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 CoNLL 2003

CoNLL 2003 (Conference on Computational Natural Language Learning) - конференция по машинной обработке естественного языка, прошедшая в Канаде в 2003 году. Общей задачей конференции было решение проблемы NER, Распознавания Именованных Сущностей, для двух языков - немецкого и английского. Для измерения точности использовались метрики точности (precision), полноты (recall) и F-мера (F-measure), участие приняли 16 различных систем, наилучшим результом стали 88.76 для английского и 72.41 для немецкого от системы FIJZ03 (здесь и в дальнейшем результаты указаны по метрике F1, если не указано обратное, кроме того, данные результаты вычислялись путем усреднения качества по всем типам сущностей). Ниже в таблицах 1 и 2 приведем также качество в разбивку по точности/полноте/F-мере (5 наилучших результатов для каждого языка):

Таблица 1: Первый датасет

English	Precision	Recall	F-measure
FIJZ03	88.99%	88.54%	88.76±0.7
CN03	88.12%	88.51%	88.31±0.7
KSNM03	85.93%	86.21%	86.07±0.8
ZJ03	86.13%	84.88%	85.50±0.9
CMP03b	84.05%	85.96%	85.00±0.8
baseline	71.91%	50.90%	59.61±1.2

Таблица 2: Второй датасет

German	Precision	Recall	F-measure
FIJZ03	83.87%	63.71%	72.41±1.3
KSNM03	80.38%	65.04%	71.90±1.2
ZJ03	82.00%	63.03%	71.27±1.5
MMP03	75.97%	64.82%	69.96±1.4
CMP03b	75.47%	63.82%	69.15±1.3
baseline	31.86%	28.89%	30.30±1.3

1.1 Описание корпуса

Датасет состоит из 6 файлов - это файлы testa, testb, train для каждого из языков. testa использовался для проверки модели при разработке, testb - для итоговой оценки модели. Файлы содержат 4 столбца, разделенные пробелами. Первый элемент каждой строки - это само слово, второй - POS (part-of-speach) tag, третий - syntactic chunk tag, четвертый - named entity tag. Также интересной особенностью 2003 года стали предоставленные списки именованных сущностей и неразмеченные данные, которые предлагалось как-то использовать для улучшения системы. Английский корпус был представлен коллекцией новостных статей из Reuters Corpus. Аннотация была произведена в University of Antwerp. Немецкий корпус - коллекция статей от Frankfurter Rundschau. Пример содержимого файла train, а также сводная таблица (3) по размерам датасетов приведены ниже.

```
WORD        POS  CHUNK NE
U.N.        NNP  I-NP  I-ORG
official    NN   I-NP  O
Ekeus       NNP  I-NP  I-PER
heads       VBZ  I-VP  O
for         IN   I-PP  O
Baghdad     NNP  I-NP  I-LOC
.           .    O     O
```

Таблица 3: Размеры датасетов

	Learning	Validating	Testing
Articles	946	216	231
Sentences	14987	3466	3684
Tokens	203621	51362	46435
LOC	7140	1837	1668
MISC	3438	922	702
ORG	6321	1341	1661
PER	6600	1842	1617

1.2 Итоги дорожки

Большинство систем на английском языке показали результаты в районе 88 - 80 при baseline в 71. На немецком языке системы проявили себя хуже - максимум 72, большинство работ в районе 60-70, однако и baseline тут значительно ниже - 30. Рассмотрим трех участников, показавших лучшие результаты для английского языка:

1.2.1 FIJZ03 - 88.76

Первое место заняла модель команды FIJZ03, достигшая результата в 88.76 [3]. Авторы модели использовали комбинацию четырех различных классификаторов - линейный классификатор, максимальной энтропии, основанное на трансформации обучение и скрытую Марковскую модель. Без газетиров и других дополнительных ресурсов они достигли результата в 91.6 на тренировочных данных, с использованием дополнительных данных сумели получить дополнительное уменьшение ошибки на 15 - 20 процентов. Также авторы отмечают, что устойчивый классификатор минимизации риска "выглядит особенно подходящим для обработки дополнительных источников признаков, и потому является хорошим кандидатом для комбинации классификаторов". Результаты работы данной модели приведены в таблицах 4 и 5.

Список рассматриваемых признаков:

- слова и их леммы в окне размеров в пять слов около текущего
- POS тег текущего и окружающего слов
- текстовые чанки в окне -1..1
- префиксы и суффиксы длины до 4 букв текущего и окружающего слов
- флаги, отражающие наличие заглавных букв (firstCap, 2digit and allCaps)
- информация из газетира
- результат работы двух других классификаторов, натренированных на более богатом датасете с большим числом категория

Таблица 4: FIJZ03 English Test

English test	Precision	Recall	F1
LOC	90.59%	91.73%	91.15
MISC	83.46%	77.64%	80.44
ORG	85.93%	83.44%	84.67
PER	92.49%	95.24%	93.85
overall	88.99%	88.54%	88.76

Таблица 5: FIJZ03 German Test

German test	Precision	Recall	F1
LOC	80.19%	71.59%	75.65
MISC	77.87%	41.49%	54.14
ORG	79.43%	54.46%	64.62
PER	91.93%	75.31%	82.80
overall	83.87%	63.71%	72.41

1.2.2 CN03 - 88.31

Авторы модели использовали подход, основанный на принципе максимума энтропии, причем использовали в качестве признаков не только локальный контекст, но также использовали и остальные вхождения этого слова для извлечения полезных признаков (т.н. глобальные признаки) [1]. Для этого они обработали датасет и создали несколько списков слов - Frequent Word List, Useful Unigrams, Useful Bigrams, Useful Word Suffixes, Useful Name Class Suffixes, Function Words, которые в дальнейшем использовались для выделения глобальных признаков.

К сожалению, в предоставленной авторами статье нет никакой информации по отбору признаков, только их перечисление, а название Useful, например, Useful Unigrams, говорит о том, что лист содержит 1-граммы, которые часто предшествуют определенному типу сущностей, а потому могут быть полезны при классификации. Однако в статье довольно подробно описаны листы и получаемые из них признаки, так что она может послужить основой для дальнейшего изучения возможностей по использованию глобальных признаков. Результаты работы данной модели приведены в таблицах 6 и 7.

Таблица 6: CN03 English Test

English test	Precision	Recall	F1
LOC	90.88%	91.37%	91.12
MISC	80.15%	78.21%	79.16
ORG	83.82%	84.83%	84.32
PER	93.07%	93.82%	93.44
Overall	88.12%	88.51%	88.31

Таблица 7: CN03 German Test

German test	Precision	Recall	F1
LOC	69.23%	59.13%	63.78
MISC	62.05%	33.43%	43.45
ORG	76.70%	48.12%	59.14
PER	88.82%	75.15%	81.41
Overall	76.83%	57.34%	65.67

1.2.3 KSNM03 - 86.07

Авторы рассматривают две модели - скрытую марковскую модель и conditional markov model, рассматривая в качестве базовых единиц не слова, а символы и n-граммы [4]. При разработке первой модели использование контекста было минимально, а при разработке второй использовался подход максимальной энтропии, после чего добавили дополнительные признаки и объединили модели в CMM. Результаты работы данной модели приведены в таблицах 8 и 9.

Таблица 8: KSNM03 English Test

English test	Precision	Recall	F1
LOC	90.04	89.93	89.98
MISC	83.49	77.07	80.15
ORG	82.49	78.57	80.48
PER	86.66	95.18	90.72
Overall	86.12	86.49	86.31

1.2.4 Итоги

Подводя итоги, отметим, что в 2003 году часто использовались и хорошо себя проявили такие классификаторы, как HMM и максимальной энтропии. Кроме того, многие авторы отмечали тот факт, что категория MISC довольно сильно повлияла на снижение качества работы моделей, связывая это с обобщенностью данной категории.

Таблица 9: KSNM03 German Test

German test	Precision	Recall	F1
LOC	78.01	69.57	73.54
MISC	75.90	47.01	58.06
ORG	73.26	51.75	60.65
PER	87.68	79.83	83.57
Overall	80.38	65.04	71.90

1.3 CRF и современные работы

1.3.1 CRF

Рассмотрим статью Andrew McCallum and Wei Li, в которой они обращаются к CRF, индуцированию признаков и методу WebListing для создания лексиконов [6]. Их система показала неплохой результат в 84.04 (F-мера), что доказывает применимость CRF в задачах выделения именованных сущностей.

Таблица 10: CRF English Test

English test	Precision	Recall	F1
LOC	87.23%	87.65%	87.44
MISC	74.44%	71.37%	72.87
ORG	79.52%	78.33%	78.92
PER	91.05%	89.98%	90.51
Overall	84.52%	83.55%	84.04

Таблица 11: CRF German Test

German test	Precision	Recall	F1
LOC	71.92%	69.28%	70.57
MISC	69.59%	42.69%	52.91
ORG	63.85%	48.90%	55.38
PER	90.04%	74.14%	81.32
Overall	75.97%	61.72%	68.11

1.3.2 Современные работы

В последние годы появилось довольно много статей, рассматривающих использование LSTM-CNNs, LSTM-CRF, LSTM-CNNs-CRF для датасета CoNLL2003. На данный момент один из наилучших результатов (State Of Art) был достигнут в 2016 году Xuezhe Ma и Eduard Hovy, используя BLSTM-CNNs-CRF, они смогли добиться результата в 91.21 (F-мера) без использования сторонних данных [5]. В 2015 году была достигнута планка в 91.62 при помощи LSTM-CNNs и использовании двух наборов данных, полученных из публично доступных источников,

авторы второй статьи также называют свой результат наилучшим [2].

Схема BLSTM-CNNs-CRF:

1. Используя CNN, извлекают морфологическую информацию, кодируют ее в символьное представление.
2. Отправляют результат первого шага в BLSTM (отмечается важность dropout слоя).
3. Результат работы BLSTM отправляется CRF.

Результатов для каждого из типов сущностей авторы не предоставили.

2 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.

2.1 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 12: 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
```

2.2 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.

2.2.1 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 [7], и используют его вместо предоставленного. Результаты команды можно увидеть в таблицах ниже.

Таблица 13: 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 14: 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 15: 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

2.2.2 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 16: 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 17: 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

2.2.3 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 18: 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 19: 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

3 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.

3.1 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
```

3.2 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.

3.2.1 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 20: 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

3.2.2 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 21: 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

3.2.3 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.

3.2.4 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.

3.3 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.

4 Настройка окружения, пакет NLTK, корпуса

Исследовательская работа проводилась в ОС Windows 10, использовалась среда для научных исследований Anaconda, Python3. Дополнительно установил пакет pymorphy2 для POS-тегирования токенов из датасетов на русском языке, для английского датасета (CoNLL2003) использовались предоставленные организаторами теги (chunk и POS). Для удобной работы с датасетами они были приведены к модифицированному формату датасетов CoNLL (были добавлены столбцы OFFSET и LEN - отступ токена и его длина), в коде они хранятся в виде модифицированных NLTK Corpus-ов. (код обработки можно найти в файлах .ipynb, а код NLTK Corpus-а - в файле corpus.py)

5 Выделение признаков

В первоначальной итерации было решено рассмотреть следующие признаки:

1. Часть речи (POS-tag, для CoNLL2003 датасета - и chunk-tag)
2. Капитализация (normal-case, Proper-case, CAPITAL-case, Camel-case)
3. Флаг, является ли слово числом
4. Флаг, является ли слово знаком пунктуации
5. Начальная форма слова

Для обработки датасетов и генерации признаков был написан отдельный класс Generator (файл generator.py), который на основании входных данных создает матрицу признаков, эта матрица обрабатывается OneHotEncoder-ом из пакета sklearn, опционально сохраняется в файл и возвращается в вызывающий код. Матрица признаков получается очень разреженной, поскольку ее размер напрямую зависит от размера словаря датасета (так, для датасета FactRuEval ее размеры достигают 35.000 * 27.000)

6 Отбор признаков

Теперь, когда у нас есть матрица признаков, можно заняться их отбором. Для начала, исключаем самые малоинформативные признаки - те, которые встретились в датасете менее 5 раз. Далее сортируем признаки по весам, присвоенным им классификатором, и отберем признаки, составляющие 90 процентов веса. Отбор признаков происходит при их генерации, код находится в классе Generator. После отбора признаков их число резко снижается - с 27 тысяч до порядка 700 (датасет FactRuEval) в случае 90 процентов веса. Приведем графики качества на тестовой и обучающей выборке в зависимости от оставляемого процента признаков.

7 Baselines

Попробуем обучить основные классификаторы на полученных данных. Рассмотрим такие классификаторы, как GradientBoostingClassifier, RandomForestClassifier, LogisticRegression, LinearSVC, без подбора параметров. Приведем

результаты нестрогой оценки классификации для всех датасетов (в случае BSNLP воспользуемся данными для обучения FactRuEval и приведем предсказанные имена классов к именам BSNLP по правилам LocOrg ⇒ LOC, Other ⇒ MISC):

Таблица 22: CoNLL2003 Results

	ORG	LOC	MISC	PER	Total
LogReg	0.760	0.820	0.787	0.888	0.826
RF	0.676	0.725	0.682	0.849	0.752
LinSVC	0.777	0.823	0.799	0.887	0.832
GB	0.677	0.729	0.748	0.834	0.758

Таблица 23: FactRuEval Results

	Per	Loc	Org	LO	Total
LogReg	0.802	0.551	0.460	0.555	0.577
RF	0.763	0.475	0.364	0.460	0.501
LinSVC	0.804	0.553	0.512	0.559	0.603
GB	0.759	0.493	0.413	0.509	0.530

Таблица 24: BSNLP EU Results

	ORG	LOC	MISC	PER	Total
LogReg	0.652	0.504	0.000	0.313	0.525
RF	0.386	0.403	0.000	0.225	0.328
LinSVC	0.668	0.540	0.000	0.359	0.543
GB	0.655	0.497	0.000	0.368	0.529

Таблица 25: BSNLP Trump Results

	ORG	LOC	MISC	PER	Total
LogReg	0.426	0.820	0.000	0.883	0.756
RF	0.289	0.700	0.000	0.823	0.674
LinSVC	0.391	0.805	0.000	0.860	0.734
GB	0.285	0.782	0.000	0.840	0.703

8 Общий подбор параметров

Список литературы

[1] Chieu, H. L., and Ng, H. T. Named entity recognition with a maximum entropy approach. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (Stroudsburg, PA, USA, 2003), CONLL '03, Association for Computational Linguistics, pp. 160–163.

[2] Chiu, J. P. C., and Nichols, E. Named entity recognition with bidirectional lstm-cnns. CoRR abs/1511.08308 (2015).

[3] Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. Named entity recognition through classifier combination. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (Stroudsburg, PA, USA, 2003), CONLL '03, Association for Computational Linguistics, pp. 168–171.

[4] Klein, D., Smarr, J., Nguyen, H., and Manning, C. D. Named entity recognition with character-level models. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (Stroudsburg, PA, USA, 2003), CONLL '03, Association for Computational Linguistics, pp. 180–183.

[5] Ma, X., and Hovy, E. H. End-to-end sequence labeling via bi-directional lstm-cnns-crf. CoRR abs/1603.01354 (2016).

[6] McCallum, A., and Li, W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (Stroudsburg, PA, USA, 2003), CONLL '03, Association for Computational Linguistics, pp. 188–191.

[7] 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.