Динамическая грамматика и ее употребление в автоматической обработке языка[1]

Дэвид Тагуэлл
Факультет Английской Лингвистики, ELTE, Будапешт
david.tugwell@itri.brighton.ac.uk

В данной работе мы представляем *динамическую грамматику*, новаторский подход к генеративному моделированию синтаксиса. Динамическая грамматика отказывается от одного из основных предположений генеративной грамматики: то есть, что в моделировании синтаксиса мы можем, и даже должны, абстрагироваться от потока времени, и что при этом некая синтаксическая структура (либо фразовая структура, либо структура зависимостей) определяет деривацию предложения. В данном подходе мы моделируем синтаксис в виде пословного накопления семантической интерпретации, обходясь без какого-либо уровня синтаксической структуры (см. [Hausser 1999], [Kempson et al. 2001], [Tugwell 1998]).
Чтобы конкретно проиллюстрировать функционирование динамической грамматики, мы приведем деривацию примерного предложения. Потом мы обсудим некоторые преимущества динамического подхода для моделирования синтаксиса. В конце мы покажем, как можно употреблять данную модель в автоматическом анализе языка.

Динамические Грамматики

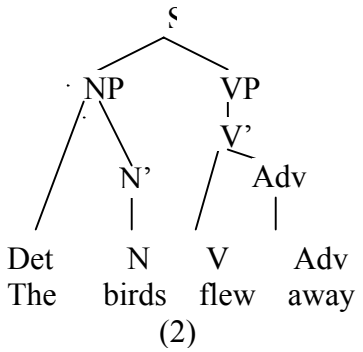
Можно сказать, что цель любой генеративной (т.е. формальной) грамматики какого-либо языка является приписыванием каждой последовательности слов, принадлежащей к этому языку, ее возможных значений. Как известно, при создании этого нового подхода, Chomsky [Chomsky 1957] основал свою трансформационную грамматику на основе анализа по непосредственно составляющим, созданного структуралистами (см. дискуссию в [Matthews 1993]). Вследствие того, до сих пор генеративная грамматика обладает статическим характером, она пользуется неким уровнем синтаксической структуры как основой модели и таким образом избегает потребности в прямом представлении семантики и одновременно абстрагируется от потока времени. Пользуясь статической синтаксической структурой, можно установить деривацию какой-то последовательности слов, работая и в обратную сторону, начиная с последнего слова.
Однако подход к генеративной грамматике, основанный на синтаксических структурах (либо фразовой структуре, либо структуре зависимостей), хотя едва ли не общепринят, далеко не единственно возможный. В последние годы ряд исследователей (см. [Hausser 1989], [Hausser 1999], [Milward 1995], [Kempson et al. 2001], [Tugwell 1998]) независимо друг от друга

[1] Я очень благодарен Наталье Кигай за терпеливую языковую помощь при писании этого текста.

предложили альтернативное построение генеративной грамматики, моделируя синтаксис в виде пословного накопления семантической интерпретации. Можно сказать, цитируя [Hausser 1989], что это динамическое представление синтаксиса основано на "возможных продолжениях" какой-то последовательности слов, по сравнению с "возможными заменами" грамматики непосредственно составляющих.
Все эти новые динамические подходы исходят из общего ощущения, что динамическое моделирование синтаксиса позволяет строить более аккуратные, компактные и пояснительные модели синтаксиса, но цели и конкретные формулировки этих моделей разные. Например, [Kempson et al 2001] показывает, что их логически-ориентированная динамическая модель может объяснять ряд проблем в теории референции и сочетания. [Hausser 1989, 1999] утверждает, что его динамическая грамматика имеет характеристики, выгодные с точки зрения формальной теории языка. [Milward 1995] элегантно решает ряд проблематичных примеров "соединения не-составляющих" (non-constituent coordination). [Tugwell 1998] представляет практическую модель, основанную на когнитивных-семантических структурах (см. [Jackendoff 1990]), и обсуждает ряд конструкций английского и других языков, включая разрывные составляющие и конструкции с дистантными зависимостями.
Чтобы конкретно представить действие динамической грамматики, в следующем разделе мы приводим простой пример, пользуясь вариантом модели в [Tugwell 1998].

Функционирование Одной Динамической Грамматики

Мы исследуем как пример анализ простого английского предложения (1) ("птицы улетели"):
(1) The birds flew away.
Типичная грамматика непосредственно составляющих может предложить что-то вроде следующей синтаксическую структуру (2) для предложения (1) (мы здесь не говорим о более современных «хомских» подходах, где предлагаемая структура была бы гораздо сложнее, а о некоторым более практичном, применимом варианте грамматики непосредственных составляющих). Семантическая интерпретация предложения (1) оказывается зависимой от данной синтаксической структуры.



(2)

Наш динамический подход отказывается от таких структур, проводя интерпретацию пословным накоплением семантической информации. Прежде всего нам необходимо подробно разработанное представление семантической интерпретации, или семантического содержания предложения. Следовательно первый вопрос -- какой метод мы выбираем для репрезентации семантических интерпретаций. Допустим, что мы воспользуемся неким вариантом системы "Minimal Recursion Semantics" (MRS, букв. семантика с минимальной рекурсией, [Copestake et al 2001]), характеризующейся "плоским" нерекурсивным представлением семантики, идеальным для представления неоконченных интерпретаций,

составляемых нарастающей структурой нашей модели.[2] Итак, допустим, что желанная интерпретация предложения (1) может быть представлена следующим образом:

(3) ситуация(s) & в_прошлом(s) & подтверждает(s,e) & событие(e) & fly'(e) & существо(x) & bird'(x) & множество(x) & определенное(x) & тема(e,x) & путь(y) & away'(y) & цель(e,y)

где

- {ситуация, событие, существо, путь}: семантические (когнитивные) примитивы
- {подтверждает, тема, цель}: семантические отношения
- {s, e, x, y}: имена переменных

и т.д.

Можно читать (3) как: "Утверждение, что имела место ситуация, в которой происходит событие перелета и темой этого события является множество птиц и цель этого события – путь из данного места в другое".

Выражение (3) представляет собой целью анализа предложения (1). Наша динамическая грамматика должна определить, как можно прийти к нему через ряд пословных переходов. В самом деле, для большинства этой семантической информации нетрудно определить, на каком слове в последовательности эта информация включится в интерпретацию, так как данная информация прямо связана с одним или другим словом предложения. Итак, допустим, что следующяя таблица показывает правильную последовательность состояний, начиная с пустой ситуации и кончая полной интерпретацией предложения. В каждом состоянии новая информация маркирована подчеркиванием, старая информация сокращена; мы пользуемся числами как именами переменных.

слово	**словоформа**	**нарастающая семантическая интерпретация**
		ситуация(1)
the	det	сит(1), существо(2), определенное(2)
birds	*bird'*, noun plural	сит(1), сущ(2), опред(2), bird'(2), множество(2)
		сит(1), сущ(2), опред(2), bird'(2), множ(2), подлежащее(1,2)
flew	*fly'*, verb preterite	сит(1), сущ(2), опред(2), bird'(2), множ(2), подл(1,2), в-прошлом(1), подтверждает(1,3), событие(3), fly'(3), тема(3,2)
away	*away'* adverb	сит(1), сущ(2), опред(2), bird'(2), множ(2), подл(1,2), в-прош(1), под(1,3), событ(3), fly'(3), тема(3,2), путь(4), away'(4)
		сит(1), сущ(2), опред(2), bird'(2), множ(2), подл(1,2), в-прош(1), под(1,3), событ(3), fly'(3), тема(3,2), путь(4), away'(4), цель(3,4)

Таблица (4)

Надо заметить, что также необходимо одно чисто синтаксическое отношение *подлежащее(x,y)*, так как в английском каждому личному лексическому глаголу требуется предназначенное подлежащее. К тому же, надо заметить, что прибавление подлежащего и

[2] В данной модели мы более склонены рассматривать семантическую структуру как изображение когнитивной действительности (как например в [Jackendoff 1990]), но это не имеет большого значения для функцировании грамматики.

прибавление отношения "цель" происходит независимо от какого-либо слова в предложении, то есть при пустых переходах.
Как правило, длина состояния увеличивается в соответствии с длиной предложения. Итак, модель имеет (счётное) бесконечное число возможных состояний. В таком случае, кажется, совсем нелегко разработать систему правил, описывающую все возможные переходы между этими бесчисленными состояниями. Однако можно представить ту же самую последовательность состояний эквивалентным, более кратким способом, как в следующей таблице.

		добавления к семантической интерпретации		
		первое место на стэке	***второе место на стэке***	***третье место на стэке***
		ситуация(1)		
the	det	**..**	существо(2), определен(2)	
birds	*bird'*,n,plur	**..**	bird'(2), множество(2)	
		подлежащее(1,2)		
flew	*fly'*, v, pret	в-прошлом(1), подтверж(1,3)	событие(3), fly'(3), тема(3,2)	
away	*away'*, adv	**..**	**..**	путь(4), away'(4)
		..	цель(3,4)	

Таблица (5)

Тут мы показываем на каждом этапе деривации только новодобавленную информацию, а также предполагаем, что новые семантические элементы (или семантические составляющие) поставлены на *активном стэке*, который с когнитивной точки зрения можно рассматривать как запас когнитивных предметов во внимании. Когда какая-то семантическая составляющая снимается со стэка, ее нужно немедленно соединить с активной структурой каким-то отношением. Например при пустом переходе между "birds" и "flew", составляющая (2) соединится с составляющей (1) отношением *подлежащее*, и при последним переходе (4) соединится с (3) отношением *цель*.
Если мы пользуемся репрезентацией в (5), не трудно видеть, каким образом можно составить правила, определяющие возможные переходы, учитывая часть речи, к которой относится данное слово, информацию о нем в лексиконе и текущее состояние интерпретации. В случае пустого перехода, надо учитывать только текущее состояние.
Очень важно подчеркнуть, что это чисто декларативная грамматика, определяющая все возможные переходы. Это не модель обработки языка или парсер, которые должны определять, *какой* из этих возможных переходов надо выбирать на данном этапе деривации. Правда, отношение между моделью обработки и самой грамматикой весьма близкое, так как переходные состояния, используемые в модели синтаксиса, точно такие, как и в парсере.

Некоторые преимущества динамического подхода

i. Семантическое обоснование и экономичность

Во-первых, можно утверждать, что данная модель более экономична, чем модель синтаксиса основанная на непосредственно составляющих, в том смысле, что динамический подход использует минимальное количество абстрактных элементов, избегая двойственности

синтаксических и семантических предметов. Чтобы убедиться, что необходимо принимать во внимании семантику в моделировании синтаксиса, рассмотрим примеры в (6):

(6) i He put it **on the shelf**.
ii He put it **there**.
iii He put it **too high for me to reach**.
iv He put it **the place he always puts it**.

В любой практической грамматике непосредственно составляющих типично видится нечто вроде правила вывода для глагола "*put*": "***VP --> NP PP***", то есть внешними аргументыми глагола "put" могут быть один существительный оборот и потом один предложный оборот. Но в действительности из наших примеров только предложение (6i), в котором цель помещения реализована предложным оборотом "*on the shelf*", соответсвует этой схеме. В остальных предложениях цель реализована соответсвенно AdvP ("*there*"), AdjP ("*too high for me to reach*")[3], или даже NP ("*the place he always puts it*"). Конечно, чтобы породить все примеры в (7), мы можем расширить грамматику, добавляя правила для "*put*": "***VP --> NP AdvP***", "***VP --> NP AdjP***" и "***VP --> NP NP***", но в результате, грамматика бы сильно перепроизводила. Лучше было бы определить, что первый аргумент глагола "*put*" -- некий предмет, а второй аргумент – некая локация; такое решение является единственно возможным для нашей динамической модели.

ii. Разрывные составляющие

Для большинства языков в мире характерно, что слова, добавляющие информацию в одну и ту же семантическую составляющую, имеют тенденцию появляться рядом в предложении. Имея в виду коммуникативную эффективность, это вполне понятно. Это явление также дает внешнее правдоподобие грамматике непосредственно составляющих. Но во всех языках есть случаи, когда эти группировки слов, по разным причинам, разделены, и, следовательно, возникают так называемые «разрывные составляющие».

Следует отметить, что разрывные составляющие не являются до такой степени проблематичными для динамической грамматики, так как у нее вовсе нет синтаксических составляющих. Чтобы понять различие между данными подходами, рассмотрим примеры в (7):

(7) i. The birds did fly away.
ii. Did the birds fly away?

С точки зрения грамматики непосредственно составляющих, в (7i) мы имеем VP "*did fly away*", а в вопросительном предложении (7ii), в результате расположения вспомогательного глагола "*did"* впереди подлежащего, этот VP разрывен. В динамической модели, чтобы позволить деривацию (7ii) нужно только добавить одно новое переходное правило – что при всяком вспомогательном глаголе и состоянии с пустой ситуацией можно создать вопросительное ситуацию, маркированную здесь знаком *Q*. Все остальные переходы такие же, как в повествовательном предложении (7i).

	добавления к семантической интерпретации		
	первое место на стэке	***второе место на стэке***	***третье место на стэке***

[3] Надо признаться, что слово "*high*" обычно считается здесь наречием, а не прилагательным. Однако, эта дезигнация основана только на возможности его употребления в таких ситуациях. Проще сказать, что оно всегда прилагательное, но притом может и имеет значение *локация*.

		ситуация(1)		
did	*do',*aux, pret	Q(1), в-прошлом(1)		
the	det	**..**	существо(2), определенное(2)	
birds	*bird'*, n, plur	**..**	bird'(2), множество(2)	
		подлежащее(1,2)		
fly	*fly'*, v, bare	подтверждает(1,3)	событие(3), fly'(3), тема(3,2)	
away	*away'*, adv	**..**	**..**	путь(4), away'(4)
		..	цель(3,4)	

Во многих отношениях роль переходных правил в данной модели соответсвует роли *конструкций* в конструкционной грамматике (см. [Goldberg 1995], [Fillmore et al. 2004]). У обоих подходов также есть практическая ориентация и убеждение, что грамматика должна относиться одинаковом способом ко всем синтаксическим конструкциям в данной языке.

Употребление Модели В Автоматической Обработке Языка

Можно заметить, что предлагаемая модель синтаксиса близко соотносится с моделью обработки языка, так как сама грамматика работает пословно "слева направо", определяя все возможные следующие состояния нарастающей семантической структуры. Чтобы превратить ее в модель, применимую для обработки текстов, нужно определить вероятность этих пословных переходов. Уже доказано (Chelba & Jelinek, 1998), что вероятностная модель, использующая синтаксическую информацию, превосходит любую модель N-грамм в распознавании речи. В данной модели существует возможность использовать еще более широкий запас синтаксической и семантической информации.

Мы опять приведем пример предложения (1): "*the birds flew away*". Любая модель N-грамм, сформированная на достаточно обильном количестве текста, может определить, что все пары слов в предложении[4] (*"the birds", "birds flew", "flew away"*) довольно вероятны и, следовательно, что (1) является довольно вероятным предложением англйского языка. Таким образом, обратясь к динамической модели, если у нас достаточно анализорованного текста, можно вычислить, что лексема *bird'* вполне вероятна как *тема* лексемы *fly'*: то есть, что комбинация {*fly', тема, bird'*} имеет высокую вероятность, как и комбинация {*fly', цель, away'*}. Пользуясь этими и другими статистическими данными, можно вычислить вероятность каждого перехода в деривации предложения, и следовательно вероятность самого предложения.

Явные преимущества динамической грамматики перед модели N-грамм станут очевидны, когда мы рассмотрим примеры в (8), где вышеупомянутой лексической зависимостью {*fly', тема, bird'*} уже не может пользоваться никакая модель N-грамм, потому что существенные слова (подчеркнутые) слишком далеки друг от друга.

(8) i. The **birds** seemed to be about to **fly** off.
ii. This is the **bird** that has been **flying** since morning.
iii. Which **bird** did you think might not be able to **fly**?
iv. **Flying** up into the evening sky, the **birds** screeched and cawed.

[4] Здесь мы говорим для простоты о модели биграмм.

Чтобы воспользоваться такими зависимостями, обязательно иметь полную модель синтаксиса. Употребляя данную динамическую модель в предложениях (8 i-iv), мы можем учитывать вероятность комбинации {*fly', тема, bird'*} немедленно, когда мы встретим другое подчеркнутое слово.

Литература

1. [Chelba & Jelinek 1998] Chelba, Ciprian & Frederick Jelinek (1998). Exploiting Syntactic Structure for Language Modeling. In Proceedings 36th ACL.
2. [Chomsky 1957] Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.
3. [Copestake et al 2001] Ann Copestake, Dan Flickinger, Ivan A. Sag (2001). Minimal Recursion Semantics: An Introduction. *Language and Computation,* Vol I, No. 3, 1-47.
4. [Fillmore et al 2004] Fillmore, Charles, Paul Kay, Laura Michaelis & Ivan Sag (2004). *Construction Grammar*. CSLI Publications.
5. [Goldberg 1995] Goldberg, Adele E. (1995). *Constructions: A Construction Grammar Approach to Argument structure*. University of Chicago Press.
6. [Hausser 1989] Hausser, Roland. 1989. *Computation of language: an essay on syntax, semantics and pragmatics in natural man-machine communication*. Berlin: Springer-Verlag.
7. [Hausser 1999] Hausser, Roland (1999). *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*. Springer Verlag.
8. [Jackendoff 1990] Jackendoff, Ray (1990). *Semantic Structures*. MIT Press.
9. [Matthews 1993] Matthews, Peter H. 1993. *Grammatical theory in the United States from Bloomfield to Chomsky*. Cambridge University Press.
10. [Kempson et al 2001] Kempson, Ruth, Wilfred Meyer-Viol & Dov Gabbay (2001). *Dynamic Syntax: The Flow of Natural Language Understanding.* Blackwell.
11. [Tugwell 1998] Tugwell, David (1998). Dynamic Syntax. PhD Thesis, University of Edinburgh.