# Named Entity Recognition на лексических признаках с учётом всех вхождений упоминания в текст

Латыпов Зуфар
Московский Физико-Технический Институт

## Abstract

Статья по распознаванию именованных сущностей.

## 1 BSNLP

BSNLP (Balto-Slavic Natural Language Processing) - конференция по языконезависимой обработке естественного языка, в котором участники работают с славянскими и балтийскими языками - (Croatian, Czech, Polish, Russian, and Slovene, slovak, Ukrainian). В 2017 году общей задачей конференции стало выделение именованных сущностей, их нормализация и межязыковое связывание.

### 1.1 Описание корпуса

Организаторами были подготовлены два датасета, первый содержит документы, относящиеся к Дональду Трампу, текущему президенту США, а второй - документы, упоминающие Европейскую Комиссию. Документы для датасетов были созданы следующим образом: для каждой из тем были произведены поисковые запросы в Google на каждом из семи языков, результаты запроса были очищены от дубликатов и обработаны HTML парсером для извлечения текста (большинство ресурсов были новостями или фрагментами их). Полученный набор частично очищенных документов был использован для отбора 20-25 документов для каждого языка и темы для подготовки финального тестового датасета. Аннотации в основном были сделаны носителями языков, межязыковое связывание - носителями двух языков.

Организаторы не предоставляли данных для обучения алгоритмов, участники были вынуждены решать этот вопрос самостоятельно. Входные данные были представлены в следующем формате:

```
<DOCUMENT-ID>
<LANGUAGE>
<CREATION-DATE>
<URL>
<TITLE>
<TEXT>
```

В качестве выходных данных от участников ожидались документы с выделенными и нормализованными именованными сущностями, указанием их типов и межъязыковых идентификаторов:

```
<DOCUMENT-ID>
<MENTION> TAB <BASE> TAB <CAT> TAB <ID>

16
Podlascy Czeczeni Podlascy Czeczeni PER 1
ISIS ISIS ORG 2
Rosji Rosja LOC 3
Rosja Rosja LOC 3
Polsce Polska LOC 4
Warszawie Warszawa LOC 5
Magazynu Kuriera Porannego Magazyn Kuriera\
Porannego ORG 6
```

### 1.2 Итоги соревнования

В соревновании приняли участие более 11 команд, но только 2 из них сумели предоставить свое решение в поставленный организаторами срок. В связи с этим соревнование было продлено, и на данный момент сайт конференции содержит информацию о четырех различных системах: JHU, Liner2, LexiFlexi, Sharoff. Приведем краткое описание этих систем и достигнутых ими результатов.

#### 1.2.1 JHU

Авторы системы JHU приняли участие только в задачах по выделению и межъязыковому связыванию именованных сущностей. Для создания своей модели они проделали следующие шаги:

1. Получили из публично доступных датасетов параллельные тексты для языков соревнования и английского
2. Применили к текстам на английском уже готовую модель выделения именованных сущностей (Illinois Named Entity Tagger)
3. Спроецировали полученные результаты с английских текстов на целевые языки при помощи Giza++
4. На полученных датасетах обучили SVMLattice named entity recognizer
5. Использовали систему Kripke для межъязыкового связывания

Система показала хорошие результаты во всех трех задачах соревнования, заняв первые - вторые места практически во всех языках.

Таблица 1: F1 scores by type and language

| | PER | ORG | LOC | MISC |
|---|---|---|---|---|
| ces | 53.30 | 21.77 | 68.12 | 0.00 |
| hrv | 60.10 | 29.36 | 63.19 | 3.39 |
| pol | 35.29 | 13.19 | 68.73 | 0.00 |
| rus | 41.77 | 14.55 | 65.03 | 0.00 |
| slk | 57.52 | 18.67 | 63.20 | 2.94 |
| slv | 55.92 | 18.18 | 65.63 | 0.00 |
| ukr | 29.56 | 6.45 | 56.83 | 0.00 |
| all | 49.26 | 18.16 | 64.80 | 1.08 |

#### 1.2.2 Liner2

Авторы модели использовали фреймворк Liner2, который предоставляет набор модулей, основанных на статистических моделях, словарях, правилах и эвристиках и аннотирующих различные типа фраз. Команда работала только с польским языком, и сумела достичь лучших результатов в задачах выделения и нормализации именованных сущностей. Результаты работы системы можно увидеть ниже в таблице.

Таблица 2: Liner2 Results

| Task | P | R | F |
|---|---|---|---|
| Names matching | | | |
| Relaxed partial | 66.24 | 63.27 | 64.72 |
| Relaxed exact | 65.40 | 62.78 | 64.07 |
| Strict | 71.10 | 58.81 | 66.61 |
| Normalization | 75.50 | 44.44 | 55.95 |
| Coreference | | | |
| Document level | 7.90 | 42.71 | 12.01 |
| Language level | 3.70 | 8.00 | 5.05 |
| Cross-language level | n/a | n/a | n/a |

#### 1.2.3 LexiFlexi

К сожалению, авторы системы не предоставили статьи о своей системе, поэтому приходится довольствоваться ее кратким описанием. LexiFlexi применяет 3 лексико-семантических ресурса ко входному тексту в следующем порядке:

1. Сопоставляет имена из базы данных JRC Variant Names
2. Сопоставляет имена из огромной коллекции названий сущностей на различных языках полу-автоматически полученной из применения BabelNet к еще не обработанному тексту
3. Сопоставляет топонимы из газетира GeoNames в еще не обработанном тексте

В конце работы системы применяются несколько языко-независимых эвристик для нахождения вариантов (аббревиатур) именованных сущностей, распознанных на предыдущих шагах.

Данная модель показала средние результаты в задачах выделения и нормализации сущностей, однако заняла практически все первые места в задаче межъязыкового связывания на уровне документов, и половину первых мест - на уровне языка.

#### 1.2.4 Sharoff

Система Сергея Шарова - пример применения метода адаптации языка к задаче выделения именованных сущностей. Автор модели создал мультиязычное пространство embedding-ов для слов, основываясь на модели Dinu [] с добавлением взвешенного расстояния Левенштейна. Это пространство было использовано для обучения NER таггера, созданного при помощи нейронной сети, базируясь на Словенском NER корпусе. Простыми словами, главная идея - если мы можем адаптировать модель отзывов к фильмам для работы с

отзывами к отелями, то наверное мы можем адаптировать модель для NER одного языка к работе с родственными языками.

Модель добилась неплохих результатов в задачах выделения и нормализации именных сущностей, заняла первые места на чешском и словенском языках, заметны также и сильные падения качества на русском и украинском. На задаче межъязыкового связывания система проявила себя не очень хорошо, везде осталась на последних местах.

## 1.3 Выводы

На мой взгляд, на данной конференции были показаны две интересные системы - система Сергея Шарова и система JHU. Система Liner2 была применена только к польскому языку, а система LexiFlexi, если верить ее описанию, просто работает с большим набором газетиров. Результаты, достигнутые системой Шарова и системой JHU, показывают применимость методов адаптации языка и параллельной обработки текстов к задачам NER.

## Список литературы