ИСПОЛЬЗОВАНИЕ РАЗМЕЧЕННОГО КОРПУСА ТЕКСТОВ ПРИ АВТОМАТИЧЕСКОМ СИНТАКСИЧЕСКОМ АНАЛИЗЕ[1]

И. М. Богуславский, Л. Л. Иомдин, В. Г. Сизов, И. С. Чардин
Институт проблем передачи информации РАН
{bogus,iomdin,sizov,ic}@cl.iitp.ru

Ключевые слова: обработка текстов на естественном языке, машинный перевод, синтаксический анализ и синтез, разрешение неоднозначности, теоретическая грамматика русского языка

Предлагается комбинированный алгоритм синтаксического анализа, используемый в лингвистическом процессоре ЭТАП-3 и, в первую очередь, в системе машинного перевода. При разрешении языковой неоднозначности составляющие ядро процессора эвристические правила динамически взаимодействуют со специально разработанным статистическим модулем, который на основе данных корпуса текстов с синтаксической разметкой приписывает веса гипотетическим синтаксическим связям. Для сбора корпусных данных были использованы русские тексты с синтаксической разметкой общим объемом в 6900 предложений (около 104000 слов). В ходе экспериментов по машинному переводу текстов с русского на английский язык с помощью данного комбинированного алгоритма выявлены локальные улучшения в работе лингвистического процессора, стимулирующие качественное развитие синтаксического анализатора и открывающие перед его разработчиками новые перспективы. В то же время количественное сравнение результатов работы комбинированного и эвристического алгоритмов синтаксического анализа не показало существенных различий в результатах их работы.

Вводные замечания

Не будет преувеличением сказать, что проблема разрешения языковой неоднозначности по сложности занимает первое место среди всех задач автоматической обработки текстов. Системы автоматического перевода не являются тут исключением: сколь бы тщательно ни были разработаны грамматики такой системы и ее словари, анализирующие модули системы на любом этапе непрерывно сталкиваются с необходимостью выбора вариантов анализа текстового материала – будь-то на уровне морфологии, синтаксиса или семантики. При этом по большому счету неважно, имеем мы дело с реальной омонимией обрабатываемого фрагмента текста или же с неоднозначностью мнимой, возникающей из-за несовершенства действующих в системе правил – тем более, что между этими разновидностями неоднозначности нет резкой границы (то, что представляется неоднозначным в коротком сегменте текста, может получить разрешение в более широком контексте; напротив, вполне недвусмысленное высказывание может оказаться омонимичным при его расширении, ср. однозначное *Я ждал его* и неоднозначное *Я ждал его вчера*).

[1] Авторы выражают признательность Российскому фонду фундаментальных исследований, благодаря поддержке которого (гранты 01-07-90405 и 02-06-80085) эта работа могла быть выполнена.

Неудовлетворительная ситуация с разрешением неоднозначности в системах автоматической обработки текста носит универсальный характер и, вообще говоря, не зависит от лингвистической модели, лежащей в основе текстового анализатора. Удивляться здесь не приходится: 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. Информация о мироустройстве, по-видимому, вообще не поддается сколько-нибудь масштабной формализации. Да и степень формализации чисто языковой семантики, достигнутая в компьютерной лингвистике, пока далеко не достаточна для того, чтобы эксплицировать те нетривиальные сведения о смысле высказывания, которыми необходимо располагать для разрешения неоднозначности текста.

Сказанное, однако, отнюдь не означает, что попытки решить проблему неоднозначности при автоматическом переводе вообще лишены перспективы. Частичное решение этой проблемы вполне возможно, и всякая система анализа естественного языка располагает арсеналом более или менее действенных средств, направленных на сокращение неоднозначности в ходе обработки текста – от простого игнорирования редких лексических единиц или синтаксических конструкций до использования масштабных статистических процедур, определяющих частотность встречаемости отдельных языковых элементов.

В ходе развития лингвистического процессора ЭТАП-3 и в первую очередь модуля автоматического перевода (Апресян *и др.* 1989, 1992) его авторы самостоятельно или в сотрудничестве с другими коллегами теоретически разрабатывали гибридную стратегию, сочетающую правиловый и статистический подход к анализу текста (Carl *et al.* 2000, Streiter *et al.* 2000a), а также предпринимали попытки практически реализовать эту стратегию: в частности, был предложен статистический способ использования двуязычных корпусов текстов для оптимизации выбора переводных эквивалентов словосочетаний (Streiter *et al.* 2000b) и разработана система статистически обоснованных приоритетов, динамически приписываемых элементам строящейся структуры предложения на разных этапах синтаксического анализа и ориентирующая анализатор на построение оптимальной структуры (Иомдин *и др.* 2001, Iomdin *et al.* 2002).

В настоящей работе делается попытка построить комбинированный алгоритм синтаксического анализа для русского языка, сочетающий в себе правиловую стратегию и статистическую информацию, собранную на основе корпуса русских текстов с синтаксической разметкой.

Размеченный корпус ИППИ РАН

В течение нескольких последних лет Лаборатория компьютерной лингвистики ИППИ РАН разрабатывает первый в истории синтаксически размеченный корпус русских текстов (Boguslavsky *et al.* 2000, Богуславский и др. 2001). Научную и практическую ценность корпусу придает глубина аннотации текста: каждое его предложение, помимо морфологической разметки, снабжено синтаксической структурой (СинтС) в виде дерева зависимостей. Корпус текстов строится полуавтоматически – каждое предложение вначале пропускается через синтаксический анализатор системы ЭТАП-3, а затем полученная СинтС проверяется и при необходимости корректируется редакторами – экспертами-лингвистами, так что полученная разметка текста представляет собой вполне качественный продукт. Для удобства работы редактора используется специальный

программный комплекс, состоящий из модуля разбивки текста на фразы и графического редактора структур, позволяющего легко и быстро модифицировать древесные объекты. Весь синтаксический корпус представлен с помощью формализма, разработанного на основе языка XML. Он хорошо совместим с формализмом TEI, признанным международным стандартом для языков разметки, и обеспечивает удобный интерфейс корпуса с другими приложениями. (О применении корпусов с синтаксической разметкой см. Чардин 2003.)

В настоящее время общий объем корпуса ИППИ РАН составляет около 12 000 размеченных предложений, или свыше 180 000 словоупотреблений.

Комбинированный алгоритм

Чтобы нагляднее охарактеризовать суть комбинированного алгоритма синтаксического анализа (СинтА), включающего компонент корпусной статистики, коротко напомним основные стадии главного, правилового алгоритма, используемого в системе ЭТАП-3.

В процессе СинтА текста ЭТАП-3 использует результаты морфологического и предсинтаксического анализа. Единицей СинтА всегда является предложение.

Морфологический анализ в ЭТАПе-3 осуществляется без учета какого-либо контекста. Это означает, например, что слово *дуло* в предложении *Из щели дуло* будет представлено в морфологической структуре как пара объектов {(ДУТЬ, прош, несов, сред, ед);(ДУЛО, ед, род)}.

На стадии предсинтаксического анализа происходит частичное разрешение лексической и грамматической омонимии по линейному контексту. В частности, в данном предложении блок предснтаксического анализа оставит для *дуло* лишь глагольную интерпретацию, а именная интерпретация будет исключена.

Результатом работы блока собственно синтаксического анализа является построение дерева зависимостей для исходного предложения, в котором каждый узел соответствует одному, и только одному, слову предложения, а все ветви помечены именами синтаксических отношений.

На начальной стадии этого блока синтаксические правила, или синтагмы, порождают множество минимальных поддеревьев (два узла, связанных синтаксическим отношением) – своего рода кирпичиков, из которых будет построено все дерево зависимостей. Затем осуществляется выбор вершины будущего дерева зависимостей и его непосредственное построение. На этой стадии происходит снятие оставшейся после предсинтаксического анализа лексической и грамматической неоднозначности, а также разрешение синтаксической неоднозначности (иными словами, удаление части минимальных поддеревьев). Здесь используется система различных фильтров, центральную роль в которой играет механизм приоритетов.

В этот момент в действие вступает новый статистический блок, играющий роль дополнительного фильтра. Основные идеи, легшие в основу его создания, были предложены в работе (Чардин 2001). Статистический блок взвешивает минимальные поддеревья, а также цепочки минимальных поддеревьев длиной в три слова в пространстве поиска алгоритма синтаксического анализа на основании частоты встречаемости фрагментов такого вида в деревьях зависимостей корпуса. В итоге в систему возвращаются новые значения приоритетов связей, которые вычисляются с учетом вновь приписанных весов.

Цепочки длиной в три слова взвешиваются в случаях, когда между двумя словами сформированы две или более альтернативных синтаксических связи и у

них нет альтернатив на один уровень выше по дереву. Если такие альтернативы есть, то взвешиваются минимальные поддеревья. Такой подход связан с тем, что в противном случае пришлось бы учитывать условную вероятность полученных приоритетов, что привело бы к чрезмерному усложнению работы. Цепочки минимальных поддеревьев можно рассматривать как n-граммы 2-го и 3-го порядка, или, соответственно, древесные биграммы и триграммы. Однако требование приписывания весов конкретным связям не позволяет в полной мере реализовать преимущества n-граммной модели.

Значения статистических приоритетов связей заведомо меньше единицы (поскольку за единицу принимается вся совокупность встречаемостей конкретных конфигураций в корпусе), в то время как значение положительного приоритета связи, присваиваемого регулярными правилами, целочисленное и всегда не меньше единицы. Благодаря этому обстоятельству при конфликте между эвристической и статистической стратегиями в механизме учета приоритетов предпочтение отдается результатам работы эвристических правил, созданных экспертами-лингвистами.

Следует отметить, что при наличии соответствующих корпусных данных предлагаемый комбинированный алгоритм может использоваться при СинтА предложений не только на русском, но и на английском языке (как, впрочем, и на любых других языках, анализ которых осуществляется на принципах ЭТАПа).

Реализация алгоритма

Описанный комбинированный алгоритм был реализован в виде особого модуля системы ЭТАП-3, написанного на языке C++. В графическом интерфейсе системы присутствует кнопка-флажок, позволяющая включать и выключать этот модуль, а ход его работы можно проследить по общему протоколу работы системы.

Конфигурация модуля задается через текстовые файлы вида {b|t}_stat.txt, где "b", и "t" соответственно обозначают биграммы и триграммы. Конфигурационные файлы представляют собой массивы строк. Простой пример строки из файла b_stat.txt – "(S)(опред)(A) 0.119527972833853". Эта запись означает, что из всех биграмм, присутствующих в корпусе, на долю биграмм, состоящих из существительного (S) и зависящего от него по определительному синтаксическому отношению прилагательного (A) на основании встречаемости в корпусе приписан вес 0.119527972833853 (что примерно соответствует 13% всех биграмм).

Реализованный статистический модуль позволяет учитывать полные или сокращенные наборы морфологических характеристик слов в предложении, имена синтаксических отношений, линейные расстояния между хозяином некоторого отношения и его слугой, а также направления отношений относительно вершины.

Лексикализация модели в текущей версии модуля не осуществлялась, хотя в перспективе разумно было бы учитывать конкретный лексический состав взвешиваемых биграмм и триграмм или же принадлежность входящих в него слов к тем или иным лексическим или семантическим классам.

Наборы данных

Для использования в экспериментах по корпусу были собраны данные трех видов. Данные первого вида ("обедненные") помимо информации о названии поверхностно-синтаксической связи включали информацию о частеречной принадлежности слова-хозяина и слова-слуги, куда, помимо, традиционных частей речи, заносились причастия и деепричастия). Данные второго вида ("умеренные")

дополнительно содержали информацию о падеже существительных. Данные третьего вида ("обогащенные") включали информацию о расширенном наборе характеристик (род, число, падеж, одушевленность) для именных частей речи, за исключением числительных, а также для причастий; для остальных глагольных форм включалась информация о репрезентации (т.е. указывалось, личный это глагол, инфинитив или деепричастие).

При сборе данных было использовано немногим более половины материалов корпуса: тренировочные тексты с синтаксической разметкой имели общий объем 6 900 предложений, или около 104 000 словоупотреблений).

Экспериментальные результаты

Серия экспериментов с комбинированным алгоритмом СинтА показала, что в большинстве случаев результаты его работы совпадают с результатами анализа, проведенного ЭТАПом-3 с отключенным корпусно-статистическим модулем. Очевидно, что в таких случаях тождественными оказываются и выполняемые системой переводы.

Тем не менее, в определенных ситуациях синтаксические структуры, произведенные комбинированным алгоритмом, отличаются от результатов работы стандартного алгоритма ЭТАПа-3. В последнем случае выполненные переводы могут различаться. (Подчеркнем, что при определенных условиях переводы не различаются даже тогда, когда СинтС входного предложения различны – например, когда различия структур сводятся к локальным расхождениям в именах синтаксических отношений, установленных между словами предложения.)

Остановимся чуть подробнее на той части результатов экспериментов по машинному переводу, в которых обнаруживаются расхождения в работе комбинированного и стандартного алгоритмов. В собранном нами экспериментальном материале имеются как примеры положительного эффекта подключения корпусно-статистического модуля, так и примеры отрицательного эффекта, когда подключение модуля ухудшает результат работы анализатора.

Приведем сначала примеры положительного эффекта внедрения статистического модуля. Для предложения

(1) *В Китае казнен владелец частного детского сада - отравитель детей*

СинтС, построенная с использованием корпусной статистики, имела вид

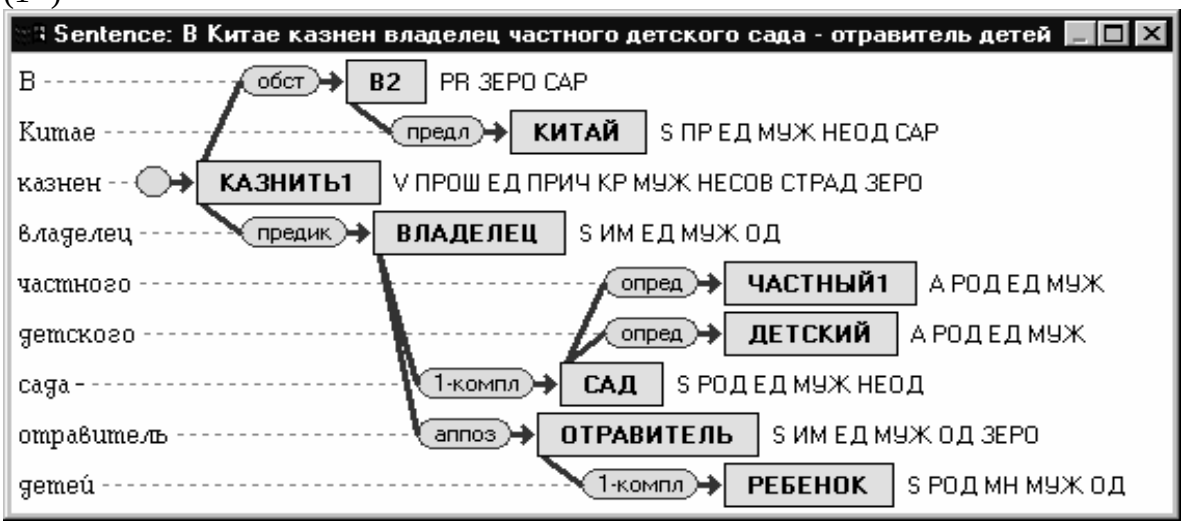
(1′)

Как легко видеть, слово *отравитель* (а с ним и именная группа *отравитель детей*) подчиняется в (1′) слову *владелец* по аппозитивному синтаксическому

отношению. Соответственно, перевод этой фразы на английский язык выглядел так: *In China the owner of a private kindergarten - the children's poisoner is executed.*

При отключении корпусно-синтаксического модуля синтаксическая структура предложения (1) имела вид

(1´´)



Здесь слово *отравитель* (и его именная группа) по разъяснительному отношению подчинено слову *сад*. Перевод, построенный для этой структуры, выглядел значительно хуже: *In China is executed the owner of a private kindergarten – the children's poisoner.*

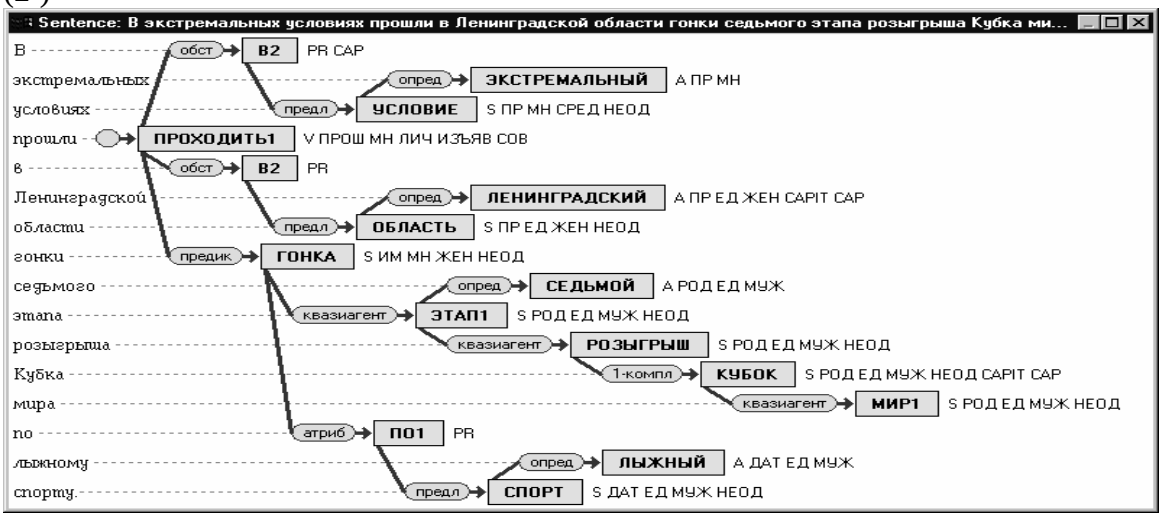
Разница в СинтС (1´) и (1´´) была обусловлена тем, что в аннотированном корпусе конфигурации из двух существительных, соединенных аппозитивным синтаксическим отношением, встречались чаще, чем конфигурации из существительных, соединенных разъяснительным синтаксическим отношением. Из двух конкурирующих отношений было выбрано более частотное.

Еще один пример положительного эффекта внедрения корпусно-статистического модуля можно было наблюдать при переводе предложения

(2) *В экстремальных условиях прошли в Ленинградской области гонки седьмого этапа розыгрыша Кубка мира по лыжному спорту.*

При включенном статистическом модуле синтаксическая структура для (2), построенная с использованием корпусной статистики, имела вид

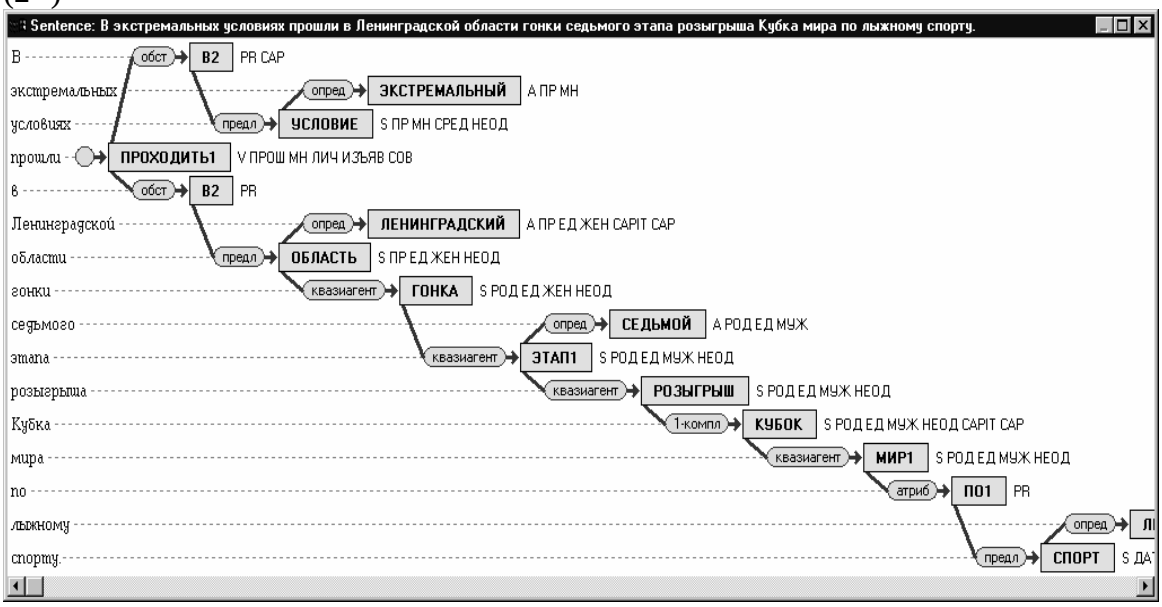
(2´)



Если отвлечься от неточности в присоединении предложной группы *по лыжному спорту* (которую следовало бы подчинить не слову *гонки*, а слову *кубок*),

СинтС (2′) была построена правильно. В частности, совершенно верно была определена синтаксическая роль существительного *гонки* как подлежащего при вершине предложения пришли. Соответственно, вполне удовлетворительным оказался и его перевод: *In extreme conditions in the Leningrad region the races of the seventh phase of playoff of the world Cup in ski sport have taken place.*

При отключении статистического модуля алгоритм построил для (2) структуру (2′′)



Как видим, в (2′′) слово *гонки* оказалось подвешенным к своему непосредственному соседу области, а вершина предложения (2) *прошли* осталось без подлежащего. Поскольку структура эта явно ошибочна, совершенно неадекватным получился и построенный на ее основе английский перевод: *In extreme conditions one elapsed in the Leningrad region of the race of the seventh phase of playoff of the world Cup in ski sport.*

При построении структуры с использованием статистического модуля была использована корпусная информация об относительно частотной встречаемости конфигурации «вершинный глагол + подлежащее в именительном падеже», в результате чего квазиагентивному синтаксическому отношению в конфигурации «существительное, подчиняющее другое существительное по квазигентивной связи» оказался приписанным меньший вес, и был предпочтен предикативный хозяин слова *гонки*.

Приведем теперь пример, когда использование корпусно-синтаксического модуля дает худший результат, чем отказ от него. Рассмотрим предложение

(3) *Кризис в КНДР не представляет непосредственной угрозы для безопасности России.*

СинтС (3′), построенная стандартным анализатором ЭТАПа, была безукоризненной. Соответственно, оказался вполне приемлемым и английский перевод (3): *The crisis in the People's Democratic Republic of Korea does not present an immediate threat for the security of Russia.*

(3′)

При подключении статистического модуля, однако, ситуация в корне меняется. Порожденная для (3) СинтС принимает вид

(3′′)

Как видим, предложная группа *в КНДР* перестала подчиняться слову *кризис* и оказалась ошибочно подвешенной к вершине предложения – глаголу представлять по 2-ой комплетивной связи. Вдобавок вместо локативного предлога *в2* (и соответственно, предложного падежа слова *КНДР*) анализатор выбрал предлог направления *в1* (и, соответственно, винительный падеж слова *КНДР*). В результате английский перевод (3) оказался неверным и совершенно неадекватным исходному тексту: *The crisis does not present to the People's Democratic Republic of Korea an immediate threat for the security of Russia* (получилось нечто вроде **Кризис не представляет Корее непосредственной угрозы для безопасности России*).

Это произошло потому, что синтаксической гипотезе

ПРЕДСТАВЛЯТЬ1 –2-компл –> В1

статистический модуль совсем немного (на 0,005412) прибавил веса из-за соответствующей частотности подобных гипотез в корпусе. Между прочим, этот модуль прибавил веса и гипотезе

КРИЗИС –атриб– > В2,

однако здесь прибавка была еще меньше и составила 0,000713.

Существенно, что различия в работе анализатора зависят от того, какой вариант корпусной статистики используется в статистическом модуле - «обедненные», «умеренные» или «обогащенные» N-граммы (см. выше). По нашему мнению, выбор между этими вариантами должен соответствовать объему корпуса, по которому

собирается статистика: ограниченный корпус, имеющийся в нашем распоряжении в настоящее время, оптимально соответствует обедненному варианту статистики. Чем богаче информация, учитываемая при сборе статистики, тем больше образуется типов биграммных и триграммных шаблонов, и встречаемость многих из возможных типов биграмм и триграмм окажется слишком малой, чтобы являться базой для принятия статистически мотивированных решений.

Статистическая оценка осуществлялась следующим образом. Около 80% задействованного корпуса было использовано при сборе данных для комбинированного алгоритма (т.е. при его обучении). Остающиеся 20% текстов суммарным объемом около 26 000 слов и 2 000 предложений использовались как тестовые данные, причем они были также разделены на две приблизительно равные части. Первая из частей использовалась не только для автоматической оценки, но и для рассмотрения и сравнения полученных структур вручную, в результате чего правила синтаксического анализатора могли подвергаться – и в действительности подвергались – некоторым изменениям. Вторая часть использовалась только для автоматической оценки. Результаты, полученные при оценке второй части являются в наибольшей степени объективными.

Количественное сравнение результатов работы комбинированного и эвристического алгоритмов анализа не выявило статистически значимых различий, вне зависимости от вида корпусных данных, использовавшегося комбинированным алгоритмом. Выяснилось, что стандартный анализатор ЭТАПа-3 правильно проводил 73-78% синтаксических связей и правильно выбирал 78-80% вершин деревьев зависимостей. 18-20% получаемых синтаксических структур совпадало со структурами корпуса.

В заключительном эксперименте была проведена дополнительная оценка на текстах сетевых новостей объемом около 29 000 слов и 1 700 предложений, размеченных в самое последнее время и представляющих собой дополнительную часть аннотированного корпуса. Результаты работы комбинированного и эвристического алгоритмов опять-таки не продемонстрировали существенных различий. В то же время на данном материале был зафиксирован рост количества правильно проведенных связей (до 78-80%) и количества правильно выбранных вершин (до 84-85%). При этом, однако, количество совпадающих структур снизилось до 9-10%.

Среди возможных причин сходства работы эвристического и статистического алгоритмов, на наш взгляд, наиболее важными оказываются две:

1) Объем аннотированного корпуса был недостаточным для получения более значимых статистических данных.

2) Внедрение статистических методов в систему оказалось недостаточно глубоким: в точке системы, где действует корпусная статистика, в пространстве поиска алгоритма синтаксического анализа содержится сравнительно немного гипотетических связей.

Последнее предположение, как нам представляется, заслуживает особого внимания. Установка на некоторое недопорождение связей в данной точке или в системе в целом в ходе разработки эвристического алгоритма анализа являлась своего рода компенсацией нехватки механизмов эффективного снятия синтаксической неоднозначности. В настоящее время авторы готовят серию экспериментов по количественной оценке порождения связей в системе посредством поиска эталонных связей корпуса во множестве всех гипотетических связей, содержащихся в пространстве поиска алгоритма синтаксического анализа на момент применения статистического модуля.

В свете полученных результатов перспективными видятся такие задачи, как расширение аннотированного корпуса, проведение экспериментов по комбинированному синтаксическому анализу с учетом длин связей в словах и направления связей, а также продуманная лексикализация описанной лингвостатистической модели.

Использование модуля комбинированного алгоритма синтаксического анализа в системе ЭТАП-3 по умолчанию на данной стадии представляется нецелесообразным.

Литература

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

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

3. Богуславский И.М., Григорьев Н.В., Иомдин Л.Л., Крейдлин Г.Е., Фрид Н.Е., Чардин И.С. Разработка синтаксически размеченного корпуса русского языка // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб, изд-во Санкт-Петербургского университета, 2002, с. 40-50.

4. Иомдин Л.Л., Сизов В.Г., Цинман Л.Л. (2001) Использование эмпирических весов при синтаксическом анализе. // Обработка текста и когнитивные технологии. Труды конференции «Когнитивное моделирование». Казань, Отечество, № 6, С. 64-72.

5. Чардин И.С. (2001) Использование аннотированного корпуса для снятия синтаксической неоднозначности в лингвистическом процессоре ЭТАП-3. Материалы 2-ой Всероссийской конференции "Теория и практика речевых исследований" (АРСО-2001). Москва, Издательство МГУ, с.26-27.

6. Чардин И.С. (2003) Лингвистические корпуса с синтаксической разметкой и их применение. // Научно-техническая информация. (В печати)

7. Boguslavsky I.M., Grigorieva S.A., Grigoriev N.V., Kreidlin L.G., Frid N.E. (2000). Dependency Treebank for Russian: Concepts, Tools, Types of Information. // Proceedings of the 18th Conference on Computational Linguistics. Vol 2, 987-991, Saarbrücken.

8. Carl M., Pease C., Streiter O., Iomdin L. (2000). Towards a Dynamic Linkage of Example-Based and Rule-Based Machine Translation // Machine Translation

9. Iomdin L., Sizov V., Tsinman L. (2002). Utilisation des poids empiriques dans l'analyse syntaxique: une application en Traduction Automatique // META, vol. 47, No 3. P. 351-358/

10. Streiter O., Iomdin L., Carl M. (2000a) .A Virtual Machine for Hybrid Machine Translation. // Труды Международного семинара Диалог'2000 по компьютерной лингвистике и ее приложениям. Т. 2. Протвино, С. 382-393.

11. Streiter O., Iomdin L., Sagalova I. (2000b). Learning Lessons from Bilingual Corpora: Benefits for Machine Translation. // International Journal of Corpus Linguistics. Vol. 5(2), pp. 199-230.

PARSING WITH A TREEBANK

I.M.Boguslavsky, L.L.Iomdin, V.G.Sizov, I.S.Chardin

Institute for Information Transmission Problems, RAS, Moscow

Key words: natural language processing, machine translation, syntactic analysis and synthesis, ambiguity resolution, theoretical grammar of Russian

A hybrid parsing algorithm has been developed and integrated into the ETAP-3 multifunctional NLP environment, to be used primarily in machine translation. When resolving language ambiguity, the heuristic rules that constitute the system's core

dynamically interact with the customized statistical module. The latter assigns weights to dependency links that constitute hypothetical parse trees employing data derived from the syntactically tagged corpus. The statistical module was trained on approximately 104000 words in 6900 sentences of syntactically annotated Russian texts. The analysis of experiments in machine translation from Russian into English with the help of the hybrid statistical module has shown local improvements in the performance of the NLP environment, which stimulates qualitative development of the parser and opens new vistas for the developers. At the same time, a quantitative comparison between the hybrid parser and the rule-based one has revealed no significant difference in their performance.