Синтаксический анализ текстов естественного языка: симбиоз формально-грамматических и вероятностных методов

Михайлов Д. В.

Новгородский государственный университет
имени Ярослава Мудрого

Дополнительные разделы к учебному курсу
«Системы искусственного интеллекта»

по направлению
230100.62 — Информатика и вычислительная техника

2015 г.

С чего начать изучение данного раздела ?

1. *Турдаков Д. Ю.* Основы обработки текстов: спецкурс для студентов ВМК МГУ. Лекция 2: Регулярные выражения и конечные автоматы / Д. Ю. Турдаков. Режим доступа: прямая ссылка.
2. *Турдаков Д. Ю.* Основы обработки текстов: спецкурс для студентов ВМК МГУ. Лекция 4: Статистические методы поиска словосочетаний / Д. Ю. Турдаков. Режим доступа: прямая ссылка.
3. *Турдаков Д. Ю.* Основы обработки текстов: спецкурс для студентов ВМК МГУ. Лекция 6: Формальные грамматики и синтаксический анализ / Д. Ю. Турдаков. Режим доступа: прямая ссылка.
4. *Ветров Д. П.* Курс «Графические модели». Лекция 3: Скрытые марковские модели / Д. П. Ветров. Режим доступа: прямая ссылка.

Основные идеи

- использование стохастического островного табличного анализатора;
- синтаксический разбор предшествует процессу «создания мостов» для идентификации исходных «островов» с целью объединить смежные анализы;
- использование покрывающей грамматики, в множестве правил которой выделяют предсказывающее, проецирующее и распространяющее множество.

Использование контекстно-свободных грамматик

- порождение регулярной аппроксимации исходной грамматики и генерация языка, который может быть либо подмножеством, либо надмножеством языка, порождаемого исходной грамматикой;
- разделение процесса разбора на последовательность более простых шагов, каждый из которых покрывается более простой (в общем случае регулярной) грамматикой;
- управление процессом разбора с помощью конечного числа эвристик, в общем случае основанных на стохастических моделях.

Сильные и слабые стороны контекстно-свободных грамматик

Наиболее актуальные проблемы

- в случае регулярной апроксимации исходной грамматики мы получаем лишь апроксимацию исходного языка, что недостаточно для некоторых приложений, задействующих расстояние между языком, порождаемым исходной грамматикой, и его апроксимацией;
- для разбиения процесса разбора на шаги исходная грамматика должна быть структурируема на каскад более простых.

Ключевое требование

Независимость от используемого источника знаний и языка.

Вариант решения

Использование вероятностных контекстно-свободных грамматик (probabilistic context-free grammars, PCFGs) для назначения вероятностей составляющим и использование этих вероятностей как отправных пунктов вычисления показателя качества.

Два шага разбора

1. «Дробление» (англ. chunking) — использование грамматики «кусков» (chunks), автоматически извлекаемых из исходной. На этом шаге выполняется частичный разбор «входной» грамматики.
2. «Островной» анализ — здесь запускается двунаправленный разбор, начиная от «островов», отождествляемых с предварительно выделенными «кусками».

В роли «островов» могут быть:

- неомонимичные слова;
- основные (базовые) именные группы — в случае текстового ввода;
- точно распознанные фрагменты — для звучащей речи.

Замечание

Chunking есть поиск синтаксически связанных групп соседних слов.

Базовое предположение

Показанное выше разделение процесса разбора позволяет использовать любую полнопокрывающую стохастическую контекстно-свободную грамматику (stochastic context-free grammar, SCFG) в качестве входа.

Традиционный фрагментарный разбор «слева направо» в островном анализе имеет две отличительные особенности:

- двунаправленность — разбор может идти как слева направо, так и справа налево;
- динамически определяемые позиции начала процесса анализа в предложении (т. е. сами острова).

Используемые при этом эвристики основаны на двух стохастических моделях: *локальной* и *граничной*, которые позволяют выбрать наиболее вероятный остров для расширения в наиболее вероятном направлении.

«Локальный» подход основан на предположении о вероятности расширения границы (и предсказания) как вероятности следующего символа на расширение при наличии терминального символа (символов) в соответствующей позиции предложения от правого/левого конца.

Рассматриваемый далее вариант разбора воплощает комбинацию расширения «снизу вверх» и прогноз «сверху вниз», управляемую стохастическими параметрами.

- *локальная модель* — статична, рассматривает только грамматическую информацию;
- *граничная (neighbouring)* — рассматривает также непосредственное окружение каждого «острова» — соседние острова и сегменты входного предложения, отвечающие промежуткам между рассматриваемым островом и его «соседями».

При этом с помощью $SCFG$-грамматики моделируется наиболее вероятное расширение (вправо или влево) контура (активного или неактивного), а также частичный разбор и рост островов «наибольшей уверенности».

Будем далее в формулах при обозначении контуров использовать *двойную точечную нотацию*.

Пусть

G есть $SCFG$-грамматика, T — множество её терминальных, N — нетерминальных символов, R_i — i-я продукция грамматики G, $P(R_i)$ — вероятность её присоединения;

$[A, i, j]$ — остров категории A, охватывающий позиции с i-й по j-ю.

$P(A >> a/G)$ — вероятность того, что, начавшись с нетерминала A, успешное применение правил грамматики G порождает последовательность, начинающуюся с терминала a.

$\{l|r\}\,c\colon N \times T \to [0,1]$ и, соответственно, $\{l|r\}\,c^*\colon N \times T^* \to [0,1]$, причём $\{l|r\}\,c(A,a)$ есть вероятность того, что порождаемое при разборе дерево с корнем в A может иметь символ a крайним правым/левым:

$$\forall\, A \in N, a \in T\colon rc(A,a) = P(A >> a/G).$$

Аналогично для списка символов la

$$rc^*(A,la) = \sum_{a \in la} rc(A,a).$$

$lc^*(A,la)$ и $lc(A,a)$ определяются симметрично.

Пусть

lt — список терминальных символов категорий слова w_{i-1},
R_i — i-я продукция грамматики G.

Тогда вероятности расширения влево:

- острова A до слова w (оно здесь расссматривается как элемент цепочки, контур неактивен)

$$P_{island}^{left}\left([A,i,j]\,/G,w\right) = \sum_{R_i:\ X\to\alpha A} P\left(R_i\right);$$

- активного контура (предсказания слева)

$$P_{arc}^{left}\left([A\to\alpha B.\beta.\gamma,i,j]\,/G,w\right) = rc^{*}\left(B,lt\right).$$

Расширения и предсказание вправо определяются симметрично.

Замечание

Вышеуказанные правила учитывают возможность пустой цепочки α или β.

Значения $\{l|r\}\, c$ и $\{l|r\}\, c^*$ являются предварительно вычисляемыми вероятностями и сохраняются в двух таблицах, именуемыми далее $Lreachability$ и $Rreachability$, соответственно.

Вычисление указанных таблиц — далеко не тривиальная задача.

Известный подход на рекурсивных грамматиках, расширенных с учётом двунаправленности разбора, предполагает учёт взаимозависимостей терминалов и нетерминалов посредством системы линейных уравнений.

Проблема — неограниченный рост размера грамматик.

В силу сказанного процесс вычисления таблиц достижимости включает следующие шаги:

1. вычисление сильносвязанных компонент;
2. решение системы линейных уравнений для каждой компоненты;
3. совершенствование алгоритма для комбинации полученных результатов.

Здесь решение о расширении острова принимается на основе информации от соседей и расстояния до них (длины «разрывов» — сегментов исходных предложений между смежными островами).

Иными словами, при принятии решения здесь руководствуются моделью расстояний, измеряемых числом терминальных символов между узлами в дереве разбора.

Поэтому вероятности распределений длин для каждого правила грамматики должны быть получены на основе тренировочного корпуса.

Граничная модель: расстояние между островами

Даны два острова $[A, i, j]$ и $[B, j + d, l]$, разделённые расстоянием d, причём между островами существуют три типа отношений:

$$R_1 = \{r\colon X \to \alpha A \beta B \gamma,\ d = |\beta|\};$$
$$R_2 = \{r\colon X \to \alpha A \beta H \gamma,\ H \xrightarrow{*} \delta B \mu,\ d = |\beta| + |\delta|\};$$
$$R_3 = \{r\colon X \to \alpha H \beta B \gamma,\ H \xrightarrow{*} \delta A \mu,\ d = |\mu| + |\beta|\}.$$

Замечание

Эти случаи отношений составляют лишь те ситуации, в которых есть минимум одно правило, включающее непосредственно минимум один из рассматриваемых островов согласно принятому нами определению соседства. По этой причине для получения полного покрытия (исходной) грамматики здесь нужен другой метод.

Обозначим далее каждую из вероятностей для $i = 1, \ldots, 3$:

$P^i\,(d/r, A, B)$ и $P^i_{acc}\,(d/A, B) = \sum_{r \in R_i} P^i\,(d/r, A, B)$, соответственно.

Эти вероятности вычисляются предварительно для каждой потенциальной пары островов, с расстоянием, принимающим значения от 0 до некоторого предела (далее возьмём его равным 3).

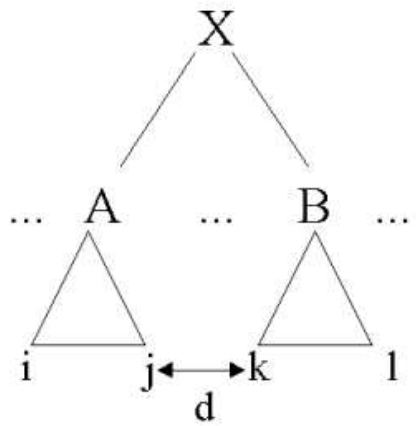


Рис. 1: Отношение R_1

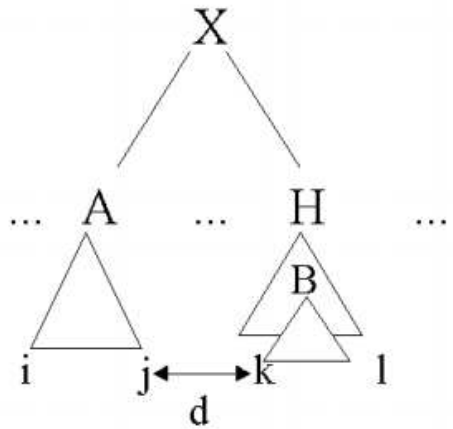


Рис. 2: Отношение R_2

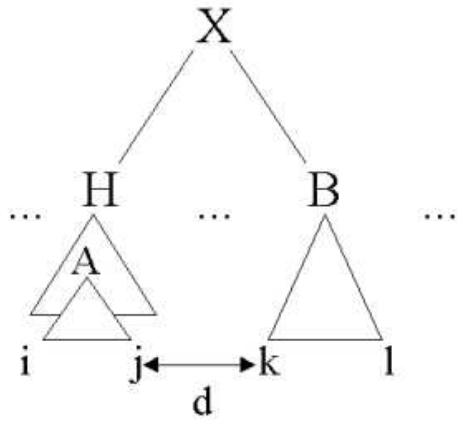


Рис. 3: Отношение R_3

- $[B, j+d, k]$ — ближайший остров справа для расширения $[A, i, j]$:

$$P_{island}^{right}\left([A,i,j]/G, w, [B, j+d, k]\right) = P_{acc}^{1}\left(d/A, B\right) + \sum_{Z \in N} \sum_{l=0}^{\min(3,d)} P_{acc}^{1}\left(d-l/A, Z\right) \cdot P_{acc}^{2}\left(l/Z, B\right) + P_{acc}^{1}\left(d-l/Z, B\right) \cdot P_{acc}^{3}\left(l/Z, A\right),$$

где N — множество нетерминалов грамматики G, $A, B \in N$.

Расширение влево определяется симметрично.

Первое слагаемое определяет случаи A и B в правой части одного и того же правила, второе — предусматривает варианты порождения B за один и более шагов из Z, который в правой части правила присутствует как A, плюс варианты A, порождаемого из Z в той же правой части как B.

Граничная модель: предсказание («активный контур»)

Идея близка расширению острова, но конкретизирует правило «активного контура».

- для расширения активного (т. е. «прогнозирующего») контура $[A \to \beta.A_l \alpha A_r.\gamma, i, j]$ пусть $[B, j+d, k]$ будет ближайшим островом справа Тогда

$$P_{edge}^{right}\left([A \to \beta.A_l \alpha A_r.\gamma, i, j]\,/G, w, [B, j+d, k]\right) = \\ = P_{acc}^{1}\left(d/r, A_r, B\right) + \\ + \sum_{\gamma_i \in N,\, i \le d,\, 0 \le d-l \le 3} \mathrm{prob}\Big(|\gamma_1 \ldots \gamma_{i-1}| = l\Big) \cdot P_{acc}^{2}\left(d - l/\gamma_i, B\right),$$

  где функция prob — рекурсивная, учитывая степень «обученности» грамматики G, задаёт распределение вероятностей длин любой подпоследовательности терминальных и нетерминальных символов.

Формула предсказания влево симметрична показанной выше для P_{arc}^{right}, различие состоит в использовании P_{acc}^{3} вместо P_{acc}^{2} (см. предыдущий слайд), а также A_l вместо A_r.

Граничная модель: некоторые замечания по использованию эвристик

1. При начальном задании прогнозирующих контуров вероятности, используемые локальной моделью, выступают в роли фильтра (т. е. только при ненулевой «локальной» на основе «граничной» вероятности можно определять, где и когда использовать контур); для расстояний $d > 2$, «локальный» подход используется напрямую.
2. Последовательное рекурсивное предсказание руководствуется только «локальными» вероятностями, ограниченными, в свою очередь, пороговым значением. Этот порог устанавливается эмпирически.

Замечание

С целтю снижения вычислительных затрат вышеупомянутые вероятности рассчитываются предварительно с использованием частот распределений длин, полученных из тренировочного корпуса. Эти данные записываются в две таблицы, содержащие вероятности каждой пары категорий от 0 до некоторого предела $limit$, а также единственный случай расстояний, превышающих $limit$. Здесь же вычисляются и более простые таблицы для случаев расширения/предсказания от первого острова предложения.

- процент предложений, покрытых островами;
- число циклов расширения/предсказания.

Замечание

Возможности управления расширением контура у «граничного» подхода выше, но в силу разреженности данных многим потенциально возможным случаям присваивается нулевая вероятность и их «предсказательный» приоритет оказывается ниже ожидаемого. Кроме того, когда необходим откат к локальной модели, все лексические контуры (а не только острова) должны быть представлены заново в расширенной структуре $heap$ с целью уверенности в возможности получения полного покрытия. В некоторых случаях это указывает на неправильно выбранный контур.

- слова, имеющие в заданном естественном языке не более одного значения, рассматривать в роли начальных островов;
- использовать в качестве островов-кандидатов именные группы (ИГ, группы существительного, noun phrases, NP);
- в рассмотрение брать только те именные группы, которые могут управлять ходом остроного анализа;
- отбираемые именные группы не должны быть рекурсивными, т. е. не содержат подчинённых именных групп.

Алгоритм извлечения грамматики для выделения именных групп (базовой NP-грамматики)

1. из исходной грамматики выбираются те правила, в которых левая часть отвечает нерекурсивной именной группе;
2. частичный разбор с выделением всех возможных базовых именных групп для каждого предложения входного корпуса. В сравнительных целях части речи у слов тестового множества неоднозначны и, таким образом, в конечном итоге многие найденные именные группы могут быть некорректны;
3. отбор найденных фрагментов (chunks) согласно их типам.

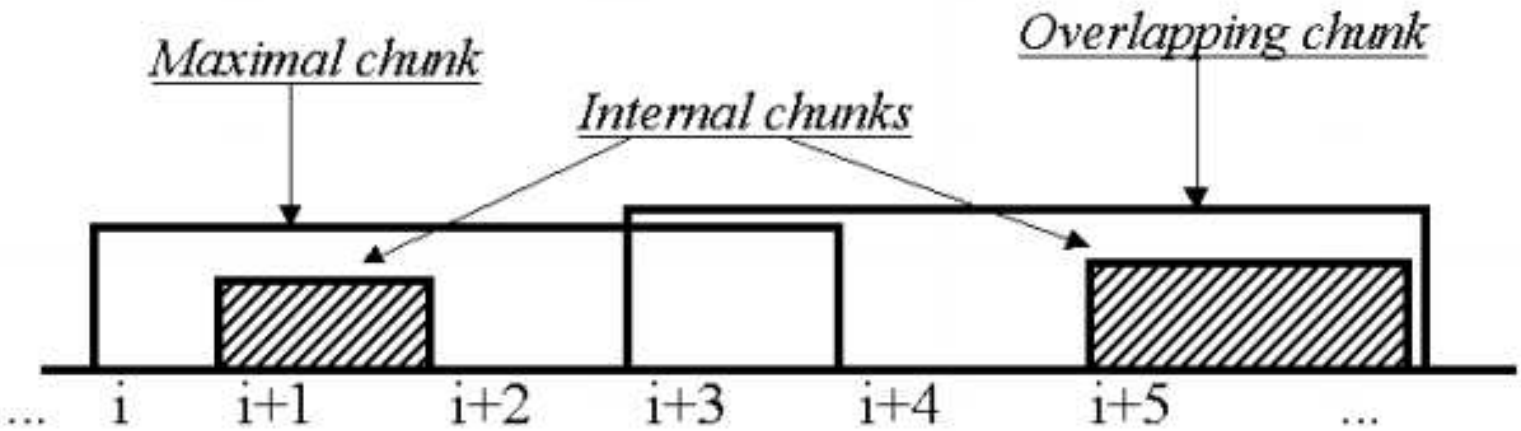
Замечание

Будем далее для вышеуказанных синтаксически связанных фрагментов текста (chunks) использовать сокращение СФТ.

Комментарий

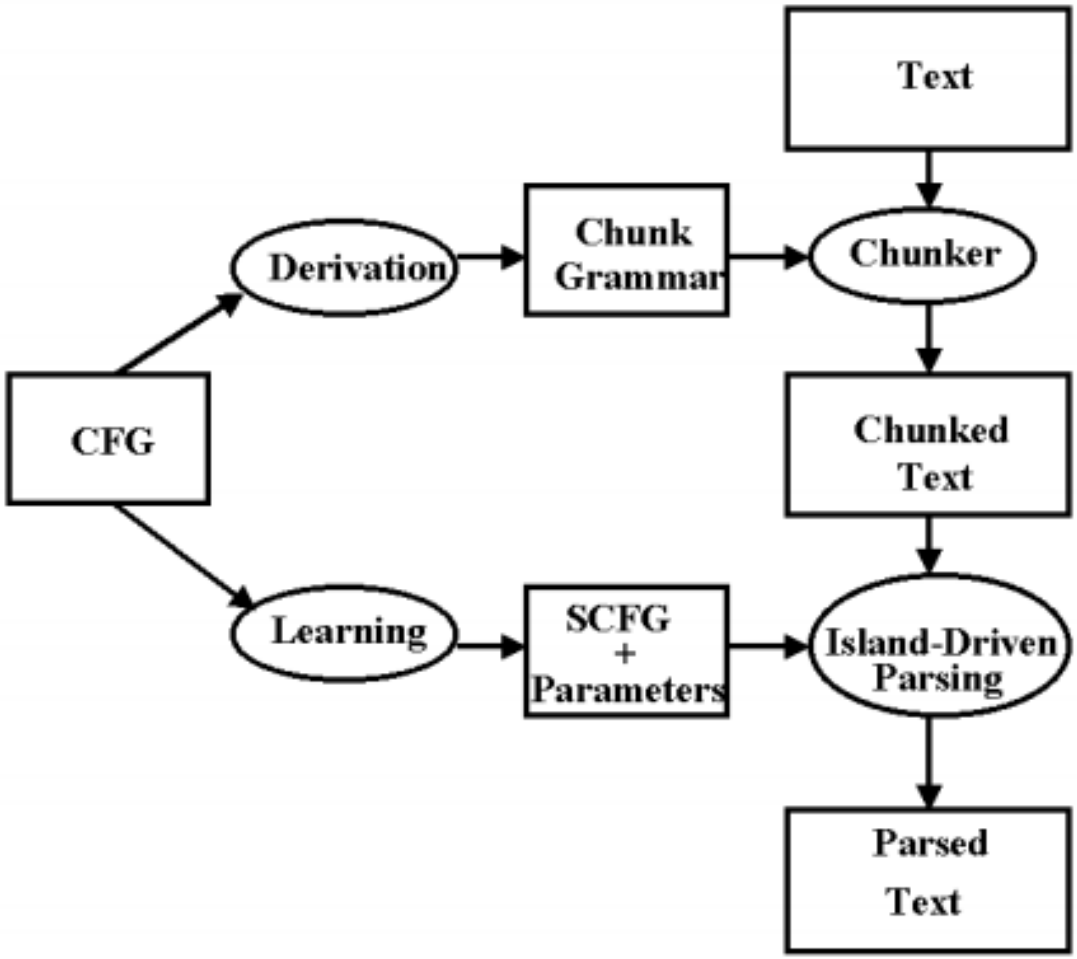
В первом приближении СФТ отвечает вершине синтаксического дерева с зависимыми, соседними в линейном ряду, причём в полученном поддереве не должно быть иных вершин той же части речи, что и корень. Пример СФТ: *green colorless thoughts*, контр-пример: *a boy kissing Mary*.

Пример реализации вероятностного островного подхода
Три найденных типа базовых именных групп



- *maximal NPs* — самые длинные из тех, которые начинаются с определённой позиции в предложении;
- *overlapping NPs* — те, которые частично покрывают предыдущие *maximal NPs*;
- *internal NPs* — остальные именные группы из найденных.

Пример реализации вероятностного островного подхода
Разбор в два этапа



BIO chunking

Задача

Идентификация именованных сущностей выделением индивидуальных токенов многословных обозначений (*Bill Gates*, *Hewlett-Packard* и т. п.).

Идея

Помечать каждый токен либо как начальный (*англ. Beginning*, **B**), либо как внутренний (*Inside*, **I**), либо как наружный (*Outside*, **O**) по отношению к текстовому фрагменту, представляющему интерес.

Исходное предложение: «*Minjun is from South Korea*».
Аннотирование:

- *Minjun* **IS_A** «Person»
- *South Korea* **IS_A** «Location»

Классификатор, реализующий BIO chunking, выделил бы следующие аннотации уровня токенов:

I-Person	**O**	**O**	**B-Location**	**I-Location**	**O**
Minjun	*is*	*from*	*South*	*Korea*	.

Данная мера определяется на основе значений точности (precision) и полноты (recall) модели.

Точность в пределах заданного класса равна отношению числа верно идентифицированных объектов класса к общему числу объектов, отнесённых моделью к этому классу.

Полнота равна отношению числа верно идентифицированных объектов заданного класса к числу объектов заданного класса в тестовой выборке.

F_1 мера (её также называют традиционной F-мерой) содержательно есть среднее гармоническое точности и полноты:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Определение 1

Базовая именная группа есть последовательность слов в предложении S:

- она является подпоследовательностью некоторой именной группы в дереве непосредственных составляющих;
- в выводе данной подпоследовательности в грамматике G нетерминальный символ NP, соответствующий вершине именной группы, встречается единственный раз.

Замечание

В английском языке определения имеют тенденцию помещаться слева в линейном порядке от вершины дерева, например, *USA Supreme Court*. В русском аналогичные конструкции оформляются родительным падежом, ср. *Верховный суд Российской Федерации*.

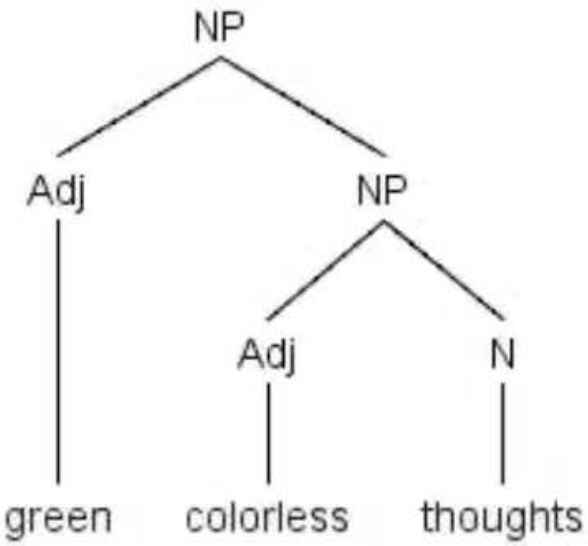


Рис. 1: Именная группа, являющаяся базовой именной группой

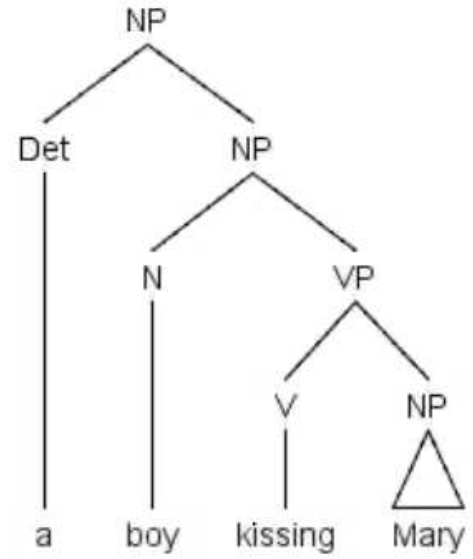


Рис. 2: Именная группа, не являющаяся базовой (нарушено условие проективности)

Построение базовой именной группы длиной более одного слова для дерева на рис. 2 невозможно.

Рис. 1: Дерево зависимостей и границы базовых именных групп для проективного предложения

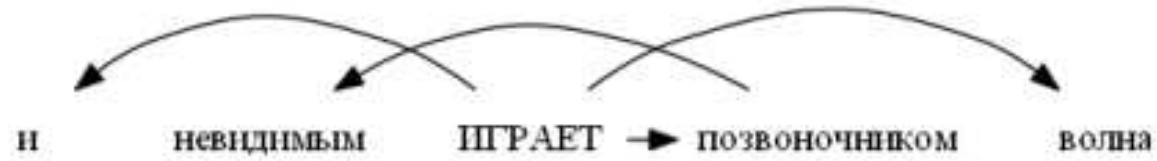


Рис. 2: Дерево зависимостей для непроективного предложения

Определение 2

Последовательность слов в предложении S *на русском языке* есть *базовая именная группа* при выполнении двух условий:

- она является подпоследовательностью некоторой именной группы в дереве непосредственных составляющих;
- в дереве непосредственных составляющих для данной подпоследовательности существует не более одной именой группы, в т. ч. вершина данной группы стоит не в родительном падеже.

Замечание

Данное определение учитывает два наиболее важных случая:

- синтаксической зависимости между двумя существительными;
- употребление родительного падежа в отрицательных предложениях.

Базовое предположение

Большинство предложений в русском литературном языке проективны.

Замечание

Проективность предложения не гарантирует сохранение синтаксических групп.

Следствие

В русских предложениях должна прослеживаться тенденция к сохранению базовых именных групп.

Шаг 1. Каждое синтаксическое дерево обрабатывается алгоритмом, который восстанавливает порядок слов, соответствующий сильно проективной конструкции.

Шаг 2. Просмотр вручную с целью выявления причин нарушения исходного порядка слов.

Пример: эксперимент с 500-ми фразами из интервью в Internet-изданиях:

- 17 случаев изменений вызвана ошибками в синтаксическом разборе;
- 6 — слабопроективные конструкции с составным глагольным сказуемым (например, *захотел пойти*);
- 3 — непроективные конструкции, среди которых дважды встретилась конструкция *друг к другу*, разобранная как полноценное дерево.

Морфологический анализ — TreeTagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

Лемматизация — CSTlemma, http://corpus.leeds.ac.uk/mocky/.

Синтаксический анализ — Malt Parser, http://www.dialog-21.ru/digests/dialog2011/materials/en/pdf/58.pdf.

Условные случайные поля: некоторые базовые определения

Определение 3

Простой граф — граф, в котором нет кратных рёбер и петель.

Определение 4

Кратные рёбра — рёбра, инцидентные одной и той же паре вершин.

Определение 5

Полный граф — простой граф, где каждая пара различных вершин смежна.

Определение 6

Кликой в неориентированном графе называется подмножество его вершин, таких, что любые две вершины подмножества соединены ребром.

Определение 7

Случайные величины x и y называются условно независимыми от z, если

$$p(x, y|z) = p(x|z)\, p(y|z),$$

т. е. вся информация о взаимозависимостях между x и y содержится в z.

Условные случайные поля: некоторые базовые определения

Определение 8

Конкретное значение, которое случайная величина (СВ) принимает в результате испытания, называется её реализацией.

Пусть $\mathcal{A}$ — конечное множество индексов, $V = \{V_i | i \in \mathcal{A}\}$ — многомерная СВ, где каждая компонента V_i есть одномерная СВ, принимающая значение v_i и определенная в некотором вероятностном пространстве.

Обозначим реализацию V как $v \in \Omega$, где Ω — множество всех возможных конфигураций.

Определение 9

Введённая вышеописанным образом случайная величина V называется *случайным полем (СП)*, англ. *Random Field*.

Замечание

Для удобства будем считать, что $\forall\ V_i \in V$ дискретна, а множество её значений конечно. При этом $\mathcal{A}$ задаёт множество точек на плоскости и, соответственно, рассматривается реализация V в этих точках.

Пусть

V — случайное поле,
E — множество рёбер, отражающее все зависимости между $V_i \in V$.

Определение 10

Множество соседей для V_i есть множество смежных с V_i вершин:

$$\partial i = \{j \in \mathcal{A} \colon j \neq i, (i,j) \in E\}.$$

Следствие

Многомерная случайная величина V и система зависимостей её компонент образуют ненаправленный граф $G = (V, E)$.

Определение 11

Пусть $G = (V, E)$ — неориентированный граф с множеством вершин V и множеством рёбер E. Набор случайных величин $V_i \in V$, $i \in \mathcal{A}$, образует *марковское случайное поле* по отношениею к G, если:

1. для $\forall\ v \in \Omega\ P(V = v) > 0$;
2. $P\left(V_i = v_i | V_j = v_j, j \in \mathcal{A} \backslash \{i\}\right) = P\left(V_i = v_i | V_j = v_j, j \in \partial i\right)$.

Пусть c — клика в G, а $v_c = \left(v_{i_1}, \ldots, v_{i_{|c|}}\right)$ — ограничение реализации v на c, где $\forall\ i_j \in c$. Обозначим множество клик графа $G = (V, E)$ как $C\left(G\right)$ и определим функцию-фактор $\Psi_c\left(v_c\right)$ как некоторую функцию $\Psi_c \colon C \to \mathbb{R}_+$.

Определение 12

Дискретное распределение есть *распределение Гиббса*, если

$$P(V = v) = \frac{1}{Z} \prod_{c \in C(G)} \Psi_c\left(v_c\right),$$

где Z — нормирующая константа (т. н. статистическая сумма):

$$Z = \sum_{v \in \Omega} \prod_{c \in C(G)} \Psi_c\left(v_c\right).$$

Теорема 1 (Хаммерслей-Клиффорд)

V является марковским случайным полем, соответствующим $G = (V, E)$, тогда и только тогда, когда $P(V = v)$ — распределение Гиббса.

Определение 13

Условное случайное поле (УСП, *англ. conditional random field, CRF*) есть марковское случайное поле, у которого множество вершин $V = X \cup Y$ разбито на два непересекающихся подмножества: X — множество наблюдаемых переменных и Y — множество скрытых переменных.

Условные случайные поля: задача предсказания

Пусть $\mathbf{x} = \{v \in X\}$, а $\mathbf{y} = \{v \in Y\}$, где X и Y есть множества наблюдаемых и скрытых переменных, соответственно.

Будем предполагать, что значения СВ из $\mathbf{x}$ и $\mathbf{y}$ принадлежат некоторым конечным пространствам $\mathcal{X}^{|\mathbf{x}|}$ и $\mathcal{Y}^{|\mathbf{y}|}$, соответственно.

Задача предсказания

Оптимальным образом восстановить значения $\mathbf{y}$ по наблюдаемым $\mathbf{x}$. Согласно теореме Хаммерслея-Клиффорда здесь нужно максимизировать

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{c \in C} \Psi_c(\mathbf{x}, \mathbf{y}),$$

где $Z(\mathbf{x}) = \sum_{\mathbf{y}' \in \mathcal{Y}} \prod_{c \in C} \Psi_c(\mathbf{x}, \mathbf{y}')$ — статистическая сумма.

Проблема

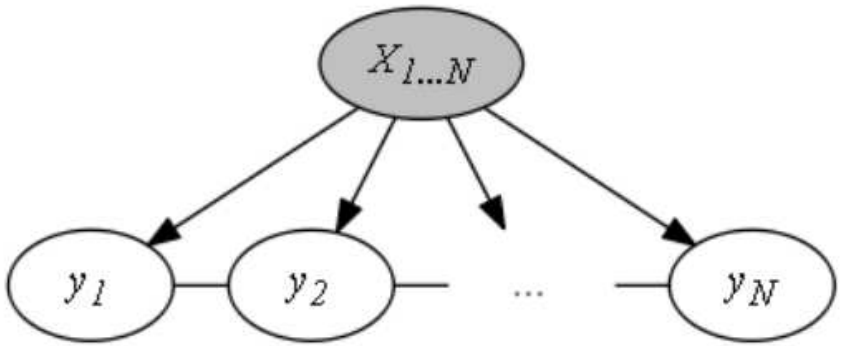
Экспоненциальный рост числа слагаемых в $Z(\mathbf{x})$ по размеру $\mathbf{x}$.

Определение 14

Линейно-цепочечное УСП (англ. *linear-chain CRF*) — это УСП, у которого множество скрытых переменных образует цепочку, т. е.

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{h}} p(\mathbf{y}|\mathbf{h}, \mathbf{x})\, p(\mathbf{h}|\mathbf{x}).$$

Будем считать $\mathbf{x}$ последовательностью слов предложения, $\mathbf{y}$ — меток BIO. Соответствующее УСП выглядит следующим образом:



Здесь имеем 2 типа факторов (соответствут дугам): одиночные и парные.

Одиночный фактор $\varphi_t^1(y_t, \mathbf{x})$ задаёт влияние, оказываемое наблюдаемой последовательностью $\mathbf{x}$ на метку y_t. Аргумент $\mathbf{x}$ указывает на то, что при вычислении φ_t^1 могут использоваться признаки любых элементов $\mathbf{x}$.

Пример признака: *«слово x_t является существительным»*.

Замечание

Значение φ_t^1 является функцией от y_t, поскольку значение $\mathbf{x}$ постоянно.

Парный фактор $\varphi_t^2(y_t, y_{t-1})$ задаёт влияние соседних меток друг на друга.

Значение $\varphi_t^2(y_t, y_{t-1})$ пропорционально *вероятности* $\mathrm{P}(y_t, y_{t-1}|X_{1\ldots N})$, т. е. вероятности *участия* этих *скрытых состояний* в генерации $\mathbf{x}$.

В рассматриваемом УСП значения в каждом из φ_t^1 вычисляется только на основе признаков текущего наблюдения, поэтому поиск наиболее вероятного списка состояний здесь выполняет *алгоритм Витерби*.

Условные случайные поля: основная идея алгоритма Витерби

Пусть имеется *скрытая марковская модель* с пространством состояний S, вероятностями π_i нахождения в i-м, а также вероятностями $a_{i,j}$ перехода из i-го в j-е состояние; $y_1 \dots y_T$ — наблюдаемые символы на выходе.

Тогда наиболее вероятная последовательность состояний (НВПС) $x_1 \dots x_T$ (*путь Витерби*) задаётся рекуррентными соотношениями:

$$V_{1,k} = \mathrm{P}(y_1|k) \cdot \pi_k$$
$$V_{t,k} = \mathrm{P}(y_t|k) \cdot \max_{x \in S} \left(a_{x,k} \cdot V_{t-1,x}\right),$$

где $V_{t,k}$ — вероятность НВПС, ответственной за появление первых t символов и завершающейся состоянием k.

Замечание

Путь Витерби находится по указателям для x во втором уравнении.

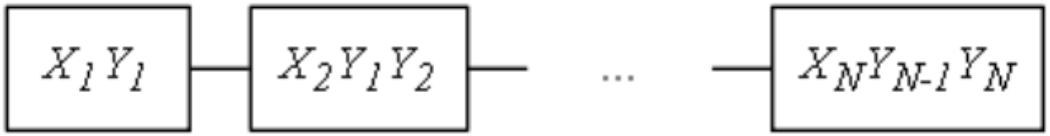
Пусть $\mathrm{Ptr}(k,t)$ — функция, которая при $t > 1$ возвращает значение x, использованное для подсчёта $V_{t,k}$, а при $t = 1$ выдаёт k. Тогда

$$x_T = \arg\max_{x \in S} \left(V_{T,x}\right), \; x_{t-1} = \mathrm{Ptr}(x_t, t).$$

Сложность алгоритма Витерби равна $O\left(T \cdot |S|^2\right)$.

Условные случайные поля: фактор кластера

Для рассматриваемого УСП кластерное дерево примет следующий вид:



Замечание

Содержательно фактор кластера $\psi_t\left(y_t, y_{t-1}, \mathbf{x}\right)$ представляет собой произведение входящих в него факторов.

Для вычисления ψ_t по наблюдаемым признакам вводятся индикаторные функции признаков f_k и соответствующие веса λ_k, а искомая условная вероятность

$$p\left(\mathbf{y}|\mathbf{x}\right) = \frac{1}{Z\left(\mathbf{x}\right)} \exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k\left(y_t, y_{t-1}, \mathbf{x}\right)\right\}.$$

На этапе обучения осуществляется подбор весов, на которых достигается максимум правдоподобия обучающей выборки.

Само обучение — с помощью различных вариаций градиентного подъёма. Один из наиболее эффективных вариантов здесь — алгоритм L-BFGS.

Здесь следует использовать обучающие выборки двух типов:

- полученная обработкой неразмеченного текста из корпуса (например, OpenCorpora) морфологическим и синтаксическим анализатором (TreeTagger, CSTlemma, Malt Parser). В результате имеем исходную выборку с синтаксической разметкой в форме дерева зависимостей;
- фрагмент синтаксически аннотированного корпуса, использующего дерево зависимостей в качестве синтаксической аннотации (пример: SynTagRus ИППИ РАН).

Далее в каждом синтаксическом дереве выделяются базовые ИГ, первое слово группы получает метку **B**, последующие — **I**, слова из не вошедших ни в одну из базовых ИГ — **O**. Т. о. обе выборки получают BIO-разметку.

Алгоритм извлечения базовых ИГ можно организовать как простой обход дерева вглубь с объединением в одну базовую ИГ поддеревьев таких, что:

1. все узлы поддерева представляют собой подпоследовательность в линейном порядке слов без разрывов;
2. вершиной каждого поддерева является слово-существительное в любом падеже;
3. в остальных узлах поддерева могут присутствовать только слова со следующими характеристиками:
   - существительное в родительном падеже;
   - прилагательное или порядковое числительное в любом падеже;
   - наречие;
4. в подпоследовательности отсутствуют знаки препинания.

- *Работа с условными случайными полями* — библиотека MALLET (Java), http://mallet.cs.umass.edu.
- Признаки элементов последовательности $\mathbf{x}$ — это признаки токенов: часть речи, падеж, число, род, заглавная буква.
- Признак «слово начинается с заглавной буквы» — для более точной обработки имён и должностей: *Президент Российской Федерации*.
- Знаки пунктуации рассматриваются как самостоятельная часть речи.
- Для каждого токена в число признаков включаются признаки правого и левого соседа, а также все конъюнкции признаков текущего и каждого из соседних токенов.

Заключение

1. Применение условных случайных полей для поиска синтаксически связанных групп соседних слов наиболее эффективно в задачах текстового анализа, не требующих полного синтаксического разбора.
2. Необходимым условием применения данного метода является формальное определение базовой синтаксической группы, которое позволило бы находить максимально длинные нерекурсивные фрагменты текста.
3. Достоинство — нетребовательность к качеству обучающей выборки.
4. Недостаток — низкая точность выделения фрагментов, содержащих предлог и/или союз, например:
   [Комиссия ООН] по [правам человека]
   вместо
   [Комиссия ООН по правам человека].
5. Внесением поправок, зависящих от конкретного языка со свободным порядком слов, качество работы метода на основе УСП сравнимо с его результатами для языков с более строгим порядком слов.

Использованные источники

1. *Ageno A.* Chunking + Island-Driven Parsing = Full Parsing / A. Ageno, H. Rodriguez. Режим доступа: прямая ссылка.
2. *Ageno A.* Probabilistic modelling of island-driven parsing / A. Ageno, H. Rodriguez. Режим доступа: прямая ссылка.
3. *Ramshaw L. A.* Text Chunking using Transformation-Based Learning / L. A. Ramshaw, M. P. Marcus. Режим доступа: прямая ссылка.
4. ClearTK: Machine Learning for UIMA. Tutorial named entity chunking classifier [Электронный ресурс]. Режим доступа: прямая ссылка.
5. *Антонова А. Ю.* Использование метода условных случайных полей для обработки текстов на русском языке / А. Ю. Антонова, А. Н. Соловьёв. Режим доступа: прямая ссылка.
6. *Романенко А. А.* Применение условных случайных полей в задачах обработки текстов на естественном языке: выпускная квалификационная работа магистра: спец. 010656 — Математические и информационные технологии. Режим доступа: прямая ссылка.
7. *Кудинов М. С.* Частичный синтаксический разбор текста на русском языке с помощью условных случайных полей / М. С. Кудинов // Машинное обучение и анализ данных. 2013. Т. 1, № 6. С. 714–724.