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.

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.