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 FactRuEval

FactRuEval - соревнование по выделению именованных сущностей и извлечению фактов, проведенное на международной конференции по компьютерной лингвистике Диалог. Само соревнование включало в себя 3 дорожки: задачей первой было простое выделение именованных сущностей, определение их типов (персоны, организации и локации, другие не рассматривались) и указание позиции и длины сущности в тексте; для решения второй дорожки нужно было связать все упоминания одной и той же сущности в рамках текста в один объект и определить атрибуты этого объекта; третья задача затрагивала вопрос извлечения фактов из текста, то есть отношений между несколькими объектами.

1.1 Описание корпуса

Корпус текстов соревнования состоит из новостных и аналитических текстов на общественно-политическую тему на русском языке. Источниками текстов являются следующие издания: Частный корреспондент, Викиновости, Лентапедия. Корпус разделён на две части: демонстрационную и тестовую. Соотношение количества текстов из разных источников в этих двух частях одинаково. Сбалансированность по каким бы то ни было другим показателям не гарантируется. Работы по разметке этой коллекции текстов были проведены силами добровольцев на сайте OpenCorpora.org под руководством экспертов в областях.

Авторами соревнования была разработана специальная аннотационная модель, которая была использована для аннотации 255 документов (122 текста в обучающей выборке и 133 текста в тестовой). Первые два слоя модели содержат аннотированые упоминания сущностей, третий слой содержит информацию об отношениях кореференции между сущностями, четвертый же слой группирует сущности в факты. Для первой дорожки использовались первые два слоя, для второй - три слоя, для третьей - все четыре слоя.

Таблица 1: Размеры датасета

Total texts		Total characters	
Demo Set	Test Set	Demo Set	TestSet
122	133	189893	460636
Total tokens		Total sentences	
Demo Set	Test Set	Demo Set	Test Set
30940	59382	1769	3138

Описание слоев: (Более полное описание может быть найдено в репозитории соревнования)

Нулевой слой - это сами токены, без обработки.

```
143783 0 1 В
143784 2 11 понедельник
143785 14 2 28
143786 17 4 июня
143787 22 1 у
143788 24 6 здания
143789 31 5 мэрии
143790 37 6 Москвы
143791 44 2 на
143792 47 8 Тверской
143793 56 7 площади
143794 64 10 состоялась
143795 75 9 очередная
143796 85 19 несанкционированная
```

В первом слое в тексте были выделены типизированные спаны. Это цепочки слов, помеченные одним или более предопределенными тегами. Предполагалось также, что каждый тип выделенного объекта может иметь свой набор тегов, например, в случае упоминания людей различали имена, фамилии, ники.

```
22763 loc_name 37 6 143790 1  # 143790 Москвы
22764 org_descr 31 5 143789 1  # 143789 мэрии
22765 loc_name 47 8 143792 1  # 143792 Тверской
22766 loc_descr 56 7 143793 1  # 143793 площади
22767 name 313 4 143831 1  # 143831 Юрия
22768 surname 318 7 143832 1  # 143832 Лужкова
```

Во втором слое спаны сгруппированы в типизированные упоминания объектов.

```
10433 Org 22763 22764 # Москвы мэрии
10547 LocOrg 22763 # Москвы
10434 Location 22765 22766 # Тверской площади
10435 Person 22767 22768 # Юрия Лужкова
```

В третьем слое упоминания объектов, содержащиеся в одном тексте и имеющие одного референта, сгруппированы в идентифицированные объекты.

```
47 10436 10437 10547
name Москва

48 10435 10441
firstname Юрий
lastname Лужков

49 10433
descriptor мэрия
name мэрия Москвы

50 10434
descriptor площадь
name Тверская площадь

51 10438
name Россия
```

В четвертом слое были выделены факты - типизированные отношения между идентифицированными объектами.

```
100-0 Occupation
Who obj48 Лужков Юрий
Position span22777 мэра
Where obj47 Москва

100-1 Occupation
Who obj168 Громов Борис
Position span22778 губернатора
Where obj637 Подмосковье
```

1.2 Итоги дорожки

Большинство систем приняло участие в двух первых дорожках, в решении второй дорожки приняли участие всего 2 команды. Авторы соревнования связывают этот факт с чрезвычайной сложностью и неочевидностью принципа решения проблемы выделения фактов. Статьи были предоставлены только тремя командами, поэтому рассмотрим их.

1.2.1 Named Entity Recognition in Russian: the Power of Wiki-Based Approach

Участники команды использовали два различных подхода - на основании только FactRuEval данных и на основании Wiki данных.

Первый подход - использование широкого набора фичей: аффиксы, сам токен, POS-тег, лемма, предикаты, флаги, характеризующие тот факт, что слово начинается с большой буквы, и другие + Word2Vec из Wiki + словари, построенные на основе Wiki.

Второй подход - конструируют датасет на основе статей Wiki [1], и используют его вместо предоставленного. Результаты команды можно увидеть в таблицах ниже.

Таблица 2: Entity Extraction (I)

Feature set	Precision	Recall	F1
basic	0.7357	0.6186	0.6720
basic+dict	0.8098	0.6988	0.7502
basic+word2vec	0.8093	0.7241	0.7643
basic+dict+word2vec	0.8257	0.7408	0.7810

Таблица 3: Entity Extraction (по типам)

Entity type	Precision	Recall	F-measure
Person	0.9340	0.8675	0.8995
Location	0.7259	0.6944	0.7098
Organization	0.7844	0.6548	0.7137
LocOrg	0.7858	0.7251	0.7542
OVERALL	0.8257	0.7408	0.7810

Таблица 4: Entity Extraction (II)

FactRuEval devset			FactRuEval testset		
Precision	Recall	F1	Precision	Recall	F1
0.88	0.64	0.74	0.85	0.69	0.76

1.2.2 Named Entity Normalization for Fact Extraction Task

Использовали rule-based подход, создали свою систему обработки текста, состоящую из токенизатора, морфологического анализатора, газетира, искателя паттернов и извлекателя фактов.

Модуль, отвечающий за поиск паттернов - основной инструмент для извлечения сущностей, спроектирован для выполнения правил, задаваемых в стиле регулярных выражений.

К сожалению, авторы статьи не указали название своей команды, поэтому приведем результаты, указанные в их статье (предварительных):

Таблица 5: Entity Extraction

Entity type	Precision	Recall	F-measure
Persons	0.9300	0.8403	0.8829
Locations	0.9535	0.8361	0.8910
Organizations	0.8181	0.5450	0.6542
OVERALL	0.9038	0.7301	0.8077

Таблица 6: Entity Normalization

Entity type	Precision	Recall	F-measure
Persons	0.8024	0.8433	0.8223
Locations	0.9017	0.7741	0.8330
Organizations	0.6490	0.5760	0.6103
OVERALL	0.7725	0.7173	0.7439

1.2.3 Information Extraction Based on Deep Syntactic-Semantic Analysis

Команда использовала уже имевшуюся у нее модель, основанную на синтаксическо-семантическом анализе, rule-based подход. Для более подробной информации авторы статьи отсылают читаталей к статьям Анисимовича и Зуева 2012 и 2013 годов с конференции Диалог. Результаты команды представлены в таблицах ниже:

Таблица 7: Entity Extraction

Entity type	Precision	Recall	F-measure
Persons	0.9450	0.9155	0.9300
Locations	0.9261	0.8698	0.8971
Organizations	0.8175	0.7564	0.7858
OVERALL	0.8931	0.8427	0.8672

Таблица 8: Entity Normalization

Entity type	Precision	Recall	F-measure
Persons	0.8817	0.8592	0.8703
Locations	0.8430	0.7942	0.8179
Organizations	0.6823	0.6763	0.6793
OVERALL	0.7903	0.7677	0.7789

Список литературы

[1] Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. Learning multilingual named entity recognition from wikipedia. Artificial Intelligence 194, Supplement C (2013), 151 – 175. Artificial Intelligence, Wikipedia and Semi-Structured Resources.