Распознавание именованных сущностей с использованием синтактико-семантических признаков и нейросетей

Юсупов Идрис (i.yusupov@phystech.edu)

1. Цель работы

Исследовать возможность использования семантико-синтаксического анализатора Compreno в качестве источника высокоуровневых признаков для задачи распознавания именованных сущностей (NER) на корпусе CoNLL 2003 в рамках нейросетевого подхода.

2. Корпус CoNLL 2003

CoNLL 2003 [Tjong Kim Sang and De Meulder 2003] - англоязычный корпус для оценки качества методов распознавания именованных сущностей.

Размечено 4 типа сущностей - персоны (PER), организации (ORG), локации (LOC) и другие (MISC). Корпус размечен по схеме Inside, Outside, Begin (IOB).

Оценка качества: F1-micro-average.

В данной работе используется схема IOBES (Inside, Outside, Begin, End, Single), поэтому получается 17 классов. Четыре для каждого из четырех типов тегов и один для Outside.

2.1. Результаты на CoNLL 2003

Модель	F1 (%)	Примечание
Florian et al. 2003	88.76	Много вручную построенных признаков
Collobert et al. 2011	89.59	-
Chiu and Nichols 2015	91.62	Обучались также на валидационной выборке и использовали сторонний газетир
Yang et al. 2016	90.94	-

3. Compreno признаки

Вершины синтактико-семантического дерева Compreno [Anisimovich et al. 2012] кодировались бинарным представлением и соотносились с токенами исходного текста, тем самым наделяя их синтактико-семантическими признаками.

Полученная размерность пространства синтактико-семантических признаков: 83950.

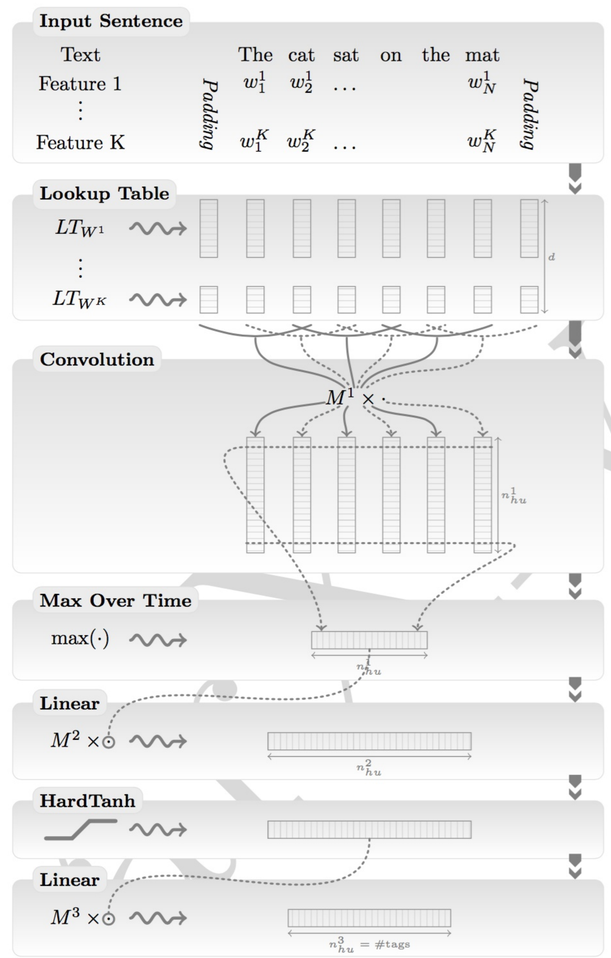
4. Модель

Для реализации была выбрана сверточная нейронная сеть из статьи [Collobert et al. 2011]. Из модели были удалены условные случайные поля для более быстрого обучения и проведения экспериментов.

4.1. Признаки

Вектора слов (senna embeddings [Collobert et al. 2011]); позиция относительно слова в предложении для которого предсказывается тег; капитализация; присутствие слова в газетире, который включен в соревнование CoNLL 2003.

4.2. Convolution Net [Collobert et al. 2011]

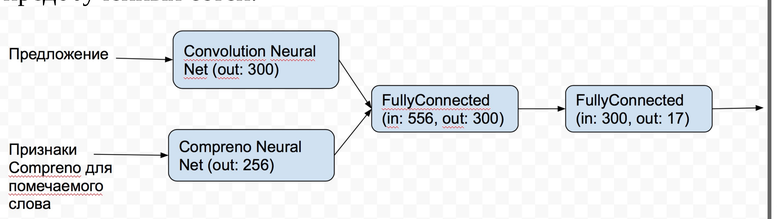


5. Compreno Net

(1): nn.SparseLinear(83951, 256)
(2): nn.Dropout(0.5)
(3): nn.HardTanh
(4): nn.Linear(256 -> 256)
(5): nn.Dropout(0.5)
(6): nn.HardTanh
(7): nn.Linear(256 -> 17)

6. Convolution Net + Compreno Net

Веса нейросетей были инициализированы весами предобученных сетей.



7. Среда для проведения экспериментов

Нейронная сеть: открытый deep learning фреймворк torch.

Обучение проводилось на Amazon AWS g2.2xlarge.

Время обучения: 4.2 часа (91 эпоха).

Код для воспроизведения экспериментов: http://github.com/sld/torch-conv-ner.

8. Результаты экспериментов

Модель	Признаки	Выборка	Метод оптимизации	Полученная F1, %
Compreno Net	Compreno sparse features	train	Mini-batch gradient descent	72.85
ConvNet	Embeddings, Capitalization, Position, Gazetteer	train	Mini-batch gradient descent	87.49
ConvNet + Compreno Net	Embeddings, Capitalization, Position, Gazetteer, Compreno sparse features	train	Mini-batch gradient descent	**88.47**
ConvNet + Compreno Net	Embeddings, Capitalization, Position, Gazetteer, Compreno sparse features	train + dev	Mini-batch gradient descent	88.81

9. Заключение

Был найден простой вариант подключения признаков Compreno к сверточной нейронной сети за счет которого F1-мера повысилась с 87.49% до 88.47%.

В будущем планируется:

- внедрить условные случайные поля в существующую модель для повышения F1-меры,

- исследовать работу предложенного решения на других корпусах.

Список литературы

Yang, Z., Salakhutdinov, R., and Cohen, W. (2016). Multi-task cross-lingual sequence tagging from scratch. CoRR, abs/1603.06270.

Chiu, J. P. and Nichols, E. (2015). Named entity recognition with bidirectional lstm-cnns. arXiv preprint arXiv:1511.08308.

Anisimovich, K. V., Druzhkin, K. J., Minlos, F. R., Petrova, M. A., Selegey, V. P., and Zuev, K. A. (2012). Syntactic and semantic parser based on ABBYY Compreno linguistic technologies.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch.

Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition.

Благодарности

Автор благодарит Анатолия Старостина, Ивана Смурова и Станислава Джумаева за ценные советы и комментарии.