Использование эмпирических весов при синтаксическом анализе

Л.Л. Иомдин, В.Г. Сизов, Л.Л. Цинман

Институт проблем передачи информации РАН,
101447, Москва, Большой Каретный пер., 19
iomdin@iitp.ru
тел. (095) 299 4927, факс (095) 209 0579

Ключевые слова: автоматический синтаксический анализ, машинный перевод, эмпирические веса и приоритеты

Empirical Weights in Parsing

Leonid L. Iomdin, Victor G. Sizov, Leonid L. Tsinman

Abstract

The paper discusses a complex of solutions aimed at ambiguity resolution in the parsing component of a multipurpose NLP system, ETAP-3. The main idea is to introduce a system of priorities, or weights, dynamically assigned to the elements of the text processed and of the structure generated during all parsing phases. These weights, empirically assigned by linguists to lexical entries and fragments of parsing rules, help tune the parser to the generation of an optimal syntactic structure of an ambiguous sentence.

1. Вводные замечания

Хотя проблема разрешения лексической и синтаксической неоднозначности в задачах автоматической обработки текстов на естественном языке по существу возникла одновременно с постановкой первой такой задачи, она до сих пор не получила сколько-нибудь удовлетворительного общего решения. Вполне вероятно, что такого решения вообще не существует, поскольку эта проблема, изучению которой в последние годы посвящено рекордное количество работ, является одной из самых сложных среди тех, что стоят перед любым текстовым анализатором, если вообще не самой сложной.

Чтобы добиться надежного разрешения неоднозначности в процессе анализа текста, во многих случаях необходимо располагать не только весьма нетривиальными данными семантики (степень формализации которой, достигнутая в компьютерной лингвистике, пока далеко не достаточна), но и разнообразной информацией о мироустройстве (возможность формального представления которого до сих пор пренебрежимо мала).

Такое положение дел с разрешением неоднозначности носит универсальный характер и, вообще говоря, не зависит ни от лингвистической модели, лежащей в основе текстового анализатора, ни от типа приложения, в котором этот анализатор применяется (ср., например, красноречивое изложение ситуации в недавней работе Oepen *et al*. 2000, 3ff). Это и не удивительно: то, что с легкостью делает человек в процессе понимания текста, опираясь при выборе интерпретации неоднозначных элементов на здравый смысл, знания о мире и широкий контекст коммуникации, пока недоступно никаким компьютерным системам лингвообработки.

Сказанное, однако, отнюдь не означает, что попытки решить проблему неоднозначности вообще бесперспективны и их следует отложить до лучших времен. Частичное решение этой проблемы вполне достижимо, и любая система анализа естественного языка располагает арсеналом более или менее действенных средств,

направленных на сокращение неоднозначности в ходе обработки текста – от простого игнорирования редких или нетипичных, по мнению разработчиков систем, изолированных языковых элементов, до использования масштабных статистических процедур, определяющих частотность встречаемости отдельных слов. По нашему мнению, наиболее успешно неоднозначность разрешается тогда, когда соответствующие механизмы в максимальной степени лингвистически обоснованы.

В настоящей работе излагается комплекс решений, применяемых при анализе естественного языка для разрешения языковой неоднозначности многоцелевым лингвистическим процессором ЭТАП-3, разрабатываемым в Лаборатории компьютерной лингвистики ИППИ РАН (см. об этом процессоре и его компонентах, в частности, Апресян и др. 1989, Апресян и др. 1992, Иомдин-Цинман 1997, Богуславский и др. 2000). Стержнем этого комплекса является система приоритетов, динамически присваиваемым элементам обрабатываемого текста и генерируемой структуры на разных этапах анализа. Эти приоритеты, с помощью которых в случае неоднозначности исходного текста анализатор ЭТАПа-3 ориентируется на построение оптимальной структуры, вычисляются благодаря разветвленному механизму весов, эмпирически приписываемых лингвистами словарным единицам, синтаксическим правилам и их фрагментам в виде набора специальных инструкций.

2. Постановка задачи

Прежде чем приступать к характеристике средств, используемых для разрешения неоднозначности в ЭТАПе, напомним вкратце существо проблемы.

Любой естественный язык, как хорошо известно, неоднозначен по своей природе. Всякий текст, созданный на естественном языке, независимо от его жанра, характера, предметной области или размера, практически неизбежно содержит элементы, которым соответствует более одного смысла. В письменном тексте (в дальнейшем речь пойдет только о них) неоднозначными могут быть, в частности,

- словоформы (форма *весла* может быть родительным падежом единственного числа, именительным падежом множественного числа или винительным падежом множественного числа лексемы ВЕСЛО) – в таких случаях говорят о **морфологической омонимии**;
- лексемы (МИР1 ‘вселенная’ vs. МИР2 ‘отсутствие войны’, ПОКОЙ1 ‘спокойствие’ vs. ПОКОЙ2 ‘комната’, УЧЕНИЕ1 ‘доктрина’ vs. УЧЕНИЕ2 ‘тренировка’, УЧИТЬ1 ‘обучать’ vs. УЧИТЬ2 ‘изучать’) – в таких случаях говорят о **лексической омонимии** или **лексической многозначности (**различия между омонимией и многозначностью в контексте рассматриваемой здесь проблемы не существенны);
- конструкции (*Российской сборной нельзя проиграть* ≈ ‘Российская сборная обязана не проиграть’ vs. ‘Невозможно потерпеть поражение от российской сборной’) - в таких случаях говорят о **синтаксической омонимии.**

Чаще всего в реальных текстах разные типы неоднозначности присутствуют в сочетании. Например, словоформа *стекла* может относиться к лексеме существительного СТЕКЛО и к лексеме глагола СТЕКАТЬ, словоформа *стекли* – к двум глагольным лексемам – СТЕКАТЬ и СТЕКЛИТЬ, причем грамматические характеристики у разных интерпретаций этих словоформ не совпадают. В таких случаях говорят о **лексико-морфологической** или **лексико-грамматической омонимии.** В коротеньком предложении

(1) *ВВС ожидают сокращения,*

которое может интерпретироваться либо как (1а) ‘Военно-воздушные силы ожидают, что произойдет сокращение’, либо как (1б) ‘Военно-воздушным силам предстоят сокращения’ имеет место, в частности (i) морфологическая омонимия словоформ *ВВС* (им. пад. в (1а) vs. вин. пад. в (1б)) и *сокращения* (род. пад. ед. ч. в (1а) vs. им. пад. мн. ч. в (1б)), (ii) синтаксическая омонимия (слова *ВВС* и *сокращения* выступают то как подлежащее, то как дополнение) и (iii) лексическая омонимия глагола ОЖИДАТЬ1 ‘находиться в состоянии ожидания’ и ОЖИДАТЬ2 ‘быть вероятным в близком будущем’[1]. В таких случаях можно говорить о **лексико-синтаксической омонимии**.

В самой общем виде задача разрешения неоднозначности в системе автоматического анализа естественного языка может быть сформулирована так: из всех возможных интерпретаций элементов анализируемого текста система должна выбрать ту или те, которые выбрал бы при восприятии этого текста человек. Соответственно, уровень разрешения неоднозначности в системе тем выше, чем ближе на пути к этому идеалу она находится. Это, в частности, означает, что система автоматического анализа должна в максимально возможной степени моделировать если не сам процесс восприятия текста человеком, то его результат: коль скоро авторы системы не могут для разрешения неоднозначности впрямую использовать знания о мире или нетривиальную семантическую информацию, они должны разработать некие компенсаторные механизмы, способствующие достижению этой цели. Именно этот подход принят в системе ЭТАП-3.

3. Система приоритетов в лингвистическом процессоре ЭТАП-3

3.1. Краткая характеристика процессора

Как уже отмечалось, ЭТАП-3 – многоцелевая система автоматической обработки текстов. Она включает в себя несколько крупных компонентов, в том числе автоматический перевод с английского языка на русский и обратно, систему общения с базами данных на естественном языке, систему синонимического перифразирования, перевод текста с универсального сетевого языка (UNL) на естественный язык, систему синтаксической коррекции русского текста и некоторые другие. Хотя комплекс решений, направленных на сокращение неоднозначности, применяется во всех этих компонентах, для определенности мы будем рассматривать лишь систему автоматического перевода, причем сосредоточимся на одном его направлении – русско-английском. Тем самым все решения будут иллюстрироваться на материале русского языка – входного языка системы русско-английского перевода.

Следует сразу же отметить, что анализ в процессе перевода в ЭТАПе-3 (как, впрочем, и обработка текстов во всех других его компонентах) осуществляется пофразно. Это означает, что в ходе анализа текста система не может выходить за пределы одного предложения. Соответственно, при разрешении неоднозначности система, к сожалению, не может воспользоваться предшествующим контекстом, в том числе и в тех случаях, когда, в каком-либо предыдущем предложении неоднозначность была успешно разрешена. Тем самым, говоря о моделировании человеческого восприятия текста, мы по существу вынуждены ограничиться ситуацией, когда человек воспринимает изолированные предложения.

Процесс анализа всякого предложения состоит из двух основных этапов: (1) морфологического анализа (МорфА) и (2) синтаксического анализа (СинтА).

---

[1] Любопытно, что несмотря на все эти различия, интерпретации (1а) и (1б) ситуативно практически равнозначны, так что в данном конкретном случае без разрешения неоднозначности система анализа текста могла бы и обойтись.

Результатом работы МорфА является морфологическая структура (МорфС) предложения, т.е. линейная последовательность МорфС каждой входящей в предложение словоформы – это, коротко говоря, имя лексемы, к которой принадлежит словоформа, плюс набор словоизменительных грамматических характеристик. Если словоформа неоднозначна, то ее МорфС представляет собой множество МорфС всех интерпретаций словоформы. Например, МорфС предложения

(2) *Необходимо знать пол ребенка*

имеет вид

(2а)

НЕОБХОДИМЫЙ$_{A,кр,сред,ед}$ — ЗНАТЬ1$_{V,инф,несов}$ — ПОЛ1$_{S,им,ед}$ — РЕБЕНОК $_{S,род,ед}$

ПОЛ1$_{S,вин,ед}$ — РЕБЕНОК $_{S,вин,ед}$

НЕОБХОДИМО$_{ADV}$ — ЗНАТЬ2$_{S,им,ед}$ — ПОЛ2$_{S,им,ед}$

ЗНАТЬ2$_{S,вин,ед}$ — ПОЛ2$_{S,вин,ед}$

Результатом работы СинтА является синтаксическая структура (СинтС) предложения – дерево зависимостей, в узлах которого стоят все слова предложения, а ветви помечены именами синтаксических отношений. Например, СинтС предложения (2) имеет вид

(2б)

Как видно из (2а), при морфологическом анализе никакого разрешения неоднозначности не происходит; тем самым вся деятельность в этом направлении сосредоточена на этапе СинтА: результирующее дерево зависимостей состоит из стопроцентно однозначных объектов (ср. 2б). Отсюда, в частности, вытекает, что всякому неоднозначному предложению соответствует более одной СинтС; поэтому одной из важнейших целей работы СинтА является упорядочение построения множества СинтС для обрабатываемого предложения таким образом, чтобы наиболее адекватные СинтС порождались п е р в ы м и. Этап СинтА, на вход которого поступает МорфС предложения, в свою очередь, распадается на несколько подэтапов:

1) подэтап предсинтаксического анализа, при котором производятся некоторые вспомогательные операции и разрешается значительная часть морфологической и лексической неоднозначности;
2) подэтап набора синтаксических гипотез – минимальных поддеревьев (два узла, связанных синтаксическим отношением) будущего дерева зависимости; этот набор производится в результате применения всех линейных условий корпуса синтаксических правил (синтагм),
3) подэтап выбора вершины будущего дерева зависимости,
4) подэтап построения дерева зависимостей – по сути дела, отбор правильных синтаксических связей из полученного на подэтапе 2) полного набора гипотез, производимый с помощью разнообразных фильтровых механизмов.

Разрешение неоднозначности производится на каждом из четырех подэтапов СинтА. Теперь мы можем перейти к непосредственной характеристике используемых в ЭТАПе-3 средств, с помощью которых эта задача решается.

3.2. Общие принципы

В ходе СинтА ЭТАП-3 фактически оперирует объектами двух сортов – с л о в о ф о р м а м и и с и н т а к с и ч е с к и м и г и п о т е з а м и (будущими синтаксическими отношениями, имена которых фигурируют в дереве СинтС). При

первом появлении каждый из этих объектов, по умолчанию, имеет нормальный приоритет (условно говоря, 0). В процессе работы анализатора этот приоритет может повышаться или понижаться. Это достигается главным образом за счет применения специальных правил, зона действий которых содержит инструкции, уменьшающие или увеличивающие текущий приоритет объекта на единицу. Поскольку к одному объекту в ходе работы анализатора на разных его подэтапах может применяться несколько правил, общий приоритет объекта может увеличиться или сократиться на несколько единиц (на практике он варьируется от –3 до +3, причем этих крайних значений приоритет достигает весьма редко, в подавляющем большинстве случаев оставаясь в пределах от –1 до +1).

На каждом подэтапе СинтА алгоритм самым непосредственным образом учитывает текущие приоритеты объектов, поочередно отфильтровывая те из них, которые имеют наименьший вес. Предположим, например, что к моменту, когда алгоритм приступает к 4 подэтапу – отбору правильных синтаксических гипотез для дерева зависимостей – какая-то группа гипотез *A* имеет приоритет –2, какая-то другая группа гипотез *B* – приоритет –1, а все остальные (группа *C*)– более высокий приоритет. В таком случае алгоритм сначала стирает все гипотезы группы *A* и пытается построить дерево зависимостей из оставшихся гипотез. Если после окончания работы всех фильтровых механизмов подэтапа 4 гипотез все равно остается чересчур много для того, чтобы однозначно построить дерево зависимостей, алгоритм стирает все гипотезы группы *B* и пытается построить дерево из гипотез группы *C*. Если на каком-то шаге итерации алгоритм приходит к ситуации, когда дерево зависимостей построить не удается, алгоритм как бы отступает на один шаг, восстанавливает некоторые из стертых гипотез, начиная с тех из них, которые имеют больший приоритет, и снова включает работу фильтровых механизмов.

На всех стадиях работы алгоритма СинтА действует следующий общий принцип: чем раньше какие-то структурные элементы с низшим приоритетом исключаются из рассмотрения, тем позже они восстанавливаются, если алгоритму приходится отступать назад.

3.3. Примеры использования системы приоритетов

Несмотря на то, что средства, устанавливающие приоритеты лингвистических объектов в ходе анализа текста, были в полном объеме введены в анализатор ЭТАПа-3 всего несколько месяцев назад, опыт применения этих средств для разрешения неоднозначности оказался исключительно плодотворным. Приведем в качестве иллюстрации несколько характерных примеров.

3.3.1. Усиление /ослабление морфологических и лексических омонимов в процессе предсинтаксического анализа

Удивительным образом, существенные трудности при анализе текста возникают вследствие того, что некоторые из весьма часто встречающихся лексических единиц омонимичны другим лексическим единицам. Эти трудности усугубляются тогда, когда омонимичные друг другу словоформы принадлежат к лексическим единицам разных частей речи. Показательными примерами являются случаи омонимии таких словоформ, как *для* (предлог vs. деепричастная форма глагола *длить*), *при* (предлог vs. императив глагола *переть*), *три* и *пять* (числительное vs. императивы глаголов *тереть* и *пятить*) *на* (предлог vs. междометие со значением ‘возьми’) и т.п. Если мы будем считать все интерпретации таких словоформ имеющими одинаковый вес, то в процессе анализа мы рискуем получить в числе первых совершенно экзотические СинтС, практически

полностью игнорируемые при восприятии текста человеком. Чтобы избежать этого, на практике оказывается достаточным просто уменьшить приоритет вторым элементам каждой из приведенных пар уже на первом из подэтапов СинтА, непосредственно обрабатывающим МорфС предложения. Технически это делается следующим образом: в словарные статьи слов, парадигмы которых содержат “редкий” омоним, вводится правило, проверяющее омонимичность этого омонима “частому” омониму, принадлежащему парадигме первого элемента пары и приписывающее редкому омониму низший приоритет. Возврат к такому омониму, если и произойдет, то на весьма далеком шаге алгоритма.

Аналогичное решение применяется в тех многочисленных ситуациях, когда внутри парадигмы одной лексемы существуют словоформы, появление которых в тексте маловероятно, причем эти словоформы омонимичны регулярно встречающимся словоформам той же лексемы. Примерами могут служить существительные *отечество* и *единство*, практически не встречающиеся в формах множественного числа, глагольные формы типа *расформирует*, практически не выступающие в несовершенном виде, и многие другие. В словарные статьи таких слов (каких в языке насчитываются тысячи!) вводятся ссылки на трафаретные правила предсинтаксического анализа, понижающие приоритет всем маловероятным формам и резко сокращающие пространство объектов – кандидатов на появление в СинтС обрабатываемого предложения.

Наконец, весьма часто на этапе предсинтаксического анализа применяются правила, усиливающие или ослабляющие те или иные лексические омонимы в зависимости от жанра текста, подлежащего обработке. В качестве примера приведем курьезную ситуацию из практической работы русско-английского перевода ЭТАПа-3, объектом которого являлись тексты общественно-политического характера (конкретно говоря, сетевая лента новостей ИТАР-ТАСС). Предложение

(3) *В 1999 году в ФРГ переехало 95 тысяч этнических немцев*

было переведено как

(3а) *In 1999 in the Federal Republic of Germany 95 thousand ethnic Germans were run over*

(т.е. ‘в 1999 году в ФРГ было задавлено (автомобилями?) 95 тысяч этнических немцев’. Этот грустный, хотя, по счастью, и не отвечающий реальности, результат перевода в принципе совершенно законен, поскольку соответствует вполне допустимой СинтС (вершиной которой является безличный глагол *ПЕРЕЕХАТЬ2*, а группа *95 тысяч этнических немцев* – дополнение при этом глаголе). Такого эффекта, однако, легко избежать, если уменьшить относительный приоритет этого достаточно разговорного глагола относительно глагола *ПЕРЕЕХАТЬ1* ‘сменить место жительства’.

3.3.2. Усиление лексических омонимов, поддержанных устойчивой лексической сочетаемостью

На третьем и четвертом подэтапах СинтА весьма часто применяются правила, повышающие приоритет лексических омонимов в случаях, если вероятно их вхождение в разнообразные устойчивые словосочетания. Например, если в предложении появляется сочетание типа *наносить поражение*, разумно повысить приоритеты и глаголу *наносить2* ‘причинять, делать фактом’, и существительному *поражение1* ‘разгром’ по сравнению с любыми другими значениями этих слов. В подобных ситуациях непосредственно после работы 2 подэтапа СинтА применяется правило повышения приоритета лексически связанным омонимам, активно эксплуатирующее понятие лексической функции И.А. Мельчука и опирающееся на тот факт, что между

соответствующими омонимами установлена гипотетическая связь соответствующего типа. Похожий прием применяется и тогда, когда нужно повысить приоритет лексической единицы, входящей в состав терминологического словосочетания. Например, абстрактное прилагательное *исполнительный1* получает больший приоритет, чем качественное прилагательное *исполнительный2* в случаях, когда оно входит в состав терминов типа *исполнительный директор, исполнительный секретарь СНГ* и т.п.

3.3.3. Усиление/ослабление синтаксических гипотез

В заключение приведем пример правила, ослабляющего некоторые синтаксические гипотезы, установленные алгоритмом СинтА. Опять-таки обратимся к практике работы русско-английского перевода ЭТАПа-3. При обработке предложения

(4) *В Петербурге отметили годовщину восстания декабристов 1825 года*

система выдала в качестве первого варианта достаточно загадочный перевод

(4а) *Approximately 1825 Decembrists of the year have celebrated in Petersburg an anniversary of uprising* (т.е. ‘приблизительно 1825 декабристов года отметили в Петербурге годовщину восстания’).

Как и в (3а), этот перевод соответствует вполне законной СинтС для (4), в которой, в частности, имеет место т.н. аппроксимативно-количественная синтаксическая связь между существительным *декабристов* и числительным *1825* – т.е связь, свойственная разговорным конструкциям типа *литров пять*. Мы не погрешим, однако, против истины, если выскажем предположение, что в таких конструкциях практически не встречаются числительные, записанные цифрами. Введя соответствующее условие в синтаксическое правило, формирующее такую связь, мы сможем активировать особое правило ослабления синтаксических гипотез, действующее на 4 подэтапе СинтА. В результате перевод типа (4а) уступит место вполне ожидаемому

(4б) *An anniversary of uprising of Decembrists of 1825 has been celebrated in Petersburg.*

а сам может быть получен на далеком шаге обработки предложения. Именно такая ситуация, в конечном счете, является целью описанного здесь комплекса решений.

Литература

Апресян и др. 1989: Апресян, Ю.Д., И.М.Богуславский, Л.Л.Иомдин, А.В.Лазурский, Н.В.Перцов, В.З.Санников, Л.Л,Цинман. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989.

Апресян и др. 1992: Апресян, Ю.Д., И.М.Богуславский, Л.Л.Иомдин, А.В.Лазурский, Л.Г.Митюшин, В.З.Санников, Л.Л,Цинман. Лингвистический процессор для сложных информационных систем. М.: Наука, 1992.

Богуславский и др. 2000: Богуславский, И.М., Л.Л.Иомдин, Л.Г.Крейдлин, Н.Е.Фрид, И.Л.Сагалова, В.Г.Сизов. Модуль универсального сетевого языка (UNL) в составе системы ЭТАП-3. // Труды международного семинара Диалог’2000 по компьютерной лингвистике и ее приложениям. М., 2000, 48-58.

Иомдин-Цинман 1997: Л.Л.Иомдин, Л.Л.Цинман. Лексические функции и машинный перевод. // Труды международного семинара Диалог’97 по компьютерной лингвистике и ее приложениям Диалог’97. М., 1997, 291-297.

Oepen *et al*. 2000: Oepen, Stephan, Dan Flickinger, Hans Uszkoreit, Jun-Ichi Tsujii. Introduction to this Special Issue. / Natural Language Engineering. Special Issue on Efficient Processing with HPSG: Methods, Systems, Evaluation. 6 (1), 1-14.