РАЗРЕШЕНИЕ СИНТАКСИЧЕСКОЙ НЕОДНОЗНАЧНОСТИ И ИЗВЛЕЧЕНИЕ СЛОВАРЯ МОДЕЛЕЙ УПРАВЛЕНИЯ ИЗ КОРПУСА ТЕКСТОВ

Александр Гельбух

Лаборатория естественного языка Центра Компьютерных Исследований (CIC)
Национального Политехнического Института (IPN) Мексики
LabLN, CIC, IPN, Av. Juan Dios Batiz s/n esq. Mendizabal, Zacatenco 07738, DF, Mexico.

gelbukh*pollux.cic.ipn.mx

ABSTRACT

An iterative statistical method of syntactic disambiguation in a large text corpus is discussed. The method is based on the automatic acquisition and use of combinatorial information, in the first place subcategorization frames of the words found in the corpus, along with the statistical weights of individual combinations of a word with all its dependent words. The dictionary acquired with this procedure is usable for (non-iterative) disambiguation of other texts. The method can be generalized on other types of combinatorial information such as collocations and grammar rules.

ВВЕДЕНИЕ*

На наш взгляд, разрешение неоднозначности — одна из ключевых проблем современной компьютерной лингвистики. С помощью хорошо известных методов морфологического и синтаксического анализа легко построить все возможные для данного слова или фразы интерпретации — в лучшем случае несколько, в худшем — сотни и тысячи. Следующая проблема — выбрать ровно одну из них, правильную, и эта проблема оказывается очень трудной. Для разрешения неоднозначности нужны словари комбинаторного типа, содержащие информацию о возможных синтагматических связях для конкретных лексем; словарь моделей управления — простейший вид такого словаря. Однако составление таких словарей вручную — дело чрезвычайно трудоемкое.

В данной работе предлагается метод извлечения такого словаря из корпуса текстов и одновременного разрешения неоднозначности в этом корпусе. Предполагается, что мы располагаем синтаксическим анализатором, строящим несколько, возможно, сотни и тысячи, гипотетических синтаксических структур для каждой фразы. Алгоритм назначает таким гипотезам веса, соответствующие вероятности того, что данная

* Работа выполнена при частичной финансовой поддержке CONACyT, грант 26424-A, REDII-CONACyT, DEPI-IPN и SNI, Мексика.

гипотеза правильна; гипотеза с наибольшим весом должна рассматриваться последующими блоками системы первой. Одновременно составляется частотный словарь вариантов оформления синтаксических валентностей лексем в правильных вариантах разбора.

Итеративные статистические методы использовались, например, в маркерах частей речи (part-of-speech taggers) для английского языка [6]. Показано, что получаемые с их помощью по корпусу текстов массивы данных могут быть использованы для анализа независимых текстов [7]. Однако в некоторых случаях, в особенности, когда математическая модель не соответствует реальному лингвистическому явлению, как, например, использование скрытых Марковских моделей в маркерах частей речи, итеративная процедура может сходиться к состоянию, не соответствующему лингвистическим фактам [8]. Мы полагаем, что в нашем случае описываемая модель достаточно хорошо соответствует лингвистическому содержанию. Эксперименты подтверждают работоспособность метода.

СЛОВАРЬ МОДЕЛЕЙ УПРАВЛЕНИЯ

В данной работе мы предлагаем метод автоматического составления списка возможных сочетаний слов, синтаксически связанных с данным, вместе с их характеристиками. В Табл. 1 приведен фрагмент такой структуры для слова *купить*.

Табл. 1. *Фрагмент статьи автоматически составленного словаря для слова* купить.

Пра-вильные	Непра-вильные	Сочетание	Пример
164782	26	вин.	*книгу*
37819	35	вин + *в* предл.	*хлеб, в магазине*
3768	47	*в* предл. + вин.	*в магазине, масло*
2826	93	вин + *на* предл.	*рыбу, на рынке*
953	643	*на* предл. + партитив + *по* дат.	*на рынке, сахару, по рублю*
632	1276	*в* предл. + вин. + *для* род.	*в книжном, учебник, для брата*

В первых двух колонках приведен пример статистических весов — частот появления данного сочетания в правильных вариантах разбора и, для сравнения, в неправильных, ошибочно построенных конкретным анализатором. Перечень характеристик, включаемых в словарь, зависит от размера корпуса и возможностей анализатора. Если анализатор определяет, например, одушевленность, или семантический класс, или какие-либо иные характеристики лексем, то они могут быть включены в словарь наряду с предлогами и падежами, показанными в Табл. 1. Может учитываться или не учитываться синтаксический тип связи. Порядок слов также может учитываться или не учитываться, в зависимости от размера корпуса.

В наш словарь входят все наблюдавшиеся в тексте сочетания узлов, синтаксически связанных с данным. При этом, вероятно, должны учитываются все связи, а не только актантные. Если анализатор вообще не определяет тип связи, а только строит дерево, то, соответственно, все возможные связи будут включены в словарь. Естественно, в

целях сокращения размера словаря в нем должны быть оставлены только наиболее частотные сочетания.

РАЗРЕШЕНИЕ НЕОДНОЗНАЧНОСТИ С ПОМОЩЬЮ СЛОВАРЯ МУ С ВЕСАМИ

Разрешение неоднозначности как определение вероятностей гипотез. Пусть имеется процедура синтаксического разбора, строящая, возможно, несколько гипотетических синтаксических структур для каждой фразы. Под разрешением неоднозначности мы понимаем определение вероятности того, что конкретная гипотеза является верной. Будем называть такое число весом гипотезы. В идеальном случае вес одной из гипотез должен быть 1, а остальных — 0.

На практике в условиях неполной информации лучше определять вес каждой гипотезы числом $w_j \in [0, 1]$, $\sum w_j = 1$ для каждой фразы. Этот метод позволяет комбинировать различные подходы к оценке гипотез. Так, если имеется несколько процедур оценки гипотез — с помощью словаря МУ, словаря словосочетаний, весов грамматических правил и т.д., то окончательный вес каждой гипотезы должен определяться, после соответствующей нормировки, произведением весов, назначенных данной гипотезе каждой из процедур оценки. Кроме того, при таком методе становится возможным упорядочение гипотез по весам. Процедура синтаксического анализа передает следующему блоку системы обработки текста наилучшую гипотезу. Если следующий блок по каким-либо причинам отвергает данный вариант разбора, рассматривается следующая по весу гипотеза.

Разрешение неоднозначности с помощью словаря МУ с весами. Предположим, имеется словарь возможных сочетаний характеристик слов, синтаксически подчиненных данному, см. Табл. 1. При этом могут учитываться семантические пометы, порядок слов и т.п. Для каждого такого сочетания известно, сколько раз оно встречалось в правильных вариантах разбора и сколько раз — в неправильных. Обозначим эти числа p^+ и p^- соответственно — вероятности появления данного сочетания в правильном и в неправильном варианте разбора.

Легко показать, что вес гипотезы определяется произведением отношений $p^+/\ p^-$ по каждому узлу синтаксической структуры, соответствующей данной гипотезе [1]. Естественно, такие веса нормируются по множеству гипотез, рассматриваемых для одной фразы. Мы вынуждены опустить доказательство, оно несложно и основано на теореме Байеса; содержательно, вклад каждого сочетания в вес гипотезы определяется тем, чаще ли анализатор строит такие сочетания по ошибке или реально находит их в тексте. Заметим, что нормировка весов гипотез внутри одной фразы позволяет ограничиться учетом только тех узлов, по которым хотя бы две гипотезы различаются.

Если сочетание вообще отсутствует в словаре, следует положить его вклад в произведение равным некоторой малой величине, как если бы оно было встречено много раз в неправильных вариантах и редко в правильных. Эта величина, однако, должна быть ненулевой, чтобы оставалось возможным сравнение гипотез по остальным сочетаниям, а из гипотез, имеющих разное количество отсутствующих в словаре сочетаний, была выбрана та, которая имеет их меньше.

Пример. Пусть для фразы *Лена купила в магазине футляр для очков* построены две гипотезы, различающиеся присоединением предложной синтагмы: 1) *купила для очков*

---

[1] Словарь строится так, что делитель в этой формуле никогда не бывает равным нулю, см. раздел 0.

и 2) *футляр для очков*. Сочетания связей для узлов *Лена*, *магазин* и *очки* при этих двух гипотезах одинаковы, следовательно, при нормировке весов по множеству всех гипотез они все равно уничтожатся. Различны у этих двух гипотез оказываются сочетания зависимых слов при узлах *купить* и *футляр*. Пусть в словаре МУ были найдены следующие частоты для этих сочетаний:

Табл. 2. *Пример результатов поиска в словаре для двух гипотез.*

Правильные	Неправильные	Сочетание	Текст
		Гипотеза 1	
632	1276	*купить*: *в* предл. + вин. + *для* род.	*в магазине, футляр, для очков*
272	6597	*футляр*: ∅	∅
		Гипотеза 2	
3768	47	*купить*: *в* предл. + вин.	*в магазине, футляр*
8902	489	*футляр*: *для* род.	*для очков*

Веса гипотез вычисляются перемножением вероятностей появления каждого сочетания в правильном варианте и делением на вероятности появления их в неправильном варианте. Для данного случая расчет следующий: $w_1 \sim (632 / 1276) \times (272 / 6597) \approx 0.02$, $w_2 \sim (3768 / 47) \times (8902 / 489) \approx 1459$, что после нормировки дает $w_1 \approx 0.00002$, $w_2 \approx 0.99998$. Вторая гипотеза явно предпочтительна.

Заметим, что фраза **Лена купила футляр для брата* неестественна, а для фразы *Лена купила для брата футляр* синтаксический анализатор просто не построит ложной гипотезы, либо другая процедура оценки, а именно, оценка по порядку слов, назначит слишком малый вес гипотезе *футляр для брата*.

Веса для разрешения неоднозначности: характеристика языка или анализатора? Как обсуждалось выше, вклад определенного сочетания в вес гипотезы определяется не частотностью этого сочетания, а отношением частоты встречаемости данного сочетания в правильных гипотезах к частоте его встречаемости в неправильных. Другими словами, важно лишь, чаще ли анализатор формирует данное сочетание по ошибке или действительно обнаруживает его в тексте.

При этом может оказаться, что очень частотное сочетание (обычно это бывает пустое сочетание, то есть отсутствие у слова подчиненных узлов) еще чаще строится анализатором по ошибке. В этом случае присутствие такого сочетания, несмотря на его частотность, должно быть скорее сигналом тревоги. В случае же использования обычного словаря сочетаемости, учитывающего только частотность появления сочетаний в тексте, присутствие в гипотезе возможного сочетания однозначно считается признаком успеха. Напротив, малочастотное сочетание может оказаться надежным признаком успеха, если анализатор почти никогда не строит такое сочетание по ошибке.

Следовательно, статистические веса в словаре, предназначенном для разрешения неоднозначности, должны отражать не только или не столько свойства самого языка, сколько свойства используемого анализатора. Это и неудивительно — ведь неоднозначность не является свойством самого языка (люди понимают язык однозначно, значит, это возможно), скорее, она — следствие несовершенства наших инструментов для его анализа. Зависимость статистических весов в словаре от версии используемого анализатора и, следовательно, необходимость постоянного обновления

словаря по мере развития используемого анализатора является еще одним аргументом в пользу применения автоматических методов составления словарей.

ИЗВЛЕЧЕНИЕ СЛОВАРЯ МУ С ВЕСАМИ ИЗ КОРПУСА ТЕКСТОВ

Предположим теперь, что уже имеется процедура разрешения неоднозначности в корпусе текстов. Таким образом, для каждой фразы не только строится некий набор гипотетических синтаксических структур, но и каждой гипотезе назначается вес, соответствующий вероятности того, что эта гипотеза является правильной. Веса могут также назначаться путем "голосования" нескольких разных процедур оценки, скажем, по порядку слов, по семантике, по лексической сочетаемости, по моделям управления. Веса всех гипотез для одной фразы в сумме составляют 1.

Предположим сначала для простоты, что для каждой фразы только одна гипотеза имеет вес 1, а остальные — 0. Для подсчета статистических весов сочетаний, входящих в хотя бы одну гипотезу, достаточно просуммировать число правильных гипотез, в которых встретилось данное сочетание, и, соответственно, число неправильных.

Если же гипотеза, в которую входит данное сочетание, имеет вес $p \in [0, 1]$, то можно считать, что это сочетание входит как бы не в целую правильную гипотезу, а в часть правильной гипотезы — соответственно доле p уверенности в том, что рассматриваемая гипотеза правильна. Оно же входит и в "часть" неправильной гипотезы — соответственно доле $q = 1 - p$ уверенности в том, что та же самая гипотеза неправильна. Обоснование этого метода также опирается на теорему Байеса.

Итак, для нахождения частотности вхождения в правильные гипотезы, а лучше сказать, веса, p^+ конкретного сочетания по корпусу текстов необходимо сложить веса p всех гипотез, в которые оно входит. Для нахождения частотности p^- вхождения сочетания в неправильные гипотезы необходимо сложить дополнения $q = 1 - p$ к весам тех же гипотез. Данные веса можно нормировать, разделив на число правильных гипотез в корпусе, равное числу фраз в нем и, соответственно, на число неправильных гипотез, равное полному числу гипотез минус число фраз.

Однако как показали эксперименты, редко встретившиеся сочетания представляют наибольшие проблемы при работе алгоритма. Действительно, сочетание, встретившееся один раз в правильном варианте и ни разу в неправильных оказывается «бесконечно хорошим». Поэтому *a priori* мы добавляем некоторую константу λ к числу p^- вхождений каждого сочетания в неправильные варианты — как если бы такие неправильные вхождения оказались ненайденными просто из-за ограниченности корпуса. Экспериментально было замечено, что наилучшие результаты достигаются при величине λ порядка числа фраз в корпусе, однако нам неизвестно теоретическое обоснование такого выбора. Таким образом решается и проблема деления на 0 в выражении p^+/p^- — величина p^- всегда положительна.

Составленный таким образом словарь получается очень большим, поскольку в него входят, наряду с "правильными", и все "неправильные" сочетания, появившиеся в вариантах разбора фраз. Поэтому огромное множество сочетаний со слишком малыми весами отбрасывается в момент выдачи словаря или даже еще раньше, во время работы[2]. Под весом может пониматься либо величина p^+/p^-, либо p^+, в зависимости от

[2] Для экономии памяти во время работы сочетания, хранящиеся в словаре, кэшируются: если при поступлении очередного элемента размер словаря превышает допустимый, то из словаря удаляется

цели составления словаря. Как было указано выше, если сочетание не найдено в словаре, то вносимый им вклад в вес гипотезы принимается равным некоторому малому числу. Это число и может быть использовано в качестве порогового значения для удаления сочетаний из словаря.

ИТЕРАТИВНАЯ ПРОЦЕДУРА

Разрешение неоднозначности методом итераций. В предыдущих разделах описаны взаимно противоположные процедуры: разрешение неоднозначности в предположении, что уже имеется словарь, и построение словаря в предположении, что неоднозначность уже разрешена. Как это часто бывает, проблема курицы и яйца хорошо решается методом итераций.

Пусть имеется большой корпус текстов и процедура синтаксического анализа, строящая для каждой фразы несколько гипотетических синтаксических структур. Итерации можно начинать с любой из двух описанных выше процедур, но проще начать с процедуры разрешения неоднозначности. На первой итерации словарь пуст. Процедура разрешения неоднозначности, обращаясь к словарю по каждому узлу каждого варианта, получает от него одно и то же очень маленькое число. Поскольку во всех вариантах разбора одной фразы имеется равное число узлов, все варианты оказываются в равных условиях, и процедура назначает им равные веса.

На втором шаге первой итерации строится словарь: для каждого обнаруженного в корпусе сочетания веса всех вариантов, содержащих его, суммируются. Уже после первой итерации словарь содержит вполне приемлемую статистику, однако коррекция весов вариантов позволит отсеять ложные сочетания. Такая коррекция проводится на первом шаге второй итерации по уже непустому словарю. На втором шаге второй итерации веса сочетаний в словаре вычисляются заново, и так далее. Итерационный процесс останавливается, когда веса элементов словаря и вариантов разбора фраз в достаточной мере стабилизируются.

Итеративный процесс не требует никакой дополнительной памяти, хотя сам словарь в каждый момент вычислений должен храниться в памяти целиком. Информация о вариантах синтаксической структуры фраз хранится в достаточно компактном виде: в виде групп ссылок на сочетания хранящиеся в словаре — такая группа соответствует одному синтаксическому дереву, и эти группы сами сгруппированы в множества, соответствующие вариантам разбора отдельных фраз. При наличии достаточного объема памяти для хранения всего словаря процесс пересчета касается только весов элементов и не требует перемещения элементов в памяти или повторного синтаксического анализа корпуса.

Экспериментальная проверка. Подробное описание методики экспериментов приведено в нашей статье [2]. Итерационный процесс сходится очень быстро. Типичные результаты эксперимента представлены для нескольких первых итераций следующими последовательностями: число правильно разобранных фраз — 37%, 85%, 89%, 90%, 90%, ..., число правильно разобранных фраз из тех, которые первоначально получили первоначально более одного варианта разбора — 16%, 80%, 86%, 87%, 87%, ... Далее эти показатели не менялись в ходе как минимум 50 итераций.

"наименее важный" элемент. Важность элемента рассчитывается с учетом веса элемента и/или времени последнего его появления в достаточно хорошем варианте.

На первой итерации некоторые фразы оказались "угаданными" чисто случайно, это и были те 16% фраз, для которых неоднозначность разрешилась "сама собой". На следующей итерации применялся словарь, составленный по корпусу с неразрешенной неоднозначностью, уже он позволил правильно разрешить 80% неоднозначности. После еще двух итераций процент ошибок сократился более чем на треть, что оправдывает применение итеративной процедуры.

Для построения словаря потребовалось гораздо больше итераций, поскольку нас интересовал не только выбор одного из многих вариантов, но и конкретные веса, использовавшиеся для удаления из словаря случайного мусора. Обычно состав словаря в достаточной мере стабилизировался после примерно 10-й итерации, хотя и после 50 итераций продолжал несущественно меняться. При анализе независимого корпуса с помощью словаря, полученного в результате 50 итераций, процент правильно разобранных фраз был 85%, или 79% неоднозначных фраз, хотя наилучшие результаты, 91% и 88%, были достигнуты со словарями, полученными на 4-й и 5-й итерациях. По-видимому, ухудшение результатов связано с удалением из словаря сочетаний с малыми весами.

Как показали эксперименты, результат применения метода почти не зависит от начального распределения весов вариантов или от начального наполнения словаря. Вся дополнительная информация о весах гипотез должна учитываться на каждом шаге процедуры путем введения дополнительных сомножителей в веса вариантов. Таким образом, приводимые в [8] рекомендации по применению итеративного метода представляются применимыми и к нашему случаю: если уже есть достаточно хороший словарь, накопленный по большому корпусу, итеративный метод на малом корпусе применять не следует.

РАЗВИТИЕ ИДЕИ

Метод допускает много вариантов и содержит параметры, выбор конкретных значений которых не вполне ясен. Так, например, можно экспериментировать с порогами весов для удаления ненужных сочетаний из словаря. Вероятно, некоторые специальные поправки к весам требуются для сочетаний, наблюдавшихся в корпусе очень малое количество раз. Выше обсуждался искусственный прием введения некоторого количества λ априорно ложных вхождений в корпус каждого сочетания. Хотя этот прием представляется слабо обоснованным, без него метод не работает.

В случае недостаточного объема корпуса могут рассматриваться не все сочетания, а только попарные: скажем, вместо "*купить в* предл. + вин. + *для* род." рассматривать три отдельных сочетания: "*купить в* предл.", "*купить* вин." и "*купить для* род." Это сделает статистику гораздо более надежной, однако ограничения на сочетаемость актантов будут при утеряны. Можно рассматривать двойные сочетания и т.д. Мы планируем также дальнейшие эксперименты с учетом не только предлогов и падежей, но и различных семантических признаков слов, прежде всего одушевленности.

Хотя МУ являются удобным полигоном для отладки метода, его применимость, по-видимому, ими не ограничивается. Тем же методом могут извлекаться и использоваться для разрешения неоднозначности и свободные словосочетания, то есть не сочетания слова с предлогом или со словом в определенном падеже, а сочетания слова с другими конкретными словами: *купить хлеб*, *купить книгу, купить в магазине* т.д. [3] Естественно, это потребует гораздо большего объема корпуса и гораздо больше памяти для хранения словаря. Напротив, даже небольшого корпуса может оказаться

достаточно для определения статистических весов правил грамматики, используемой для синтаксического анализа [5, раздел 7.5]. Наилучшие результаты, по нашему мнению, могут быть достигнуты при одновременном применении данного метода к нескольким разным наборам данных — МУ, грамматике, словосочетаниям и т.д.

ЗАКЛЮЧЕНИЕ

Предложенный чисто вычислительный метод, не требующий никакой дополнительной информации, позволяет одновременно разрешать синтаксическую неоднозначность и строить словари комбинаторного типа, в том числе словари МУ, по большому корпусу текстов. Эти словари могут быть в дальнейшем использованы для разрешения неоднозначности в других текстах. При этом метод совместим с любыми другими методами оценки весов синтаксических гипотез — дополнительная информация учитывается путем введения сомножителей в выражение для вычисления весов.

По-видимому, успех метода основан на том, что соответствующие объекты и процедуры, в данном случае МУ и их статистические веса, являются лингвистической реальностью, учитываемой говорящим в процессе генерации речи. Так, если говорящий чувствует, что построенная фраза, хотя и вполне правильная грамматически, будет неправильно проинтерпретирована слушающим путем оценки весов синтаксических гипотез, то он старается изменить структуру фразы, сделав ее однозначно интерпретируемой. При этом семантическая информация учитывается слабо. Благодаря такому механизму, если наша процедура оценки статистических весов дает для двух вариантов интерпретации фразы, скажем, веса 60% и 40%, то вероятность того, что следует предпочесть первую гипотезу, на самом деле гораздо выше 60% — иначе бы говорящий постарался перестроить фразу. Впрочем, такой вывод требует психолингвистической проверки.

ЛИТЕРАТУРА

1. Апресян Ю.Д., Богуславский И.М., Иодмин Л.Л. и др. *Лингвистическое обеспечение системы ЭТАП-2*. М: Наука, 1989.
2. Большаков И.А., Гельбух А.Ф., Галисия-Аро С. *Имитационное моделирование в лингвистике: оценка и отладка методов анализа с помощью генератора квазитекста*. Труды Международного семинара Диалог-98: компьютерная лингвистика и ее приложения, Казань, 1998.
3. Большаков И.А, Кассиди П.Дж., Гельбух А.Ф. *КроссЛексика — словарь словосочетаний и тезаурус общеупотребительной лексики русского языка*. Труды Международного семинара Диалог-95: компьютерная лингвистика и ее приложения, Казань, 1995.
4. Мельчук И.А. *Опыт теории лингвистических моделей Смысл ⇔ Текст*. М.: Наука, 1974.
5. Allen, James. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, Inc., 1995.
6. Baum, L.E. *An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process*. Inequalities, 3:1–8, 1972.
7. Cutting, D. *et al*. *A practical part-of-speech tagger*. Proc. of Third Conference on Applied Natural Language Processing, Trento, Italy, 1992.
8. Elworthy, D. *Does Baum-Welsh re-estimation help taggers?* Proc. of Fourth Conference on Applied Natural Language Processing, Stuttgart, Germany, 1994.

9. Gelbukh, A.F. *Using a semantic network for lexical and syntactical disambiguation*. Proc. of CIC'97, Nuevas Aplicaciones e Innovaciones Tecnologicas en Computacion, Simposium internacional de computacion, CIC, IPN, Mexico D.F., 1997.