Об одном способе осуществления синтеза предложения при МП /на основе синтагматического анализа/.

Ч. А. Хоар.

my first article, ? published written summer 1960 in Машинные

Перевод

По обычным представлениям, в результате ситагматического анализа получается список синтагм, т.е., список упорядоченных пар слов, первое из которых "подчиняет" второе, причем для каждой синтагмы имеется признак, указывающий на то, какой тип связи имеет место в данной синтагме. При синтезировании предложения /другом /или на том же/ языке, из списка пар надо получить линейную последовательность слов, образующую синтаксически и морфологически правильное предложение. При этом надо определить, какой именно набор признаков синтагм окажется наиболее удобным. В настоящей заметке я не буду рассматривать морфологическую сторону синтеза: я ограничусь вопросом о расстановке слов в нужном порядке /что более существенно для синтеза предложения, например, на английском языке, и вообще представляется более сложной задачей/.

На

Предлагаемый метод синтеза основывается на том, что во многих языках линии зависимости не пересекаются, т.е., если слова a и b входят в одну синтагму (a, b), и слова c и d входят в другую синтагму (c, d), то эти слова нельзя ставить в таком порядке, что линия, проведенная между a и b пересекается с линией, проведенной между c и d. [1]) Это значит, следующие возможности исключаются:

...c...b...d	...c...d...b
...d...b...c	...a...c...b
...c...a...d	...b...d...a
...b...d...c	...d...b...c...a

[1]) Это утверждение верно во всех случаях, когда легко по интуиции установить линии зависимости. В разборе более сложных случаев, надо обеспечить, чтобы осталось верным.

Мы можем использовать этот факт, если мы синтезируем фразу слева направо /разыскивая сперва первое слово, потом второе, и так далее/. Положим, что на данном шагу мы правильно расставили первые n слов /которые мы в дальнейшем обозначим расставленными словами/; мы разыскиваем теперь $n+1$ ое слово /разумеется, среди нерасставленных слов/. Здесь полезно определить понятие "косвенной зависимости". Слово a прямо зависит от слова b, если существует синтагма (b,a) ж в анализируемом предложении; слово косвенно зависит от слова b, если существует в анализируемом предложении последовательность слов $x_1, x_2 \ldots x_n$, причем x_1 прямо зависит от b, x_{i+1} прямо зависит от x_i /для каждого i/ и a прямо зависит от x_n. Иными словами, b является в некотором смысле "предком" слова a. Теперь мы можем сформулировать следующие правила:

№ I. Если существует не расставленное слово, прямо зависящее от n-ого слова, то в качестве $n+1$-ого слова нельзя выбрать слово, которое не зависит /прямо или косвенно/ от n-ого слова /если бы мы выбрали такое слово, это неизбежно привело бы к пересечению линий зависимости/.

II. Если не существует слова, зависящего от n-ого, но есть слово, подчиняющего n-ое слово, то можно выбрать в качестве $n+1$-ого слова только это слово или слово, прямо или косвенно зависящее от него.

Мы могли бы привести другие правила, ограничивающие возможности выбора следующего слова; но достаточно этих двух, чтобы показать каким именно образом выбор очередного слова может быть значительно сокращен.

Такие правила, однако, не определяют однозначно какое именно слово надо выбрать в случае если существует больше одного нерасставленного слова, зависящего от последнего /n-ого/ и расставленного слова. Для таких случаев мы вводим особые признаки синтагм. Эти признаки должны указывать, стоит ли зависящее слово налево или направо от управляющего слова, и как "тесно" они связаны. /а
В качестве таких признаков используются целые числа: отрицательные числа обозначают, что зависимое слово стоит налево, и положительные числа - направо; причем меньшая абсолютная величина обозначает более тесную связь, и наоборот.

Если слово более тесно связано со словом чем , это обозначает, что они должно стоять ближе к чем т.е., порядок слов будет /в случае левой зависимости/, или /в случае правой зависимости/.

Положим теперь, что два нерасставленных слова, b и c, зависят от последнего / ого/ слова a в синтагмах $+1(a,b)$ и $+3(a,c)$. В таком случае мы не можем выбрать слово c, так как b более тесно связано с a чем c /$|+1| < |+3|$/. /Здесь мы предполагаем, что признаки приписываются с таким расчетом, чтобы никогда не встречались два слова, зависящих от данного слова с одним и тем же признаком/. Но мы еще не можем сказать, что именно b есть следующее слово: если другое слово d "зависит налево" /т.е., с отрицательным признаком/ от слова b, то d непременно должно стоять перед b. Если два слова "зависят налево" от b, то мы выбираем в качестве "кандидата" на следующее место то слово, которое менее тесно связано с b, то есть, которое должно стоять дальше налево от него. Обозначим это слово e. Если не существует не расставленного слова, зависящего от e налево, мы можем определенно ставить e на $n+1$-ом месте, и потом весь процесс начинается с самого начала. Если, однако, существуют слова, зависящие от e налево, мы выбираем на место "кандидата" то слово, которое должно стоять дальше налево: бывший кандидат e отвергается. Мы продолжаем этот процесс, пока не найдем "кандидат", от которого ничего не зависит налево: именно это слово будет $n+1$-ым словом в синтезируемом предложении. Можно грубо описать изложенный процесс следующим образом: искать "первое" слово, зависящее направо от последнего расставленного слова; обозначим его b. Потом искать "крайнее" слово, прямо или косвенно зависящее налево от b: ставить это слово на следующем месте после a.

Я еще не рассматривал такой случай, когда не существует не расставленного слова, зависящего направо от последнего расставленного слова. Этот случай легко разбирается. Мы переходим на $n-1$-ое слово, и обрабатываем его как будто оно и является последним расставленным словом : если и у него нет зависящих направо, то мы переходим на $n-2$-ое слово, и т.д. Но когда мы нашли искомое слово n-ое, мы всё равно ставим его на $n+1$-ом месте : никаких вставок не может быть.

До сих пор я предполагал, что мы уже расставили первые n слов : теперь надо объяснить как выбирается первое слово /в случае $n=0$ /. Здесь мы рассматриваем такой метод анализа, при котором остается одно и только одно слово, которое не зависит от другого слова. В самом начале синтеза удобно взять это независимое слово в качестве "первого"слова, и, идя слева направо, синтезировать вышеописанным образом ту часть фразы, которая стоит направо от него. Потом мы синтезируем левую часть фразы, начиная еще раз с независимого слова, но идя теперь справа налево. При этом, можно использовать тот же самый алгоритм, если только заменить повсюду слова "правое" на "левое", "за" на "перед", и "следующее" на "предыдущее".

Остается теперь лишь один вопрос : каким образом приписать синтагмам признаки, т.е., как определить "тесность" связей в синтезируемом предложении. Здесь я возьму в качестве иллюстрации английский язык. Можно перечислить для примера некоторые типы синтагм, в которых существительное фигурирует как управляющий член.

Тип синтагмы	Пример	Признак
(S, артикль)	the	−3
(S, прилагательное)	new	−2
(S, S как прилагательное)	Leica camera	−1
(S, S в аппозиции)	(type III)	+1
(S, предлог)	in a leather case	+2
(S, причастие)	bought in Germany	+3
(S, соч. союз)	, and an expensive lens	+4

8 Здесь мы считаем, что, когда существительное управляет несколько прилагательных, к только самое близкое связывается с ним, а другие связываются друг с другом в цепочке. Например к "современная вычислительная машина" анализируется в синтагмы (машина, вычислительная)
(вычислительная, современная)

Здесь я не претендую на полноту : вероятно, что понадобится больше 7 степеней тесности. Я надеюсь, однако, что из этого списка будет ясно, что в стандартных случаях можно легко установить степень тесности, если только знаем, какие части речи входят в данную синтагму. В нестандартных случаях, конечно, дело обстоит сложнее, и надо будет составить более или менее сложный алгоритм, чтобы разобрать такие случаи. Например, если причастие не вводит целого оборота /т.е., от него ничего не зависит кроме, может быть, наречие/, то оно обычно стоит перед существительным на том же месте, где обычно стоит прилагательное : ~~ж~~ сравните a exciting play, a play exciting much interest.
С другой стороны, если от причастия зависит предложный оборот, оно обыкновенно стоит за существительным : сравните a saved boy, a boy saved by Nobody.
Но это уже дело лингвистов, и не касается технического метода осуществления синтеза.

Для этого метода синтеза предполагается, что язык, на котором синтезируется предложение, является регулярным в следующем смысле. Положим, что ~~три~~ слова a, b, c, d, и ~~X~~ зависят от ~~третьего~~ слова x, и что они должны стоять в порядке $a \ldots b \ldots x \ldots c \ldots d$. ~~Скажем, что в другом предложении~~ Если этот порядок остается неизменным, несмотря на возможное ~~наличие~~ отсутствие других слов, зависящих от x, то язык называется регулярным, и предлагаемый метод применяется без особенных затруднений. В таком случае, можно приписать ~~приписать~~ признак каждой синтагме на основе свойств одной /данной/ синтагмы. В обратном случае, невыгодно было бы применить предлагаемый метод.

Е

Достоинство предлагаемого метода состоит в следующем :

1 Лингвистическая проблема синтеза целого предложения /в случае регулярности/ к проблеме правильной расстановке только двух слов, зависящих от одного слова /так как правильный порядок всех слов, зависящих от данного слова однозначно определяется, если только знать правильный порядок всех пар слов. Тогда можно составить алгоритм который приписывает синтагмам их нужные признаки.

2 Алгоритм имеет циклический характер, так что программа будет очень проста, и для реализации процесса понадобится незначительное число ячеек машинной памяти.

Недостатком предлагаемого метода является то обстоятельство, что его осуществление требует много машинных операций : чтобы найти на каждом шагу крайнее слово налево, надо будет пройти несколько раз весь список синтагм. Этот недостаток может быть существенен, если мы переводим на язык, в котором вообще мало приходится переставить.

Автор хочет выразить свою благодарность И.А.Мельчуку который сделал поправил изложение во многих местах, и вел со мной разговоры, на которых эта статья основана.

thoare

Customer Order Form_N1.pdf
27-09-06 10:06

thoare

India.pdf
27–09–06 10:05