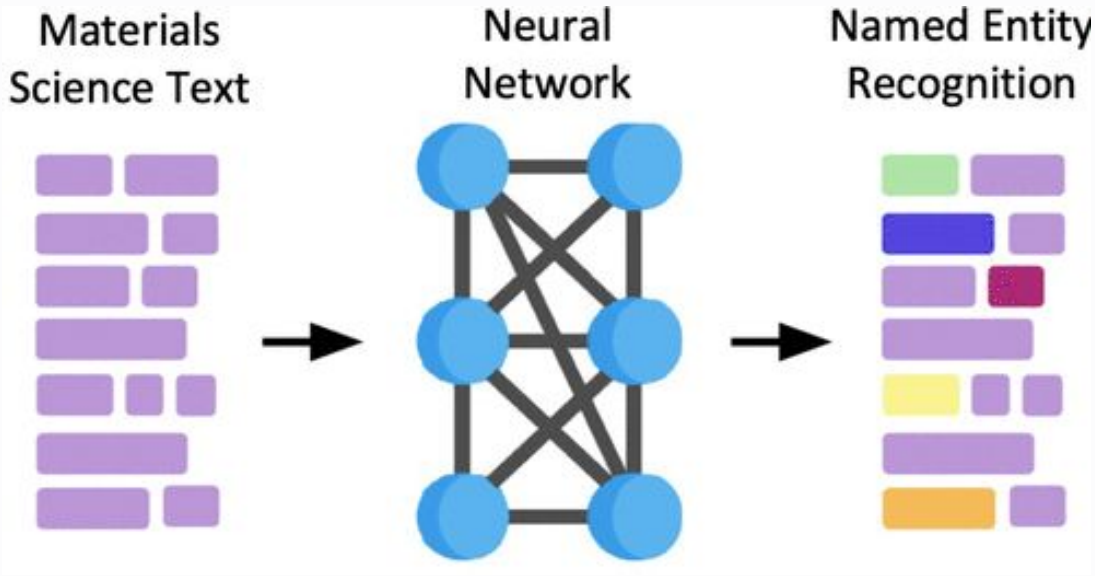


Парсинг набора данных

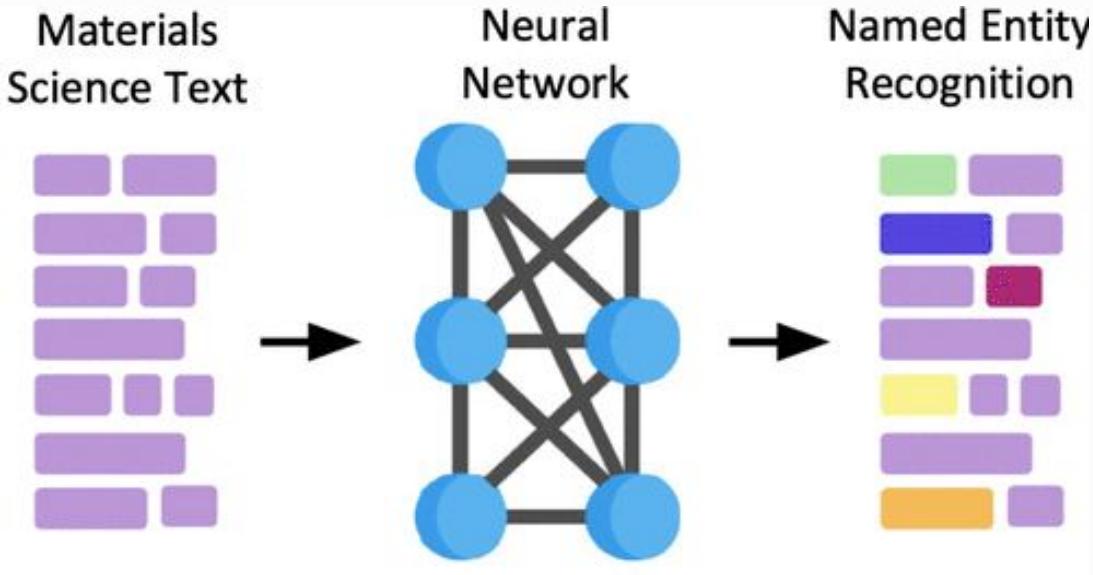


NER - Named Entity Recognition

Named Entity Recognition



Named Entity Recognition



- CNN (только для коротких текстов)
- BiLSTM
- BiLSTM - CRF
- BiLSTM - Attentions - CRF
- Transformer-Based NER(Bert, RoBERT, ALBERT)
- Neural cascade

Char BiLSTM модель

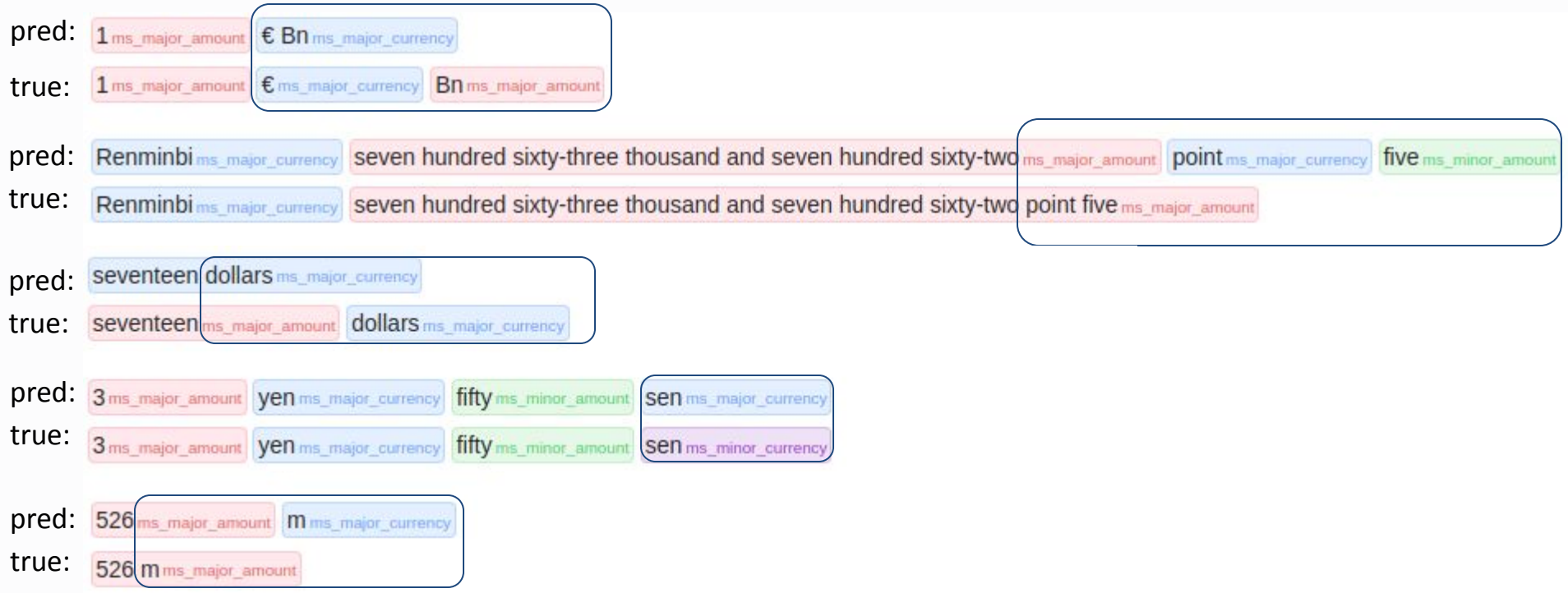
Network Layers

- Embedding - на словарь символов
- Dropout
- Linear
- RELU
- **BiLSTM**
- Linear

Classification Report

	precision	recall	f1-score	support
outer	1.00	1.00	1.00	85916
ms_major_amount	1.00	1.00	1.00	5407
ms_major_currency	1.00	1.00	1.00	5019
ms_minor_amount	0.79	0.79	0.79	29
ms_minor_currency	0.91	0.74	0.82	27
ms_major_amount_bracketed	1.00	1.00	1.00	15
outer_like_brackets	0.97	0.81	0.88	47
accuracy			1.00	96460
macro avg	0.95	0.91	0.93	96460
weighted avg	1.00	1.00	1.00	96460

Проблемы - ошибки



Синтетические данные

- Анализ ошибок модели
- генерация новых синтетических данных
  - на основе ошибках сетки
  - используя символьное представление любого числа (num2words)
  - учитывая порядок sub-аннотации
- добавить на обучение

```
ms_major_currency_syntetics = ['cents of euro', '¢', 'sen', 'manats', 'yen', 'Brunei dollars']
ms_major_amount_syntetics = ['lakh', 'a', 'A', '⅛', 'crore', 'Four', 'trillion', 'a single', 'Lacs']
ms_like_brackets_syntetics = ['(', ')', '’', 'and', ':', ';', 'of', '']
ms_minor_currency_syntetics = ['kopeks', 'd', 'shillings', 'kobo', 'quadrans']
```

Синтетические данные

- Анализ ошибок модели
- генерация новых синтетических данных
  - на основе ошибках сетки
  - используя символьное представление любого числа (num2words)
  - учитывая порядок sub-аннотации
- добавить на обучение

```
ms_major_currency_syntetics = ['cents of euro', '¢', 'sen', 'manats', 'yen', 'Brunei dollars']
ms_major_amount_syntetics = ['lakh', 'a', 'A', '⅛', 'crore', 'Four', 'trillion', 'a single', 'Lacs']
ms_like_brackets_syntetics = ['(', ')', '’', 'and', ':', ';', 'of', '']
ms_minor_currency_syntetics = ['kopeks', 'd', 'shillings', 'kobo', 'quadrans']
```

Данные до синтетики

train - 11500

val - 1000

test - 3700

Синтетические данные

- Анализ ошибок модели
- на этих ошибках генерировать новые синтетические данные
- добавить на обучение

Classification Report

	precision	recall	f1-score	support
outer	1.00	1.00	1.00	85916
ms_major_amount	1.00	1.00	1.00	5407
ms_major_currency	1.00	1.00	1.00	5019
ms_minor_amount	0.90	0.90	0.90	29
ms_minor_currency	0.96	0.85	0.90	27
ms_major_amount_bracketed	1.00	1.00	1.00	15
outer_like_brackets	0.93	0.87	0.90	47
accuracy			1.00	96460
macro avg	0.97	0.95	0.96	96460
weighted avg	1.00	1.00	1.00	96460

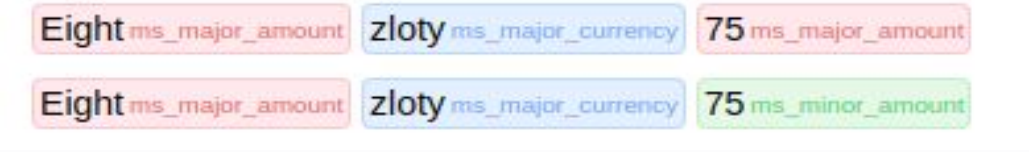
Промежуточные результаты

F1 - Score

Classes	Char BiLSTM	Char BiLSTM + synthetic data
ms.major_amount	1.00	1.00
ms_major_currency	1.00	1.00
ms_minor_amount	0.79	0.9
ms_minor_currency	0.82	0.9
ms_amount_bracke ted	0.9	1.00

Char BiLSTM

Char BiLSTM + synthetic data



Char BiLSTM

Char BiLSTM + synthetic data

Промежуточные результаты

F1 - Score

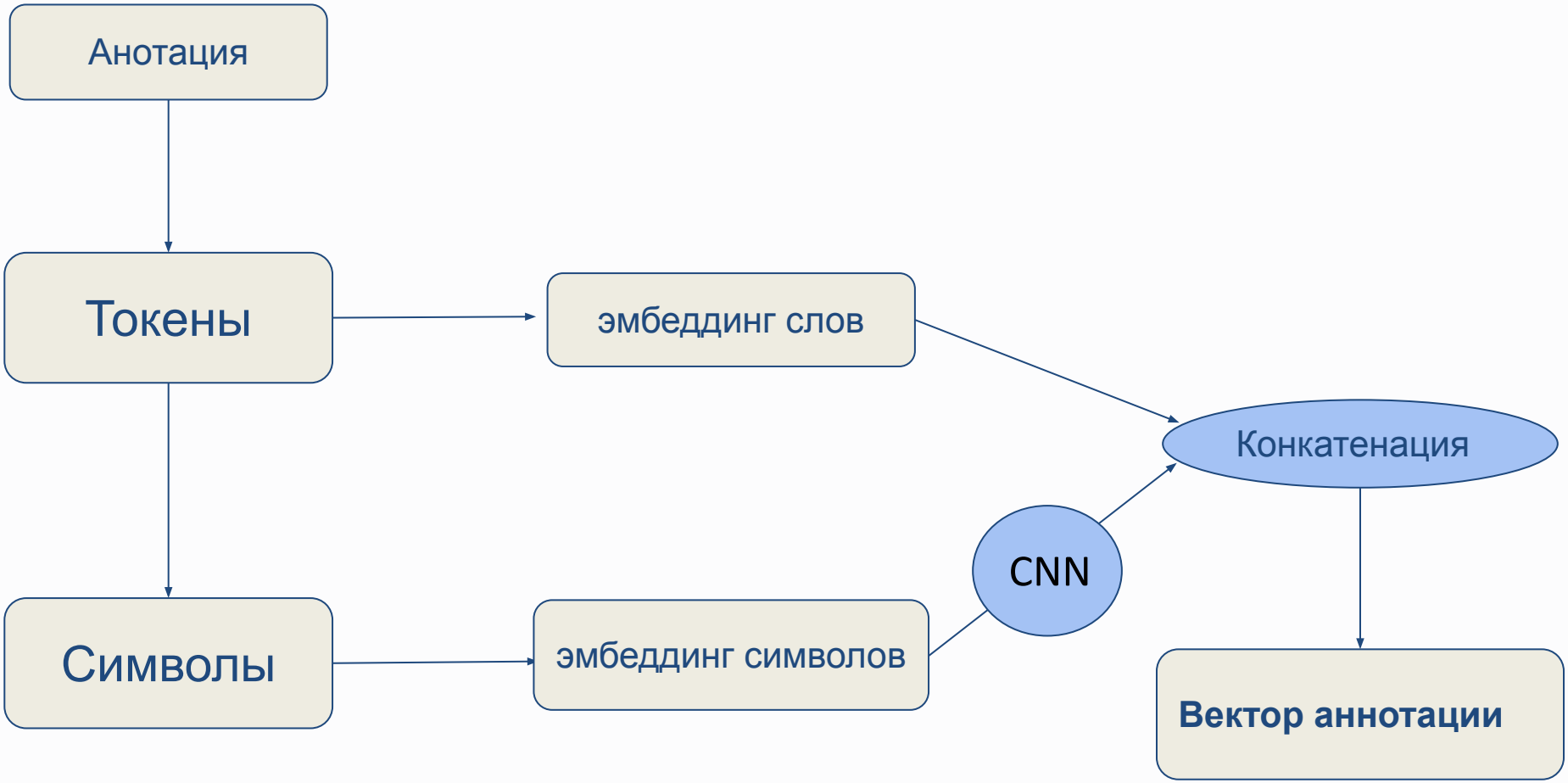
Classes	Char BiLSTM	Char BiLSTM + synthetic data
ms.major_amount	1.00	1.00
ms_major_currency	1.00	1.00
ms_minor_amount	0.79	0.9
ms_minor_currency	0.82	0.9
ms_amount_bracketed	0.9	1.00

macro average	0.9	0.96

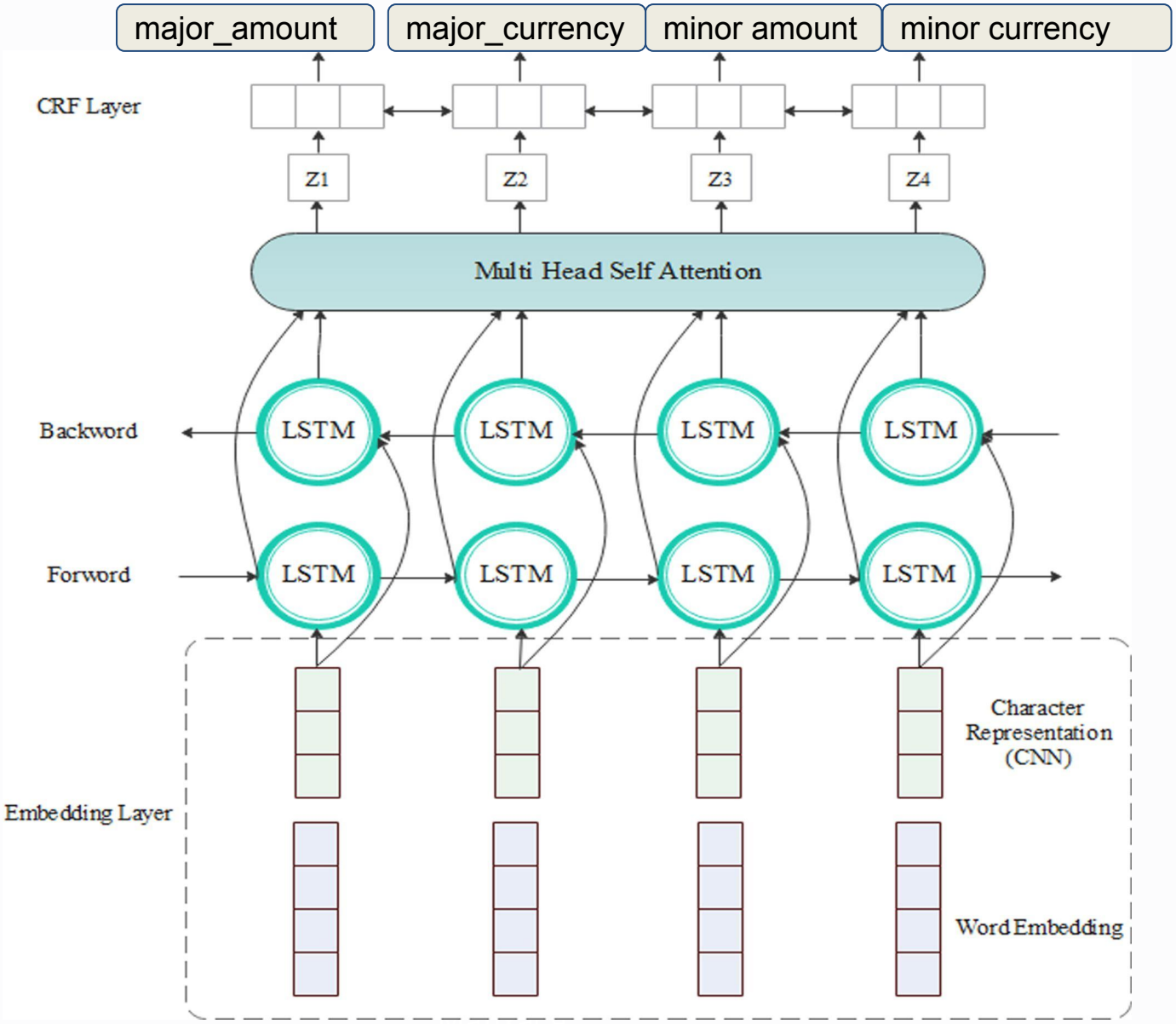
Заключение

- С добавлением синтетических данных сетка начинает предсказывать лучше, улучшая значение macro average из 0.9 до 0.96.

Обновление эмбеддинг аннотации



Добавление Attentions и CRF слой



Сравнение результатов и заключение

BiLSTM + ATTENTION + CRF

Classification Report

	precision	recall	f1-score	support
ms_major_amount	1.00	1.00	1.00	5404
ms_major_currency	1.00	1.00	1.00	5056
ms_minor_amount	0.95	0.91	0.93	22
ms_minor_currency	1.00	0.87	0.93	15
ms_major_amount_bracketed	1.00	1.00	1.00	15
accuracy			1.00	10512
macro avg	0.99	0.95	0.97	10512
weighted avg	1.00	1.00	1.00	10512

Сравнение результатов и заключение

Ошибки → Всего 15 ошибок из 3710 аннотации

Classes	FP	FN
ms.major_amount	3	4
ms_major_currency	3	4
ms_minor_amount	1	2
ms_minor_currency	0	2
ms_amount_bracketed	0	0

Char BiLSTM

Char BiLSTM + synthetic data

BiLSTM + Attention + CRF

Синтетические данные

- Анализ ошибок модели
- генерация новых синтетических данных
  - на основе ошибках сетки
  - используя символьное представление любого числа (num2words)
  - учитывая порядок sub-аннотации
- добавить на обучение

Данные до синтетики

train - 11500

val - 1000

test - 3700

Синтетические данные

- Анализ ошибок модели
- на этих ошибках генерировать новые синтетические данные
- добавить на обучение

	precision	recall	f1-score	support
ms_major_amount	1.00	1.00	1.00	5694
ms_major_currency	1.00	1.00	1.00	5226
ms_minor_amount	0.94	0.91	0.92	40
ms_minor_currency	1.00	0.92	0.96	27
ms_major_amount_bracketed	1.00	1.00	1.00	25
accuracy			1.00	11012
macro avg	0.99	0.96	0.97	11012
weighted avg	1.00	1.00	1.00	11012

Сравнение результатов и заключение

F1 - Score

Classes	Char BiLSTM	Char BiLSTM + synthetic data	BiLSTM + ATTENTION + CRF	BiLSTM + ATTENTION + CRF + synthetic data
ms.major_amount	1.00	1.00	1.00	1.00
ms_major_currency	1.00	1.00	1.00	1.00
ms_minor_amount	0.79	0.9	0.93	0.92
ms_minor_currency	0.82	0.9	0.93	0.96
ms_amount_bracketed	0.9	1.00	1.00	1.00
macro average	0.9	0.96	0.972	0.976

Заключение

- С добавлением синтетических F1-мера для класса "minor currency" повысилась с 0.93 до 0.96,. Для класса "minor amount"значение F-меры изменилось незначительно с 0.93 до 0.92. Однако, общее качество предсказаний модели улучшилось, macro avg с 0.972 до 0.976

Related papers:

- Named Entity Recognition with Bidirectional LSTM-CNNs (https://arxiv.org/abs/1511.08308)
- Named-entity recognition for Indonesian language using bidirectional LSTM- CNNs https://www.sciencedirect.com/science/article/pii/S088523080190169X
- Ma, Xuezhe. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF (https://aclanthology.org/P16-1101.pdf)
- Huang, Zhiheng. Bidirectional LSTM-CRF models for sequence tagging / Zhiheng Huang, Wei Xu, Kai Y (https://aclanthology.org/W19-3712/)
- Neural architectures for named entity recognition / Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian et al. (https://www.sciencedirect.com/science/article/pii/S1877050918314832?ref=pdf_download&fr=RR-2&rr=7c79057369629d43)