1. **On Biomedical Named Entity Recognition: Experiments in Interlingual Transfer for Clinical and Social Media Texts. Zulfat Miftahutdinov**

**Содержит множество ссылок на актуальные статьи**

Применение трансферного обучения для биомедицинских текстов. Трансфер происходит между языками и доменами(тематикой текстов). А именно – модель(Multilingual Bert), обученная на английских текстах(это high-resource язык – для него больше аннотированных датасетов для BioNER) используется для NER в русском языке.

3 основных вопроса:

- насколько качественно происходит трансфер из одного языка/домена в другие

Для русского языка EHR Multilingual Bert gave F 75%, LSTM-CRF – 62%

- при наличии малого набора данных, может ли модель NER получить эффективность больше, чем обученная на большом наборе данных

Да, предобученной MultilanguageBERT для того, чтобы достичь 99% точности в наборе обучения, необходимо 550 предложений входного датасета(EHR, Drug), когда модели без предобучения - 1500

- достигается ли увеличение эффективности при вариации данных обучения

Увеличивается стабильность девиаций F-метрики.

Модель:

BERT, MultiBERT+Softmax, (BIO tags, sentence level training), Adam optimizer + polynomial decay

Также, для сравнения: LSTM+CRF + Word2Vec[2]

**В результате сравнения типов моделей выяснилось, что MultilanBert+SoftMax выигрывает у LSTM+CRF**(подробности в статье).

1. **Size vs. Structure in Training Corpora for Word Embedding Models: Araneum Russicum Maximum and Russian National Corpus**

Word2Vec trained on the Russian National Corpus

1. **Towards reliable named entity recognition in the biomedical domain** <https://academic.oup.com/bioinformatics/article/36/1/280/5520946?login=true>

Улучшение генерализации BiLSTM-CRF при помощи трансферного обучения, вариационного дропаута и многозадачного обучения. Авторам удалось достичь наибольшего увеличения точности при многозадачном обучении и вариационном дропауте вместе.

Специфика текста – медицинский текст: хим. вещества, болезни, гены/белки. Корпус: CRAFT https://github.com/UCDenver-ccp/CRAFT и др.

1. **Извлечение информации Из клинических текстов на русском языке Шелманов А. О. 2015**

[**http://www.dialog-21.ru/digests/dialog2015/materials/pdf/ShelmanovAOetal.pdf**](http://www.dialog-21.ru/digests/dialog2015/materials/pdf/ShelmanovAOetal.pdf)

Проведено извлечение медицинских сущностей: лекарств, болезней(и связанных с ними объектов – тяжесть, лечение, часть тела и т.п.)

Методы:

- для поиска мед терминов: словари и правила

- для определения болезней и связанных объектов – SVM, деревья

Был сформирован корпус медицинских сущностей, который используется в [1]

Cвязанная презентация: <https://cs.hse.ru/data/2017/12/06/1161306556/Slides-IntroToAI-HSE-2017-09-Smirnov.pdf>

Связанный сайт http://nlp.isa.ru/index.php/component/portal/?view=projinfretrievalclinical

В статье много информации о задаче НЛП и испытаниях связанных с ней до 2015г.

Пайплайн:

* POS, токенизация, сепарация предложений - AOT.ru
* Синтаксические зависимости – MaltParser+ SynTagRus
* Собственно распознавание

1. **Cross-type Biomedical Named Entity Recognition with Deep Multi-Task Learning Xuan Wang**

<https://arxiv.org/pdf/1801.09851.pdf>

Улучшение точности BiLSTM-CRF при многозадачном обучении в BioNer

1. **Multitask Learning: A Knowledge-Based Source of Inductive Bias by Richard Caruana**

<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.57.3196>

многозадачное обучение

1. **A Survey on Deep Learning for Named Entity Recognition Jing Li**

<https://arxiv.org/pdf/1812.09449.pdf>

Описание текущей ситуации в NER. То есть, описываются ранее существовавшие методы извлечения(rule-based, ML) – их сильные и слабые стороны.

Далее, описываются методы глубокого обучения, их преимущства.

Описывается пайплайн – как организована модель и какие этапы в ее обучении присутсвуют. В организации пайплайна есть три основные стадии: векторное представление исходных данных(уровень слов, символов и гибридный), кодирование контекста и декодирование контекста. Для каждого этапа представлены описания популярных моделей DL. Сравнение эффективности этих методов.

Далее приводятся техники, успешно примененные в NLP – multitask learning, transfer learning, reinforcement learning, attention, adversarial learning.

В заключение, приводятся дальнейшие пути развития:

- применение NER в специфичных доменах слов(fine-grained)

- etc..

1. **Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text)** [**https://bmjopen.bmj.com/content/bmjopen/9/4/e023232.full.pdf**](https://bmjopen.bmj.com/content/bmjopen/9/4/e023232.full.pdf)

. Использованы словари правила(схема, готовый пайплайн GATE) для извлечения НДЭ: диагноза, тип приступа, фокальные припадки, частота припадков, лекарства

1. Основная статья BERT <https://www.aclweb.org/anthology/N19-1423/>
2. BioBERT: a pre-trained biomedical language representation model for biomedical text mining

BERT обученный на медицинских текстах – BioBERT – pre-trained+fine tuned models

1. Distributed Representations of Words and Phrases and their Compositionality. T. Mikolov and et al.

Основная статья Word2Vec

1. <https://rusvectores.org/ru/> - источник русскоязычных корпусов и моделей для NLP
2. **А.Сокирко "Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ)"** <http://aot.ru/docs/sokirko/sokirko-candid-1.html>