Неконтролируемый синтаксический анализ

Unsupervised parsing

Потемкин С. Б. (potemkin@philol.msu.ru)

Филологический факультет МГУ, Москва, Россия

Представлен статистический подход к синтаксическому анализу необработанных текстов в формализме дерева зависимостей. Приведен алгоритм, выполняющий парсинг зависимостей за время, квадратично зависящее от длины предложения, после обучения на неразмеченном корпусе.

1. Введение

Решение проблемы автоматического синтаксического анализа без предварительной ручной настройки и обучения на размеченном корпусе имеет большое теоретическое и практическое значение. Правила разбора, полученные в результате, могут пролить свет на процессы освоения языка людьми и на общую структуру языка, обеспечить предварительную обработку текстов при синтаксической разметке больших корпусов и, в перспективе, качественный анализ текстов для прикладных задач обработки естественного языка. Интерес к этой проблеме существенно повысился благодаря доступности огромных корпусов текстов, росту вычислительных мощностей и новым алгоритмам машинного обучения.

В последнее время прилагались большие усилия по использованию для синтаксического анализа размеченных корпусов, которые позволяют проводить проверку гипотез, выдвигаемых грамматическими теориями, а также формировать сами правила синтаксиса. Этот процесс называется «тренировкой» формальной грамматики и должен завершиться при достижении некоторого малого процента ошибок. Для тренировки грамматики составляются синтаксически аннотированные корпуса, которые получили название «treebank». В настоящее время имеются три-банки для языков: болгарского (BulTreeBank), польского (Проект CRIT-2), русского (ЭТАП-3, ИППИ РАН), и наиболее продвинутый для чешского языка (Prague Dependency Treebank). Ведутся работы для балканских (сербохорватского, словенского, боснийского) языков.

Большинство работ по синтаксическому анализу основано либо на правилах, либо на управляемом обучении. Хорошие синтаксические анализаторы в формализме непосредственных составляющих (НС) доступны для английского и некоторых других языков [3]. Также имеются работы, основанные на формализме дерева зависимостей [6, 7, 8, 10].

Для большинства языков мира, однако, отсутствуют хорошие синтаксические анализаторы, либо вообще какие-либо анализаторы. Это связано с тем фактом, что для большинства языков отсутствуют ресурсы, необходимые как для построения парсеров на основе правил (полные компьютерные грамматики), так и парсеров, обучаемых на примерах из три-банка. Поскольку создание таких ресурсов требует больших материальных и трудовых затрат, желательно разработать достаточно точные методы для выполнения грамматического разбора без обучения на три-банке, или для автоматического или полуавтоматического создания три-банка.

В течение последних лет наблюдается устойчивый прогресс в области неконтролируемого синтаксического анализа, но большая часть работ основана на НС-грамматиках, тогда как для описания синтаксиса русского языка традиционно используется модель зависимостей [11].

2. Достигнутый уровень разработок

Один из подходов к упрощению и облегчению синтаксической разметки национального корпуса заключается в использовании размеченного кор-

пуса другого, хорошо документированного языка, с применением специально создаваемых для этой цели алгоритмов «перевода разметки». В качестве опорного размеченного корпуса обычно выступает PennTreeBank английского языка. Поскольку для большинства языков имеются хотя бы переводные английские словари, задача разметки, в общем, упрощается, хотя результаты и неидеальны. Это особенно заметно для славянских языков с относительно свободным порядком слов, грамматика которых обычно описывается формализмом зависимостей, тогда как в PennTreeBank использован НС формализм.

Другой, чисто статистический подход, имеет определенные преимущества — необходимо иметь лишь ограниченный (около 1 млн. словоупотреблений) неразмеченный национальный корпус, без корпуса параллельных текстов и даже без двуязычного переводного словаря. Это особенно важно для малых и исчезающих языков, для описания которых отсутствуют материальные и людские ресурсы.

Статистический подход к синтаксическому разбору предложений применен в нескольких, взаимосвязанных методиках, включая DLM (dependency language model) (Gao, Suzuki, 2003) [4], U-DOP (unsupervised data-oriented parsing) (Bod, 2006) [2], CCL (common cover links) (Seginer 2007) [9].

В рамках метода Бода для выполнения парсинга необходимо:

- Построить все возможные деревья разбора для всех предложений корпуса и все поддеревья каждого дерева разбора.
- Найти наилучшее (наиболее вероятное) дерево разбора для данного предложения.

Фактически, при реализации метода возникают значительные вычислительные трудности, поскольку рост числа поддеревьев превышает экспоненциальный (каталанские числа) с удлинением предложения. Для решения этих проблем предложены методы представления поддеревьев в виде вероятностной контекстно-свободной грамматике, [5] и записи всех деревьев в виде «совместного леса» [1], что сводит задачу к обозримому, однако очень значительному, объему вычислений.

В подходе Сегинера общепринятое представление структуры предложения в виде дерева зависимостей заменяется совокупностью «общих покрывающих связей» (Common Cover Links, CCL). Разбор предложения проводится последовательно, пословно, путем анализа начальной последовательности слов предложения. Результаты такого частичного анализа в дальнейшем не изменяются, а лишь дополняются. Каждая новая связь добавляется, если она не нарушает определенные правила, заданные априори и если она обладает максимальным (среди допустимых) весом. Для определения весов создается лексикон, содержащий для каждого встреченного в тексте слова список левых и правых, связанных с ним, соседей и частоту таких связей.

По сравнению со структурой зависимостей CCL структура обладает определенными преимуществами: во-первых, для предложения типа «I know the boy sleeps» со структурой зависимостей [[I][know] [[the boy][sleeps]]] CCL не устанавливает направления связи в отношении [the boy]. Аналогично, для русского языка не будет установлено направление управления в предложно-падежной группе.

Второе отличие более существенно. В традиционном методе к моменту прочтения предложения до слова boy будет установлена зависимость между know и boy, однако после прочтения предложения до конца, придется удалить эту связь и установить новые – [know sleeps] и [sleeps boy] Эта проблема известна в психолингвистике как проблема повторного анализа. В CCL структуре эта проблема обойдена путем назначения каждой связи значения «глубины» этой связи. Этим достигается однозначность восстановления скобочной структуры, без необходимости удаления ранее установленных связей. Парсер на основе CCL, настроенный на английский язык, доступен для некоммерческого использования, http://staff.science.uva.nl/~yseginer/ccl/.

Наконец, Гао и Судзуки также предложили инкрементный подход к парсингу, при котором структура зависимостей строится последовательно, после ввода очередного слова предложения и вычеркивания связей, нарушающих ацикличность и проективность. Их метод был применен не к анализу структуры предложения, а к восстановлению иероглифической записи японского предложения (кана-кандзи) на основании слоговой записи (кана) – эта проблема и метод ее инкрементного решения характерны также для задачи распознавания речи.

Настоящая работа опирается в основном на методику Гао и Судзуки, однако для анализа зависимостей предложения разработан алгоритм, строящий покрывающее дерево всего предложения, без вычеркивания связей, работающий за время $O(n^2)$, при сохранении классического вида структуры зависимостей.

На основе представленного метода возможна автоматическая синтаксическая разметка неаннотированного корпуса, как для русского языка, так и для других языков, имеющих достаточные объемы электронных текстов, с преобладанием проективных предложений.

3. Модель локальных связей (МЛС)

В предлагаемой нами модели локальных связей стру ктура зависимостей строится снизу вверх. Вначале устанавливаются связи между соседними словами (локальность), которые объединяются в юниты, затем устанавливаются связи между соседними юнитами, и так далее, пока не достигает-

ся последний, верхний уровень объединения, чем и завершается построение дерева зависимостей. Существенным в этом процессе является выбор последовательности объединения юнитов, который определяется весом связи между ними. Аналогично модели грамматики связей (LG) [12] в нашей модели установленные связи являются ненаправленными, но в отличие от указанной работы, связи не помечаются и их установка не требует заранее подготовленного лексикона моделей управления.

3.1. Определения

Для более формального описания модели введем обозначения:

W — последовательность слов предложения;
W={w1, w2,..., wn}

T — дерево зависимостей, построенное над W;
T = {(i, j)}, где i, j – номера слов, связанных зависимостью. T является проективным деревом.

U — юнит, поддерево T над неразрывной подпоследовательностью W; U_{k0}*=wk, либо* U_{kl} *= {wk, wk+1, …, w+l}, где каждая пара слов связана ветвью, принадлежащей дереву T.*

Открытой вершиной юнита U назовем такую вершину wm, для которой не существует принадлежащей U ветви (i, j), i < m < j. Иначе вершина wm является закрытой.

Смежными юнитами U_{al}={wa, wa+1,…,wap} U_{bm} = {wb, wb+1,…, wbq} назовем такие юниты, для которых b=ap+1, то есть начало второго юнита непосредственно следует за концом первого юнита.

В принципе, модель языка должна определять вероятность предложения W по всем возможным деревьям T над W, то есть

$$P(W) = \Sigma P(W, T) \text{ по всем } T. \quad (1)$$

Практически, для оценки P(W) используется единственный член суммы, а именно P(W,T*):
где T* — наиболее вероятная структура зависимостей предложения, которая доставляет максимум выражению P(W,T):

$$T^* = \operatorname{argmax} P(W,T) \quad (2)$$

Цель парсинга состоит в том, чтобы найти самый вероятный разбор T* данного предложения W, максимизирующий вероятность P(T|W). Предполагая, что связи (i,j) независимы друг от друга (очень сильное допущение), имеем

$$P(T|W) = \Pi P((i,j)|W) \quad (3)$$

где P((i,j)|W) является вероятностью зависимости (i,j) в конкретном предложении W. Вероятность P((i,j)|W) невозможно оценить непосредственно, поскольку мы предполагаем, что корпус не содержит, или содержит очень мало тождественных предложений. Поэтому в качестве приближения P((i,j)|W) берется вероятность P(i,j), которая зависит только от встречаемости слов wi, wj в предложениях корпуса и, возможно, от расстояния (j–i).

Вероятность P(i,j) оценивается как

$$P(i,j) = C(wi, wj, R) / C(wi, wj) \quad (4)$$

где C(wi,wj,R) – число обнаружений связи R между словами wi и wj в корпусе, а C(wi,wj) — число обнаружений слов wi и wj в одном и том же предложении корпуса.

Вероятность зависимости P(i,j) можно рассматривать как вес связи d(i,j), то есть, связь с более высокой вероятностью имеет больший вес.

Проблема разреженности данных решается использованием приближения, описанного в работе [3], а именно, используется следующая оценка:

$$d(i,j) = E = \lambda 1 E1 + (1 - \lambda 1)(\lambda 2 E23 + (1 - \lambda 2) E4) \quad (5)$$

где

$$E1 = CR1/C1;\ E23 = (CR2+CR3)/(C2+C3)\ E4 = CR4/C4$$
$$CR1 = C(wi, wj, R);\ C1 = C(wi, wj),$$
$$CR2 = C(wi, *, R);\ C2 = C(wi, *),$$
$$CR3 = C(*, wj, R);\ C3 = C(*, wj),$$
$$CR4 = C(*, *, R);\ C4 = C(*, *).$$

где * означает любое слово.

Параметры λ1 и λ2 лежат в диапазоне (0,1) и определяются экспериментально. Нами приняты значения, приведенные в работе [4], а именно λ1=0,7, λ2=0,3.

3.2. Алгоритм парсинга

Традиционные методы парсинга используют алгоритм динамического программирования, который требует $O(n^5)$ операций. Для парсеров с использованием биграмной модели лексических зависимостей разработаны $O(n^3)$ алгоритмы [8].

Приведенный ниже алгоритм строит проективное дерево T* над последовательностью вершин {1, …, n} за время $O(n^2)$ при заданных значениях d(i,j), он очень эффективен и прост в реализации.

Функция *склеить*(Uap, Ubq, i, j) удаляет юниты Uap, Ubq, создает на их месте новый юнит Ua(p+q+1) и закрывает в нем все вершины, лежащие в интервале между i и j.

Предлагаемый алгоритм парсинга (рис. 1) локальных зависимостей (ПЛЗ) требует $O(n^2)$ операций для разбора предложения длиной n слов. Докажем это утверждение.

ПАРСИНГ ЛОКАЛЬНЫХ ЗАВИСИМОСТЕЙ (W)

```
1  n = длина(W)
2  do while n>0
3  dmax = max d(i, j), где i, j есть открытые вершины смежных юнитов Uap, Ubq
4  поместить = (wi, wj) в T*
5  Ua(p+q+1) = склеить(Uap, Ubq, i, j))
6  n = n–1
7  end do
8  return(T)
```

Рис. 1. Алгоритм парсинга локальных зависимостей

На последнем шаге цикла требуется установить связь между двумя юнитами, покрывающими все предложение. Для этого требуется найти максимальную по весу связь между открытыми вершинами этих юнитов. В наихудшем случае юниты имеют равную длину (или их длины отличаются на 1) и все их вершины открыты. Для выбора максимальной связи потребуется выполнить n/2 * n/2 т.е. $n^2/4$ сравнений.

На предпоследнем шаге каждый из юнитов делится пополам и потребуется выполнить $2*n^2/16$ сравнений. На шаге n-i потребуется выполнить $2^i *(n^2/2^{2i}) = n^2/2^i$ сравнений.

Суммируя по i, получаем общее число сравнений для наихудшего случая разбора:

$$n^2 * \Sigma\, 1/2^i$$

Ряд сходится к 1, а общее число операций = $O(n^2)$

3.3. Создание корпуса для обучения

В этом разделе описаны два метода, которые использовались, чтобы разметить необработанный текстовый корпус для обучения МЛС:

(i) сбор статистики грамматических признаков n-грам, n=3, и

Грамматические признаки кодировались согласно Грамматическому словарю Зализняка. Морфологическая омонимия не снималась, вместо этого грамматические признаки омонимичной словоформы расщеплялись: если словоформе может быть приписано m различных грамматических кодировок, статистика каждой из этих кодировок увеличивалась на 1/m.

(ii) сбор статистики k-буквенных окончаний n-грам, k=4, n=5.

Поскольку русский язык относится к флективным языкам, статистика k-буквенных окончаний использовалась параллельно статистике грамматических признаков, а также для внутренней проверки метода.

Для сбора статистики использовалась часть коллекции www.lib.ru объемом около 2 Гбайт.

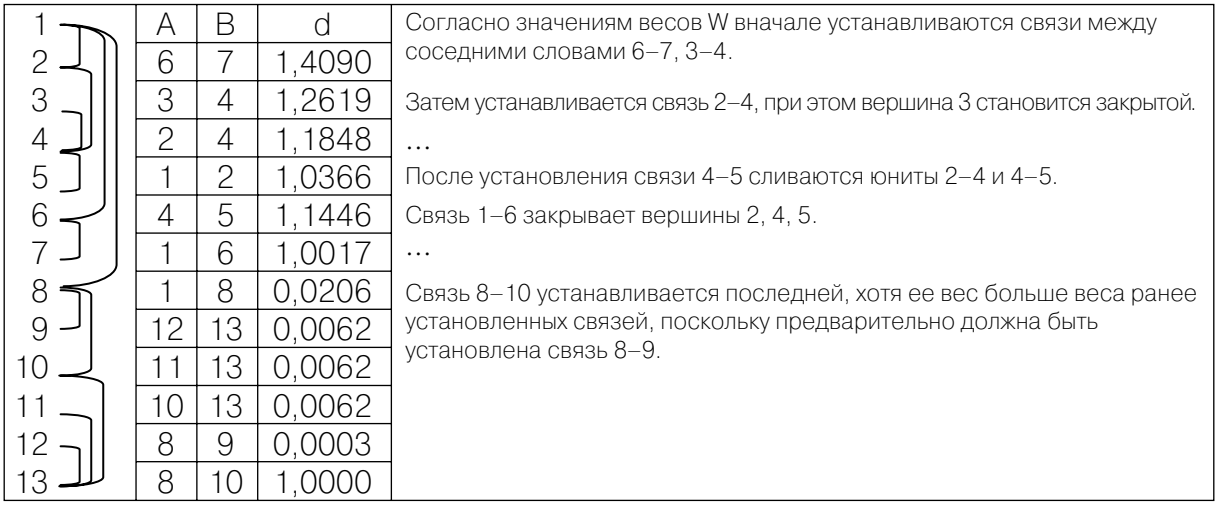

Рис. 2. Пример работы алгоритма

Итеративное обучение модели.

1. Каждое предложение тренировочного корпуса подвергается синтаксическому анализу согласно алгоритму Рис. 1. В качестве начальных значений веса зависимостей приняты величины $P(d(i,j)) = C(wi, wj, R) / C(wi, wj)$, $|i-j| < 5$ на основе собранной статистики (i) или (ii)

2. По результатам парсинга согласно (5) подсчитываются и записываются новые значения для E1, E23, E4 и E.

Выполняется парсинг каждого предложения с новыми значениями весов зависимостей.

Шаг 2 повторяется до тех пор, пока изменение весов зависимостей не станет меньше заданного порога.

3.4. Результаты экспериментов

В качестве исходного корпуса текстов выбрано собрание коротких рассказов А. П. Чехова объемом около 1 Мбайт. Общее число размеченных предложений — 12255 (исключены предложения длиной менее 3 слов). Средняя длина предложения — 15 слов. Самое длинное предложение состояло из 95 слов.

Все знаки препинания опускались.

Вид размеченного предложения представлен на Рис. 3 (рассказ «Драматург»).

Выводятся: слова предложения и номер слова, установленные связи и таблица «ЗАВИСИМОСТИ» с перечислением связей. Столбцы A и B содержат номера связанных вершин, d — вес связи (вычисленный по формуле 5), в правом столбце - флажок разрешения данной связи. Флажок позволяет исключать ложные связи в интерактивном режиме.

Приведенный пример представляет получение правильного разбора после небольшого числа итераций. Разбор большинства предложений, однако, содержит ложные зависимости, которые не устраняются после 10-й итерации.

Подсчет правильных и ложных связей выполняется обычно сравнением с «золотым стандартом», т.е. с корпусом безусловно правильно разобранных предложений. К сожалению, в настоящее время такой золотой стандарт для русского языка отсутствует в свободном доступе. Поэтому предполагается выполнить экспертную проверку деревьев разбора. При этом группе экспертов будет предложено вычеркивать ложные связи, не внося других исправлений. По результатам проверки можно будет выполнить оценку работы алгоритма, и, главное, внести изменения в веса связей, что позволит улучшить результаты анализа.

4. Заключение

Представлена модель локальных зависимостей, в которой имплицитно учтены лингвистические ограничения структуры предложения — вероятностные зависимости, которые выражают отношения между словами в предложении в виде ненаправленного проективного дерева, а также проективный характер предложения. Предложен новый алгоритм грамматического разбора, выполняющий поиск дерева зависимостей снизу вверх, устанавливая локальные связи между соседними словами и группами слов.

В цикле итерации после разбора всех предложений корпуса выполняется уточнение весов связей, затем вновь выполняется разбор и т.д.

Результаты проведенных экспериментов показывают, что результаты разбора улучшаются после нескольких итераций работы алгоритма, однако не для всех вариантов грамматической структуры и лексического состава предложений.

Имеется несколько возможностей для совершенствования модели.

В частности, при образовании очередного юнита, можно проверять, является ли он устойчивым или терминологическим словосочетанием, и обрабатывать его соответственно. Предполагается также эксплицитно включить в алгоритм проверку грамматических ограничений (напр., согласование существительного и прилагательного, запрет на связь предлога с более чем одним существительным и т.п.)

	A	B	d	
Доктор 1				
мгновенно 2	5	6	14,136	✔
проникается 3	3	4	1,7116	✔
уважением 4	8	9	1,0711	✔
к 5	2	3	1,0730	✔
пациенту 6	1	3	1,7056	✔
и 7	6	7	0,7046	✔
почтительно 8	4	6	0,7260	✔
улыбается 9	7	9	0,4719	✔
	—	—	—	**OK**

	A	B	d	
Доктор 1				
мгновенно 2	5	6	14,136	✔
проникается 3	3	4	1,7116	✔
уважением 4	8	9	1,0711	✔
к 5	2	3	1,0730	✔
пациенту 6	1	3	1,7056	✔
и 7	6	7	0,7046	✔
почтительно 8	4	6	0,7260	✔
улыбается 9	7	9	0,4719	✔
	—	—	—	**OK**

Рис. 3. Разметка предложения после 1-й и 4-й итерации

Далее, возможно преобразование ненаправленного дерева зависимостей в направленное — путем поочередного назначения каждой открытой вершины дерева (то есть, вершины, над которой не проходит ни одна связь) — корнем дерева, подсчета статистики образованных направленных связей и выбора наиболее вероятного варианта.

Поскольку модель локальных зависимостей применима к языку, основная часть предложений в котором — проективные, и благодаря высокой скорости парсинга, эту модель и алгоритм ПЛЗ можно использовать для языков с ограниченными лингвистическими ресурсами, даже в отсутствии морфологического анализатора.

Литература

1. *Billot S., Lang B.* The Structure of Shared Forests in Ambiguous Parsing // Proceedings ACL 1989.
2. *Bod R.* An all-subtrees approach to unsupervised parsing // Proceedings of COLINGACL
3. *Collins M., Hajic J., Brill E., Ramshaw L., Tillmann C.* A statistical parser for czech // Proceedings of the 37th Meeting of the Association for Computational Linguistics (ACL), pp. 505–512
4. *Gao J., Suzuki H.* Unsupervised learning of dependency structure for language modeling // ACL 2003, pp. 521–528.
5. *Goodman J.* Efficient algorithms for parsing the DOP model // Proceedings Empirical Methods in Natural Language Processing 1996, Philadelphia, PA: 143–152.
6. *McDonald R., Satta G.* On the complexity of non-projective data-driven dependency parsing // Proceedings of the International Conference on Parsing Technologies (IWPT)
7. *Nivre J.* An efficient algorithm for projective dependency parsing. // Proceedings of International Workshop on Parsing Technologies, pp. 149–160
8. *Smith D.A., Eisner J.* Bootstrapping feature-rich dependency parsers with entropic priors // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 667–677
9. *Seginer Y.* Fast Unsupervised Incremental Parsing // Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 384–391, Prague, Czech Republic, June 2007.
10. *Ножов И. М.* Реализация автоматической синтаксической сегментации русского предложения // Дисс. Канд. Техн. Наук. — М.: РГГУ, 2003
11. *Mel'čuk I.* (1988) Dependency Syntax: Theory and Practice // Albany, N.Y.: The SUNY Press, 428 pages.
12. *Ginter F., Pyysalo S., Boberg J., Salakoski T.* Regular Approximation of Link Grammar // T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 564–575, 2006.