Частичный синтаксический разбор текста как пометка частей речи

Реферат

Обработка частичного синтаксический разбор как пометки дают результаты, сравниваемые с другими, более сложными подходами. Используя обучение и тестирование материала CoNLL 2000, наша лучшая модель имела точность 94,88%, с общим счетом FB1 – 91,94%. Отдельные показатели FB1 для NP были 92.19%, VP – 92.70% и PP – 96.69%.

1 Введение

Частичный синтаксический разбор получил достаточное количество внимания за последние несколько лет (например (Ramshaw и Маркус, 1995)). В этой статье, вместо того, чтобы изменить какую-то существующую технику, или предложить совершенно новый подход, мы решили построить частичный синтаксический парсер с использованием готового теггера частей речи (POS теггер). В любом случае, мы сознательно не меняли внутренние операции POS теггера. Наши результаты свидетельствуют о том, что для достижения разумной производительности частичного синтаксического парсера, в целом, не требуется ничего более сложного, чем простой POS теггер. Тем не менее, анализ ошибок предположил существование небольшого набора конструкций, которые не так легко характеризуются конечным числом состояний подходов, таких как наш.

2 Теггер

Мы использовали POS теггер (Ratnaparkhi, 1996) на основе максимальной энтропии. При пометках модель пытается восстановить наиболее вероятную (ненаблюдаемую) последовательность тегов, учитывая последовательность наблюдаемых слов.

Для наших экспериментов мы использовали только теггер двоичного распределения (Ratnaparkhi, 1996).

3 Убеждение теггера частичному синтаксическому разбору

Пониманием здесь является то, что можно посмотреть в (некоторые) различия между пометками и (частичным) разбором в качестве одного из контекста: частичный синтаксический разбор требует доступ к большей части окружающей лексической/POS синтаксической окружающей среды,

чем делает простой POS теггер. Эта дополнительная информация может быть закодирована.

Однако, необходимо сбалансировать этот подход тем, что как количество информации в состояниях увеличивается с ограниченным уровнем подготовки материала, шанс увидеть такое состояние снова в будущем уменьшается. Поэтому мы ожидали бы увеличение производительности, поскольку мы увеличили количество информации в состояниях, а затем уменьшили, когда переобучили и/или сделали более редкой статистику доминирующих факторов.

Мы обучили теггер, используя «слова» различной «конфигурации» (конкатенаций) фактических слов, POS-теги, типы чанков, и/или суффиксы или префиксы слов. По образованию при этих сцеплений, мы помогаем преодолеть разрыв между простой POS пометки и мелкой разбора. Тренируясь этими конкатенациями, мы помогаем преодолеть разрыв между обычным POS теггером и частичным синтаксическим разбором.

В оставшейся части работы, мы имеем в виду то, что теггер считает слово в качестве конфигурации. Конфигурация будет объединением различных элементов обучающей выборки соответствующих принятию решений относительно назначения чанка. "Слово" будет означать слово, найденное в обучающей выборке. 'Пометки' относятся к POS тэгам, найденным в обучающем наборе. Опять же, эти теги могут быть частью конфигурации. Мы имеем в виду то, что теггер считает теги как прогнозы. Прогнозы будут метками чанков.

4 Эксперименты

Теперь мы дадим детали экспериментов, с которым столкнулись. Чтобы было понятней, рассмотрим следующий фрагмент обучающего набора:

Word	W1	W2	W3
POS tag	T1	T2	T3
Chunk	C1	C2	C3

Слова – это w1, w2 и w3, POS теги – t1, t2 и t3, метки чанков – c1, c2 и c3. мы построили различные конфигурации при прогнозировании метки чанка для слова W1.

Что касается ситуации только что упомянутой (прогнозирования метку для слова W1), мы постепенно увеличивается количество информации в каждой конфигурации следующим образом:

1. Конфигурация, состоящая из просто слов (слово W1). Результаты:

Тип чанка	P	R	FB1
Общий	88.06	88.71	88.38
ADJP	67.57	51.37	58.37
ADVP	74.34	74.25	74.29
CONJP	54.55	66.67	6O.00
INTJ	100.00	50.00	66.67
LST	0.00	0.00	0.00
NP	87.84	89.41	88.62
PP	94.80	95.91	95.35
PRT	71.00	66.98	68.93
SBAR	82.30	72.15	76.89
VP	86.68	88.15	87.41

Общая точность: 92.76%

2. Конфигурация, состоящая из просто меток (метка T1). Результаты:

Тип чанка	P	R	FB1
Общий	88.15	88.07	88.11
ADJP	67.99	54.79	60.68
ADVP	71.61	70.79	71.20
CONJP	35.71	55.56	43.48
INTJ	0.00	0.00	0.00
LST	0.00	0.00	0.00
NP	89.47	89.57	89.52
PP	87.70	95.28	91.33
PRT	52.27	21.70	30.67
SBAR	83.92	31.21	45.50
VP	90.38	91.18	90.78

Общая точность: 92.66%

3. Оба слова, теги и текущая метка чанка (W1, T1, C1) в конфигурации. Мы предоставляли доступ теггеру к текущей метке чанка путем обучения другой модели с конфигурацией, состоящей из тегов и слов (W1 и T1). Обучающий набор затем восстанавливали к состоянию из конфигураций слов-тегов и отметками использования этой модели. После этого мы собрали прогнозы для использования в второй модели. Результаты:

Тип чанка	P	R	FB1
Общий	89.79	90.70	90.24
ADJP	69.61	57.53	63.00
ADVP	74.72	77.14	75.91
CONJP	54.55	66.67	60.00
INTJ	50.00	50.00	50.00
LST	0.00	0.00	0.00
NP	89.80	91.12	90.4
PP	95.15	96.26	95.70
PRT	71.84	69.81	70.81
SBAR	85.63	80.19	82.82
VP	89.54	91.31	90.41

Общая точность: 93.79%

4. Окончательная конфигурация сделала попытку принять дело с разреженными статистиками. Она состояла из текущего тега t1, следующего тега t2, текущей метки чанка c1, двух последних букв следующей пометки чанка c2, первых двух букв текущего слова w1, и последних четырех букв текущего слова w1. Эта конфигурация была результатом многочисленных экспериментов и дала лучшую производительность. Результаты могут быть увидены в таблице 1.

Мы отметили наши эксперименты в разделе комментариев.

5 Анализ ошибок

Мы изучили работу нашей окончательной модели по отношению к тестируемому материалу и обнаружили, что ошибки, допущенные нашим частичным синтаксическим парсером могут быть сгруппированы в три категории: тяжелые синтаксические конструкции, ошибки, допущенные в обучении или тестировании материала аннотаторами, и ошибки, свойственные нашему подходу.

Принимая каждую категорию из трех, в свою очередь, проблемные конструкции включают: координацию, знаки препинания, обработки двухобъектных переходных VP, как переходные VPS. Путаница в отношении прилагательных или адвербиальных фраз рассматривюется как притяжательные.

Ошибками (шумом) в обучении и тестировании материала были, в основном, POS метки ошибок. Дополнительным источником ошибок были нечетные решения аннотирования.

Окончательный источник ошибок был свойственнен нашей системе. Экспоненциальное распределение (которое использует наш теггер) назначает ненулевую вероятность для всех возможных событий. Это означает, что теггер будет время от времени назначать метки чанков, которые являются незаконными, например, присвоение метки слова I-NP, когда слово не в NP. Хотя эти ошибки были нечастыми, устранение их потребует "открытости" в теггере и отвержения незаконных гипотез меток чанков из рассмотрения.

6 Комментарии

Как утверждалось во введении, увеличение размера контекста дает лучшие результаты, в которых обеспечивается ограничение таких проблем, как разреженные статистики. Наши эксперименты показывают, что это было действительно так.

Мы не делаем никаких заявлений относительно общности нашего моделирования. Очевидно, что он специфичен для использования теггера.

Более подробно, мы обнаружили, что:

- PP кажется легко идентифицировать.
- ADJP и ADVP чанки было трудно определить правильно. Мы подозреваем, что улучшения здесь требуют большей синтаксической информации, чем только базовых фраз.
- Наше выступление в NP должны быть улучшены. С точки зрения моделирования, мы не относимся ни к одному чанку, в отличие от любого другого чанка. Мы также не обрабатываем слова по-разному, в отличие от любых других слов.
- Производительность в использовании только слов и POS тегов были примерно эквивалентны. Тем не менее, производительность, используя оба источника была лучше, чем при использовании любого источника информации в изоляции. Причина этого в том, что слова и POS теги имеют различные свойства, и вместе специфичность слов может преодолеть неровности меток, в то время как обилие тегов может иметь дело с малой плотностью слов.

Наши результаты не были настолько плохими, как высказал Buchholz и др. (Sabine Buchholz and Daelemans, 1999). Этот сопоставимый уровень

производительности позволяет предположить, что частичный синтаксический разбор (распознавание базовых фраз) является довольно простой задачей. Улучшения может исходить от лучшего моделирования, борьбой с незаконной последовательностью порций, что позволило бы умножить чанки с доверительными интервалами, комбинациями системы и т.д., но мы чувствуем, что такие улучшения будут малы.