Распознавание именованных сущностей с использованием синтактико-семантических признаков и нейросетей

Юсупов Идрис (i.yusupov@phystech.edu)

Московский физико-технический институт

Аннотация

В данной работе исследованы различные способы внедрения синтактико-семантических признаков в нейронные сети для решения задачи распознавания именованных сущностей. Исследование проводилось в рамках англоязычного корпуса CoNLL 2003. Полученные результаты показывают, что синтактико-семантические признаки дают ощутимый прирост оценки качества, без какой-либо инженерии над ними.

Ключевые слова: нейронные сети, распознавание именованных сущностей.

1 Введение

Сегодня глубокие нейронные сети показывают лучшие результаты в самых разных областях: от обработки изображений [Krizhevsky et al. 2012] до анализа данных пациентов [Miotto et al. 2016]. На нейронные сети обычно подают простые признаки, такие как пиксели изображения или позицию слова в словаре, таким образом освобождая человека от инженерии признаков. В то же время, например при работе с текстами, имеется большое количество доступных морфологических, синтаксических и семантических признаков. Их внедрение в нейронную сеть не должно вызывать трудностей и может положительно сказаться на оценке качества.

В данной работе исследовано влияние синтактико-семантических признаков Compreno для задачи NER на корпусе CoNLL 2003 при использовании нейросетевого подхода.

Статья организована следующим образом: в части 1 проведен обзор связанных работ. Выбранная нейросетевая модель и способы внедрения синтактико-семантических признаков описаны в части 2. В части 3 описаны проведенные эксперименты и программная реализация.

Полученные результаты показывают повышение F1 меры почти на 1% на корпусе CoNLL 2003 при использовании синтактико-семантических признаков Compreno (87.49% против 88.47%). При этом затраты на их внедрение были минимальными - инженерия над признаками не проводилась.

2 Связанные работы

Победители соревнования по NER CoNLL 2003 [Florian et al. 2003], получившие 88.76% F1, представили систему использующую комбинацию различных алгоритмов машинного обучения. В качестве признаков был использован их собственный, вручную составленный газетир, POS-теги, CHUNK-теги, суффиксы, префиксы и выход других NER-классификаторов тренированных на внешних данных.

Современные модели используют минимальное количество вручную составленных признаков, независимую от задачи и языка архитектуру.

[Collobert et al. 2011] представили комбинацию сверточной нейронной сети с условными случайными полями, получившую 89.59% F1 на датасете CoNLL 2003. Их нейросетевая архитектура не зависит от задачи и используется как для NER, так и для POS-теггинга, CHUNK-теггинга, установления семантических ролей (semantic role labelling). Для задачи NER они использовали три типа признаков - векторное представление слова, капитализацию и небольшой газетир, включенный в соревнование CoNLL 2003.

[Chiu and Nichols 2015] представили комбинацию сверточных сетей, рекуррентных сетей и условных случайных полей. Они использовали такие же признаки как у [Collobert et al. 2011], дополнительный, вручную сформированный газетир на основе DBpedia и обучались на объединенной train+dev выборке CoNLL 2003. У них получилось 91.62%

F1. Кроме корпуса CoNLL 2003 они тестировали архитектуру на более крупном англоязычном корпусе OntoNotes 5.0. На нем они получили state-of-the-art результат 86.28%.

[Yang et al. 2016] представили глубокую иерархическую рекуррентную нейросетевую архитектуру с условными случайными полями для разметки последовательностей. Они использовали такие же признаки как у [Collobert et al. 2011]. Кроме англоязычного корпуса CoNLL 2003, где они получили state-of-the-art 90.94% F1 при обучении только на train выборке, они тестировали работу нейросети на CoNLL 2002 Dutch NER и CoNLL 2003 Spanish NER. На этих датасетах они улучшили предыдущий state-of-the-art результат: 82.82% до 85.19% на CoNLL 2002 Dutch NER и 85.75% до 85.77% на CoNLL 2003 Spanish NER.

Все современные работы используют векторное представление слов и условные случайные поля в своих моделях. Из сторонних признаков используются только газетиры.

В работах [Xu et al. 2014, Bian et al. 2014] описано применение дополнительных признаков для слов (морфологических, синтаксических, семантических) для создания более совершенных векторных представлений. Такие векторные представления помогают повысить оценку качества в прикладных задачах. Например, при использовании Skip-gram модели, точность в analogy reasoning task на датасете WordRep [Gao et al. 2014] - 31.30%, а при использовании дополнительной информации для создания векторов слов - 39.85% [Xu et al. 2014].

3 Модель

Для реализации была выбрана сверточная нейронная сеть из статьи [Collobert et al. 2011]. Для быстрой имплементации и обучения из модели были удалены условные случайные поля.

В качестве признаков выступают вектора слов, позиция относительно слова в предложении для которого предсказывается тег, капитализация и присутствие слова в газетире, который включен в соревнование CoNLL 2003.

Общий алгоритм работы следующий:

1. На вход нейросети поступает набор идентификаторов признаков для всего предложения.

2. Набор идентификаторов пропускается через Lookup Table для каждого признака. Lookup Table отображает идентификатор в вектор обучаемых весов.

3. Полученная матрица признаков для всего предложения подается на следующий слой, который проходится окном размера 3 и выполняет операцию свертки (temporal convolution). Затем извлекается максимум по каждой строке (max over time). Таким образом для предложения любой длины получается фиксированный вектор признаков.

4. Полученный вектор признаков подается на полносвязный слой.

5. Затем выход полносвязного слоя подается на выходной слой, который возвращает вероятность для каждого тега (softmax).

В качестве функции потерь используется кросс-энтропия (cross entropy).

Подробная математическая модель описана в статье [Collobert et al. 2011].

3.1 Синтактико-семантические признаки

Существует много инструментов для получения дополнительных признаков для слова. Например, для извлечения синтаксических признаков есть MaltParser [Nivre et al. 2006]. Для получения семантических признаков есть BabelNet [Navigli and Ponzetto 2010]. Также существуют интегрированные инструменты, которые проводят синтактико-семантический анализ. Например, Compreno [Anisimovich et al. 2012].

Часто такие признаки кодируют бинарным представлением. Полученные вектора могут иметь большую размерность - больше 1000.

Плотные вектора большой размерности будут сильно замедлять процесс оптимизации и для хорошего обучения потребуется много данных и вычислительных ресурсов. В таких случаях часто применяют методы для уменьшения размерности, например сингулярное разложение или автоэнкодеры. Минусом таких методов является потеря информации после сжатия.

Если же вектора большой размерности разреженные, то используют специальные методы для работы с такими данными.

В данной работе предлагается 2 способа внедрения синтактико-семантических признаков:

- сжать синтактико-семантические вектора с помощью сингулярного разложения и добавить как еще один Lookup Table в сверточную нейронную сеть;
- добавить еще одну нейронную сеть для синтактико-семантических признаков и оптимизировать её вместе со сверточной нейронной сетью.

4 Эксперименты

4.1 Корпус CoNLL 2003

CoNLL 2003 [Tjong Kim Sang and De Meulder 2003] - англоязычный корпус для оценки качества методов распознавания именованных сущностей. Корпус содержит обучающую, тестовую и валидационную выборку. Размечено 4 типа сущностей - персоны (PER), организации (ORG), локации (LOC) и другие (MISC). Корпус размечен по схеме Inside, Outside, Begin (IOB). Оценка качества считается с помощью F1-micro-average.

Как и у [Collobert et al. 2011], данные были сконвертированы из схемы IOB в схему IOBES (Inside, Outside, Begin, End, Single). Во время тестирования, данные конверти-

Таблица 1: Количество статей, предложений, токенов и именованных сущностей

Выборка	Статьи	Предложения	Токены	LOC	MISC	ORG	PER
Обучающая	946	14987	203621	7140	3438	6321	6600
Валидационная	216	3466	51362	1837	922	1341	1842
Тестовая	231	3684	46435	1668	702	1661	1617

руются обратно в формат IOB и подаются на вход скрипта, включенного в CoNLL 2003, оценивающего качество классификации.

4.2 Синтактико-семантические признаки Compreno

Синтактико-семантически признаки были получены с помощью Compreno. Они представляют собой разреженные вектора размерности 83950. Они покрывают около 60% корпуса CoNLL 2003. Все слова, не покрытые Compreno, кодировались как дополнительный признак 83951.

4.3 Эксперименты без синтактико-семантических признаков

Нейросетевая модель имеет такие же параметры как и у [Collobert et al. 2011]. Небольшой модификацией является добавление Dropout слоя в качестве регуляризатора, после каждого полносвязного слоя. Размерность выходного слоя - 17. 4 для каждого из тегов и 1 для Outside.

В качестве векторного представления слов (embeddings в таблице 2), использовались Senna embeddings[1], которые находятся в открытом доступе.

[1]http://ronan.collobert.com/senna/

Таблица 2: Результаты экспериментов без использования синтактико-семантических признаков

Модель	Признаки	Датасет	Метод оптимизации	Полученная F1, %	F1 в статье Collobert et al. [2011]
Window	Embeddings, Capitalization	train	Mini-batch gradient descent	86.27	-
Window	Embeddings, Capitalization	train	Stochastic gradient descent	-	86.97
ConvNet + CRF	Embeddings, Capitalization, Position	train	Stochastic gradient descent	-	88.67
ConvNet + CRF	Embeddings, Capitalization, Position, Gazetteer	train	Stochastic gradient descent	-	89.59
ConvNet	Embeddings, Capitalization, Position	train	Stochastic gradient descent	86.77	-
ConvNet	Embeddings, Capitalization, Position, Gazetteer	train	Stochastic gradient descent	87.89	-
ConvNet	Embeddings, Capitalization, Position, Gazetteer	train + dev	Stochastic gradient descent	88.37	-
ConvNet	Embeddings, Capitalization, Position, Gazetteer	train	Mini-batch gradient descent	**87.49**	-

По таблице 2 видно, что результаты немного ниже чем у [Collobert et al. 2011]. Это связано с тем, что для Window подхода использован другой метод оптимизации, а для Convolution подхода не были применены условные случайные поля.

В качестве референсной, будет использована модель из последнего эксперимента показывающая 87.49% F1. Это сделано для чистоты эксперимента, т.к. далее обучение происходило только на train выборке по правилам соревнования CoNLL 2003 и применялся mini-batch gradient descent для ускорения экспериментов.

4.4 Эксперименты с синтактико-семантическими признаками SVD 1024

Синтактико-семантически признаки Compreno размерности 83950 были сжаты с использованием TruncatedSVD[2] до размерности 1024. После сжатия описываемая дисперсия была равна 72%. Т.е. потерялось 28% информации. Сжатые вектора были добавлены в нейронную сеть с помощью дополнительного Lookup Table слоя.

[2]http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

Таблица 3: Результаты с синтактико-семантическими признаками сжатыми SVD

Модель	Признаки	Да-та-сет	Метод оптими-зации	Получен-ная F1, %
ConvNet	Position, Compreno SVD 1024	train	Mini-batch gradient descent	75.89
ConvNet	Capitalization, Position, Gazetteer, Compreno SVD 1024	train	Mini-batch gradient descent	81.83
ConvNet	Embeddings, Capitalization, Position, Gazetteer, Compreno SVD 1024	train	Mini-batch gradient descent	86.85
ConvNet	Embeddings, Capitalization, Position, Gazetteer	train	Mini-batch gradient descent	87.49

По таблице 3 видно, что такой способ ведет к небольшому ухудшению F1 меры.

4.5 Эксперименты с синтактико-семантическими признаками для совместно-оптимизированной нейросети

Была добавлена еще одна нейронная сеть для синтактико-семантических признаков и оптимизирована вместе со сверточной нейронной сетью.

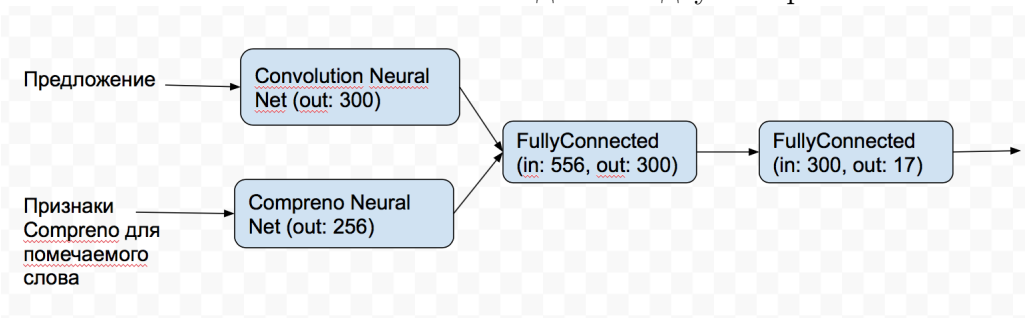
Нейронная сеть для синтактико-семантических признаков работает следующим образом:

1. На вход подается разреженный вектор признаков слова (размерность 83951) для которого предсказывается тег.
2. Далее этот вектор пропускается через 2 полносвязных слоя.
3. На выходе еще один полносвязный слой, который выдает вероятность определенного тега. Выходов также 17.

Сверточная сеть, учитывающая всё предложение, и полносвязная сеть, обрабатывающая синтактико-семантические признаки слова для которого предсказывается тег, соединяются следующим образом (рис. 1):

1. Из обеих нейросетей удаляются выходные слои.
2. Предыдущие слои из обеих сетей соединяются в новый полносвязный слой.
3. Новый полносвязный слой соединяется с выходным слоем. Выходов как и тегов 17.

Рис. 1: Объединение двух нейросетей



Веса у объединенной сети были инициализированы обученными моделями - моделью показывающую 87.49% для сверточной сети и моделью показывающую 72.85% (см. таблицу 4) для второй нейронной сети.

Таблица 4: Результаты с синтактико-семантическими признаками для объединенной нейросети

Модель	Признаки	Датасет	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

По таблице 4 видно, что признаки Compreno улучшают F1-меру почти на один процент.

4.6 Программная реализация

Нейронная сеть написана с использованием открытого фреймворка torch[3].

Код для воспроизведения экспериментов будет выложен по адресу: github.com/sld/torch-conv-ner.

Скорость обучения на машине с GPU Amazon AWS g2.2xlarge[4]:

- 1 эпоха при одиночной обработке (stochastic gradient descent): ∼450 сек.
- 1 эпоха при пакетной обработке (mini-batch gradient descent): ∼171 сек.

[3]http://torch.ch
[4]https://aws.amazon.com/ru/ec2/instance-types/

- Модель получающая 87.49% обучалась 91 эпоху (∼4.2 часа).
- 1 эпоха при пакетной обработке с использованием признаков Compreno: ∼615 сек.

Скорость классификации составляет 2500 токенов в секунду при пакетной обработке.

5 Заключение

В данной работе исследовано влияние синтактико-семантических признаков Compreno для задачи NER на корпусе CoNLL 2003 при использовании нейросетевого подхода. Признаки Compreno, при достаточно низких затратах на внедрение, увеличивают F1 меру.

В будущем планируется внедрить условные случайные поля в существующую модель для повышения F1 меры и исследовать работу предложенного решения на других корпусах. Также интересным направлением для исследований является создание векторного представления слов с учетам синтактико-семантических признаков.

6 Благодарности

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

Список литературы

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. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialog*, page 18.

Bian, J., Gao, B., and Liu, T.-Y. (2014). Knowledge-powered deep learning for word embedding. In *Machine Learning and Knowledge Discovery in Databases*, pages 132–148. Springer.

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

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics.

Gao, B., Bian, J., and Liu, T.-Y. (2014). Wordrep: A benchmark for research on learning word representations. *arXiv preprint arXiv:1407.1640*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6:26094 EP –.

Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Xu, C., Bai, Y., Bian, J., Gao, B., Wang, G., Liu, X., and Liu, T.-Y. (2014). Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1219–1228. ACM.

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