Эксперимент по построению синтаксической структуры английских предложений с использованием заранее известных фрагментарных данных.[1]

Диконов Вячеслав
ИППИ РАН
dikonov@iitp.ru

Павел Дяченко
ИППИ РАН
pavelvd@iitp.ru

Аннотация

Основная гипотеза данной работы состоит в том, что использование дополнительной информации об анализируемом предложении в процессе автоматического построения его синтаксической структуры и далее семантического графа позволяет повысить качество получаемой структуры.

Мы проверили эту гипотезу на основе материала, представляющего собой корпус английских предложений, для которого ранее были получены различные знания о свойствах отдельных слов. Обработка материала в нашей работе проводилась лингвистическим процессором ЭТАП. Получены числовые оценки количества и качества изменений в результате использования имевшихся корпусных данных при текущем уровне развития синтаксического анализатора ЭТАП, английского комбинаторного словаря и словаря UNL.

1. Введение

При построении синтаксического дерева зависимостей по тексту автоматической системой, основанной на правилах, алгоритм вынужден преодолевать ряд сложностей, связанных с существованием альтернативных вариантов разбора. При этом неоднозначность возникает на нескольких этапах – при получении морфологического разбора слов предложения, при определении лексического значения слов и при построении синтаксических связей.

Целью работы является ответ на вопрос, как можно улучшить качество синтаксических структур, получаемых с помощью синтаксического компонента лингвистического процессора ЭТАП [1,2,3] при использовании в процессе его работы некоторой заранее известной фрагментарной дополнительной информации, которую можно почерпнуть из существующих размеченных корпусов текстов. Работа проводится на материале английского языка. В качестве привлекаемой дополнительной информации используется разбиение предложений на отдельные слова, идентификаторы лексического значения - имя соответствующей слову статьи комбинаторного словаря (далее КС) системы ЭТАП - частеречные и отдельные морфологические характеристики (число, причастная форма глаголов и т.п.) некоторых слов предложения, а также сведения о том, что некоторые слова образуют группы, обозначающие одно понятие или объект реальной действительности. Примерами таких групп могут служить имена и названия: например «general Burnside», «Fulton County Grand Jury», словосочетания типа «curtain call», фразовые глаголы и т.п.

Положительный ответ на этот вопрос позволит решить две практические задачи: во-первых - автоматическим путем получить набор синтаксических структур в формате ЭТАП на материале значительного по объему корпуса английских предложений, используя все доступные средства повышения качества разбора. Такие структуры могут, к примеру, стать основой для двухступенчатого построения морфо--синтаксической разметки корпуса экспертами-лингвистами, как это делается при разработке размеченного корпуса русских текстов «СинТагРус» [10]. Вначале с помощью автоматических средств анализа текста порождается черновой вариант разметки, а затем он вручную корректируется экспертами. Это обеспечивает значительную экономию сил и времени.

Вторая практическая цель – автоматически построить корпус семантических графов в формате искусственного языка UNL [8] с помощью разрабатываемого нами UNL-конвертера. Язык UNL предназначен для записи смысла текста и позволяет исключить лексическую неоднозначность, а

[1] Работа частично поддержана средствами грантов РФФИ № 08-06-00367 и РФФИ № 07-06-00373.

также свести к минимуму возможность различного толкования текста.

Развитие этой работы позволит оценить влияние различных типов информации на процесс анализа предложения в ЭТАП и проверить возможность повышения числа правильных разборов. Кроме того, работа позволяет расширить возможности системы по использованию и дополнению частичной информации о предложении.

2. Материал исследования

Исходным материалом является свободно доступный в Интернет корпус английских предложений Semcor [4,5], который содержит 37176 предложений из корпуса Brown, и тестовый набор предложений для соревнований по разрешению лексической неоднозначности Senseval (далее все вместе называется "корпус"). Формат данных в обоих случаях одинаков. Данный материал был выбран потому, что он содержит разметку значений слов концептами словаря Wordnet v2.1. Создаваемый нами словарь UNL имеет карту соответствий с концептами Wordnet, что позволяет отождествить слова предложения с лексическими единицами UNL и далее со статьями комбинаторного словаря ЭТАП, тем самым получая доступ к лингвистической информации еще до начала морфологического и синтаксического разбора.

Помимо семантической разметки Semcor также содержит частеречные метки, автоматически расставленные парсером BRILL, однако разметка не включенных в семантическую аннотацию служебных слов не является надежной. Одновременно, в Интернете имеются и другие варианты разметки тех же самых данных, в частности данные о важнейших синтаксических связях, предоставляемые исследовательской группой IXA из университета Страны басков [9].

2.1. Типы разметки корпуса

Корпус содержит леммы большинства слов и разметку двух видов: морфологическую и семантическую. Морфологическая разметка представлена метками в формате парсера BRILL [6,7]. Они содержат информацию о части речи и важнейших морфологических характеристиках слов (См. Таблицу 1).

BRILL	**ЭТАП**	**Толкование**	**BRILL**	**ЭТАП**	**Толкование**
MD	V	Глагол (модальный)	RBS	ADV SUP	Наречие превосходной степени
VB	V	Глагол	WRB	ADV	Наречие
VBD	V	Глагол (прошедшее время)	RP	PART	Частица
VBG	V ING	Причастие наст. вр. (Participle I)	TO	PART	Частица to
VBN	V PP	Причастие прош. вр. (Participle II)	UH	INTJ	Междометие
VBP	V	Глагол (мн.ч.)	DT	ART/A	Артикль или кванторное слово - прилагательное
VBZ	V	Глагол (3-е лицо)	WDT	S	Существительное
NN	S SG	Существительное ед.ч.	CD	NUM	Числительное
NNS	S PL	Существительное мн.ч.	EX	PART	Частица
NNP	S SG	Существительное собственное, ед.ч.	FW	NID	Неизвестное или иноязычное слово
NNPS	S PL	Существительное собственное, мн.ч.	RBR	ADV COMP	Наречие сравнительной степени
PRP	S	Существительное (местоимение)	RBS	ADV SUP	Наречие превосходной степени
PRP$	A	Прилагательное (притяж. мест.)	WRB	ADV	Наречие
WP	S	Существительное (вопросит. слово)	RP	PART	Частица
WP$	A	Прилагательное (вопросит. слово)	TO	PART	Частица to
JJ	A	Прилагательное	UH	INTJ	Междометие
JJR	A COMP	Прилагательное сравнит. степени	DT	ART/A	Артикль или кванторное слово - прилагательное
JJS	A SUP	Прилагательное превосходной степени	WDT	S	Существительное
CC	CONJ	Союз	CD	NUM	Числительное
IN	PR /CONJ	Предлог или союз	EX	PART	Частица
RB	ADV	Наречие	FW	NID	Неизвестное или иноязычное слово
RBR	ADV COMP	Наречие сравнительной степни			

Таблица 1: Используемая таблица соответствий меток BRILL и ЭТАП.

Семантическая разметка состоит из записи значений слов в терминах словаря Wordnet и выделения имен собственных. Значения слов записаны с помощью атрибута «lexsn=», значением которого являются числовые идентификаторы синсета Wordnet. У имен и названий, которых нет в Wordnet, значением этого атрибута является ссылка на синсет, обозначающий один из трех классов: людей, общностей/организаций и мест. Как видно на Рисунке 1 на следующей странице, в поле леммы у таких слов стоит идентификатор класса (PERSON/GROUP/LOCATION), который дублирует rdf-ссылка. Имена собственные, которые не имеют атрибута «lexsn», можно определить по меткам «NNP» или «NNPS» в поле значения атрибута «pos», что также можно отнести к семантической разметке.

...
<wf cmd=done **rdf=group pos=NNP lemma=group** wnsn=1 **lexsn=1:03:00:: pn=group**>Fulton_County_Grand_Jury</wf>
<wf cmd=done pos=VB lemma=say wnsn=1 **lexsn=2:32:00::**>said</wf>
…

Рисунок 1: Семантическая разметка в корпусе.

В отличие от морфологической разметки, атрибут «lexsn» с номером синсета имеют не все слова в корпусе. Семантических меток нет у служебных слов, т.е. предлогов, союзов, частиц. Кроме того, плотность этой разметки неравномерна в разных частях корпуса. Semcor делится на 3 части, представляющих собой каталоги с именами brown1 brown2 и brownmv, файлы в которых имеют различную степень подробности разметки. Если файлы из brown1 содержат семантическую разметку почти для всех значимых слов, то в brownmv она есть лишь у некоторых слов.

Важной особенностью разметки является объединение словосочетаний, которые считаются одним концептом в Wordnet, например: «abdominal nerve plexus», «absence without leave» и т.д., а также ряда именных групп — названий, части фразовых глаголов и т. п. Отдельные слова групп соединяются знаком "_", а иногда "-", а вся группа считается одним словом. В корпусе нет никаких сведений об отдельных словах внутри слитых групп и о синтаксических связях между словами.

2.2. Достоверность разметки корпуса

Различные типы разметки в Semcor различаются не только количественно по числу размеченных слов, но и качественно - по достоверности. Морфологическая разметка была сделана автоматически и потому не может считаться полностью надежной. Практика показала, что в ней имеется заметное число ошибок, причем число ошибок различно для разных классов слов. В разметке служебных слов ошибок больше. Тем не менее, в дальнейшем был найден вариант использования этих частично ошибочных данных, при котором противоречивая информация игнорировалась.

Семантическая разметка была сделана вручную и считается более достоверной. Наличие семантического атрибута «lexsn» у слова автоматически повышает достоверность информации атрибута «pos», поскольку концепты Wordnet однозначно сопоставлены с частями речи.

2.3. Доступная информация

В результате преобразования имеющейся разметки из корпуса можно получить следующие данные: а) часть речи и морфологические характеристики слов, б) леммы, в) лексические значения, г) группы слов, которые соответствуют низким уровням структуры составляющих и обычно являются именами или названиями, д) семантическую классификацию имен собственных. Однако эти сведения имеются не для всех слов. Стандартной является ситуация, когда о конкретном слове известна лишь часть из упомянутых типов данных. Степень надежности данных также различна для разных классов слов и комбинаций типов разметки.

3. Ход эксперимента

Работа в рамках эксперимента была разбита на несколько последовательных частей. Сначала необходимо было обработать корпус, т. е. извлечь содержащуюся в нем лингвистическую информацию и преобразовать ее в совместимый с принятым в лингвистическом процессоре ЭТАП формат. В результате были получены файлы, содержащие фрагментарную информацию указанных выше типов о словах каждого предложения, а также о существовании некоторых групп слов внутри предложений и значениях отдельных слов. Известные группы слов были выделены и проанализированы с помощью ЭТАП отдельно от включающих их предложений. Найденные таким образом связи были встроены в окончательную версию исходного материала, которая передавались системе ЭТАП для построения полной синтаксической структуры всех предложений корпуса. После получения полных син-

таксических структур следует этап их анализа. Все структуры, которые удовлетворяют формальным требованиям правильности с точки зрения грамматических правил ЭТАП, то есть представляющие собой цельное синтаксическое дерево без фиктивных связей, впоследствии используются для построения семантических графов UNL с помощью конвертера UNL.

3.1. Преобразование данных корпуса

Для использования корпуса в нашем эксперименте необходимо было преобразовать исходные данные корпуса и записать их в XML-файлы формата TGT (tagged text). Система ЭТАП опирается на комбинаторный словарь (КС), в котором одному слову заданной части речи могут соответствовать несколько словарных единиц - лексем, которые различаются спектром лексических значений и синтаксическими свойствами. Каждая лексема может объединять несколько разных лексических значений, которым соответствуют слова UNL и синсеты Wordnet. В ходе преобразования осуществлен переход к лексемам, что подразумевает большую полноту информации об отдельных словах.

Для осуществления конверсии корпуса была написана специальная программа на языке perl. Ее функции включают а) преобразование формата файлов, б) переход от используемых в корпусе единиц (морфологических меток, словоформ, сочетаний слов, концептов Wordnet) к единицам системы ЭТАП (лексемам английского комбинаторного словаря, морфологическим характеристикам ЭТАП) а также семантического языка UNL ("универсальным словам") и в) встраивание в полученный материал дополнительных сведений об отдельных синтаксических связях в рамках предварительно выделенных групп слов внутри предложения. Помимо файлов корпуса и таблиц соответствий морфологических меток BRILL и ЭТАП программа использует информацию из словаря UNL (имена «универсальных слов» и ссылки на соответствующие им синсеты Wordnet) и английского комбинаторного словаря ЭТАП (имена лексем, частеречную принадлежность лексем, семантические дескрипторы лексем и ссылки на соответствующие лексемам «универсальные слова» UNL). Данные словарей необходимы для поиска лексем КС, соответствующих указанным в корпусе концептам Wordnet. Кроме этого, программа может использовать файлы с записью результата автономного (отдельно от остального предложения) анализа некоторых фрагментов предложений из корпуса с помощью ЭТАП.

3.2. Сегментация предложений

В ходе синтаксического анализа в ЭТАП текущей версии не используется понятие синтаксических групп. Алгоритм синтаксического анализа строит все возможные в рамках грамматики связи и выбирает один из гипотетических наборов связей без учета границ групп. В результате возможна ситуация, когда часть входящих в одну неразрывную синтаксическую группу слов оказывается подчинена какому-нибудь слову вне группы, и тем самым оторвана от нее в результате проигрыша правильной связи на этапе конкуренции связей. Это существенным образом искажает синтаксическую структуру. Использование корпуса позволяет избежать части ошибок такого рода, поскольку корпус в явном виде указывает границы и состав некоторых синтаксических групп. Эта информация сохраняется программой преобразования данных в tgt-файлах в виде атрибута GRP. Его значением является список номеров слов, которые образуют группу. В число фиксируемых корпусом групп входят многословные имена и названия, термины, фразовые глаголы и сочетания, представленные в Wordnet. Кроме того, в качестве групп нами выделяются фрагменты предложения заключенные в скобки и кавычки. Всего в корпусе имеется более 6000 предложений с обособленными этим способом фрагментами прямой речи, примечаний и т. п.

Чтобы предупредить ошибки разрыва синтаксических групп, использовался следующий прием: выделенные группы анализировались системой ЭТАП отдельно от содержащего их предложения, полученные связи запоминались и встраивались в tgt файлы вместе с данными корпуса.

Такой способ дал хорошие результаты на фрагментах, которые могли бы считаться самостоятельными предложениями, прежде всего на фрагментах прямой речи. Однако, при изолированном разборе более коротких именных групп настроенный на стандартные предложения синтаксический анализатор может делать ошибки, связанные с неверным определением вершины группы или предпочтением существительному омонимичного глагола. Например, именная группа «Miss Grant» в стандартном режиме работы анализатора была разобрана как глагольная «to miss Grant». Для корректного разбора требуется разработать особый режим анализатора. Чтобы избежать этой проблемы, из числа выделенных корпусом и проанализированных автономно именных групп

были исключены те, чья структура включала в себя глаголы.

3.3. Последовательность преобразования

Первым шагом преобразования исходных данных является переход от меток парсера BRILL к набору характеристик ЭТАП согласно Таблице 1. При этом учитываются систематические расхождения между парсерами. Так, например, слова more и most в корпусе имеют метку JJ (прилагательное), а в ЭТАП они могут быть только наречиями. Одновременно происходит исправление явных ошибок, когда какое-либо служебное слово имеет невозможную с точки зрения словаря ЭТАП метку части речи.

Следующим шагом является определение того, какая лексема комбинаторного словаря должна быть приписана каждому из слов предложения, снабженному значением по Wordnet. При этом осуществляется переход на основании данных словарей от синсета Wordnet к «универсальному слову» UNL и затем к имени лексемы. Если слово не имеет семантической аннотации или восстановить цепь соответствий не удается, то используется поиск лексем по совокупности леммы и части речи. Это позволяет определить лексему, если в словаре нет двух и более омонимичных лексем для данной леммы, принадлежащих к одной части речи.

После этого выполняется коррекция разбиения предложения на слова. Составные предлоги и имена, которые представлены в словаре ЭТАП как единые слова, например IN$ORDER$TO, объединяются в одно слово. Склеенные воедино в корпусе группы слов для совместимости с ЭТАП разделяются на отдельные слова. Некоторые из групп являются уже известными программе в результате процедуры сегментации и предварительного разбора. В этом случае данные разбора сразу включаются в получаемый tgt-файл исходных данных.

В неизвестных группах делается попытка определения части речи и подбора лексемы КС с учетом семантической классификации группы и семантических дескрипторов лексем. Так, слово Bush в группе George_Bush имеет две омонимичные лексемы КС: BUSH1 (куст) и BUSH2 (фамилия). При наличии семантического классификатора «PERSON» выбирается фамилия на основании приписанного этой лексеме в словаре КС дескриптора «ЧЕЛОВЕК».

В результате преобразования формируется представление данных, пример которого дан на Рисунке 2.

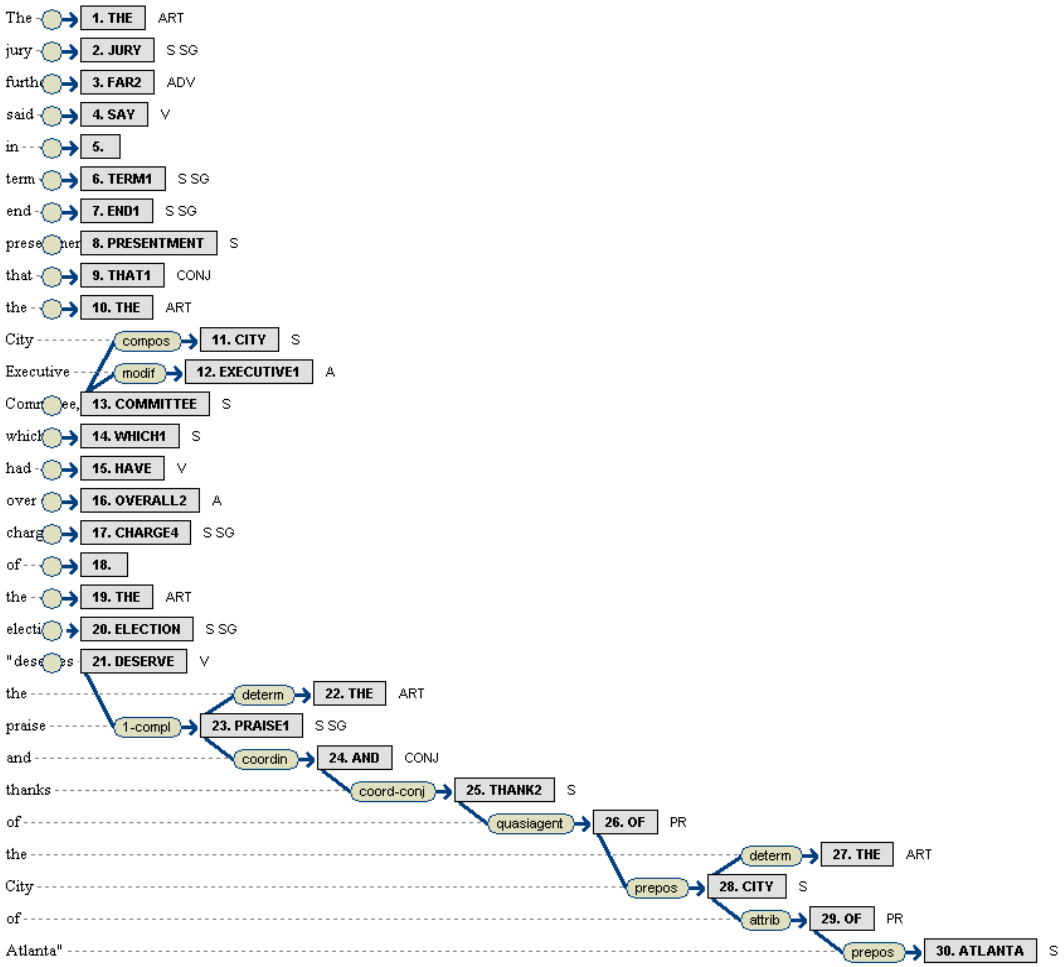

Рисунок 2: Часть предложения из корпуса после конверсии в формат ЭТАП с двумя предварительно разобранными сегментами.

Полученные на первой стадии эксперимента данные являются фрагментарными и в некоторых случаях содержат ошибки, возникающие как следствие ошибок разметки корпуса или несовершенства процесса анализа изолированных сегментов предложения.

3.4. Обработка данных процессором ЭТАП

Подготовленные данные в формате tgt в дальнейшем обрабатывались с помощью процессора ЭТАП для получения полных синтаксических структур фраз. При этом использовались возможности интеллектуальной компьютерной среды StrEd, разработанной для системы ЭТАП В. Г. Сизовым. Подробно работа StrEd и в особенности синтаксического отладчика этой среды рассмотрена в [2,3]. Для учета специфики обрабатываемого корпуса в алгоритм работы синтаксического отладчика был внесены некоторые изменения.

Ключевым моментом в работе ЭТАПа с материалом описываемого эксперимента было использование данных, которые уже были заключены внутри разметки текста, в качестве фильтра для гипотез, построенных процессором. Такая фильтрация проводилась на этапах морфологического и синтаксического анализа.

В ходе морфологического анализа ЭТАП должен определить для каждого слова предложения, какой лексической единице и в какой форме оно соответствует. При этом для дальнейшего анализа фразы необходимо существование словарной статьи данной лексической единицы в системе ЭТАП. В общем случае процессор генерирует несколько лексико-морфологических вариантов разбора, из которых в ходе дальнейшей работы процессора выбирается один. Нам приходилось принимать во внимание, что ЭТАП приписывает словам иной набор морфологических характеристик (существенно более полный), чем тот, который содержался в разметке, поэтому из всех вариантов разбора выбирался тот, у которого имя лексемы совпадало с разметкой, а морфологические характеристики не противоречили ей. Альтернативные варианты разбора в этом случае

стирались, и в результате в окончательную структуру фразы попадал вариант, согласующийся с разметкой. Однако в ряде случаев все альтернативы, построенные ЭТАПом, противоречили разметке корпуса. Для таких предложений существовало 2 решения – отказаться от попытки построить для них дерево или отказаться от требований разметки. В ходе эксперимента мы использовали возможность получить дерево, однако отмечали факт замены лексемы.

После морфологического анализа следовала процедура построения связей между словами. В основном режиме работы в этот момент ЭТАП строит все гипотетически возможные связи с учетом лексико-морфологических вариантов разбора слов. В ходе нашего эксперимента связи, установленные в ходе предварительного анализа фрагментов фраз, имели статус окончательных и не могли быть заменены конкурирующими гипотезами, также они препятствовали проведению связей внутрь сформированных групп и замене их альтернативными гипотезами.

3.5. Построение UNL-структур на основе синтаксических деревьев

В рамках работы мы планируем с помощью лингвистического процессора ЭТАП автоматически построить UNL-графы для тех предложений, для которых удалось получить правильное синтаксическое дерево.

При этом на одном из этапов английские узлы дерева заменяются на узлы UNL. В большом числе случаев одному английскому слову соответствует набор «универсальных слов», причем для каждой конкретной фразы правильный выбор «слова» единственный и зависит от значения исходного слова. Определение значений слов в предложении является ключевым моментом построения графа UNL. Оно существенным образом влияет на интерпретацию валентных отношений и конфигурацию связей в графе. Эту информацию можно считать эквивалентом человеческого суждения при интерактивном разрешении лексической неоднозначности, что является мощным инструментом коррекции автоматического анализа.

Например, в одном из предложений встретилось словосочетание «Committee approval» в значении «одобрение комитетом правительства». В словаре КС системы ЭТАП есть статья COMMITTEE, которой соответствуют два значения этого слова – «правительственная структура» и «неправительственная организация». Для синтаксического анализа различия между этими значениями не важны, но в UNL структуру должно попасть только одно из них. Без семантического анализа фразы сделать правильный выбор между этими вариантами невозможно, но в разметке есть указание на правильное имя статьи UNL – «committee{icl>administrative_unit>thing}».

Таким образом, для построения UNL-графа вновь большое значение приобретает разметка, которая содержит информацию о значениях ряда слов. Однако для использования этих данных было необходимо расширить инструментарий системы ЭТАП. Для этого была разработана новая инструкция, которая заменяет лексему в узле фразы на «универсальное слово», имя которого содержится в разметке. Также было написано правило, которое вызывает эту инструкцию для всех узлов, у которых есть атрибут с указанием значения слова. Для окончания экспериментальной работы необходимо отделить правильные синтаксические деревья от остальных и применить к ним описанный алгоритм.

4. Выводы

4.1. Количественная оценка результатов

Всего в экспериментальном режиме было проанализировано 37701 предложение английского корпуса. В результате было получено 37136 структур фраз, то есть полноценное синтаксическое дерево было построено в 98,5% случаев. В общей сложности было обработано 721240 узлов, то есть средняя длина предложения составляет 19 слов. Проведено 671727 синтаксических связей, из них 30700 (4,57% от общего числа) фиктивных. Среди построенных структур в 25397 (68,39% от общего числа полученных деревьев) не содержатся фиктивные связи, а в 11734 структурах (31,6%) есть по крайней мере одна фиктивная связь.

Для оценки этих результатов было сделано следующее – мы построили синтаксические деревья в штатном режиме работы процессора ЭТАП и затем сравнили результаты для каждой фразы. При анализе 37701 предложения в штатном режиме было получено 36384 структуры (96.5%). Эти структуры содержат 718745 узлов и 658034 связей, из них 24819 фиктивных (3,77% от общего числа связей). Построено 28132 структуры без фиктивных связей (77,32% от общего числа построенных деревьев) и 8252 структур, содержащих фиктивные связи (22,68%). Таким образом, можно заметить, что в целом учет разметки позволил незначительно увеличить

общее количество построенных деревьев, но в то же время за счет предварительно проведенных внутри групп связей задача построения правильной структуры фразы стала сложнее, в результате чего доля деревьев с фиктивными связями возросла с 22, 68% до 31,6%.

Отдельный интерес представляет попарное сравнение структур предложений, полученных в экспериментальном и штатном режимах. Всего различающихся структур 14020, то есть примерно 38% от общего числа построенных. Эти структуры содержат 342025 узлов и 328046 связей, из которых 20649 фиктивные (6,29%). Среди этих предложений у 2452 (17,49% от числа различающихся фраз) разные вершины. В 19885 случаях у узлов отличаются части речи, 19544 узлам были приписаны различные лексемы.. У 55423 узлов произошла смена слова-хозяина. Различается 51841 синтаксическая связь.

4.2. Качественная оценка результатов

После получения полных синтаксических структур необходимо было оценить их, то есть определить число произошедших изменений в структурах и дать им качественную оценку. Ручная проверка всего корпуса, а тем более построение эталонных структур для 37000 предложений слишком трудоемки. Поэтому для оценки были выбраны 2 файла из двух, наиболее и наименее тщательно размеченных, частей корпуса (brown1 и brownmv). Всего в них содержится 181 предложение. Кроме того, из третьей части корпуса (brown2) был выбран контрольный файл из 70 предложений. Два тестовых файла использовались для поиска ошибок и подсчета их числа, а контрольный - для оценки конечного результата исправлений на "девственных" данных. Результаты их оценки можно экстраполировать на весь корпус.

В начале производилось сравнение результатов разбора простого текста без какой-либо разметки с результатом использования корпусных данных. Это дало два набора разных структур для одних и тех же предложений: базовый и улучшенный за счет данных корпуса. Оба неизбежно содержат ошибки синтаксических связей. Однако их сравнение позволяет выделить, классифицировать и оценить отдельные изменения. Впоследствии был выполнен ручной разбор 181 предложения из тестовых файлов, чтобы добавить в сравнение третий - эталонный результат. Все различия между этими наборами подразделяются на 4 типа, а именно: смена вершины синтаксического дерева, замена лексем КС в узлах дерева, изменение синтаксических и морфологических признаков отдельных слов, изменение синтаксических связей (типа связи, хозяина и слуги). Каждое отдельное изменение любого из типов получает собственную оценку: правильно (результат изменения приближает структуру к эталону), неправильно (результат менее похож на эталон) или нейтрально (оба состояния не имеют ноошения к эталону). В дальнейшем основным используемым показателем качественной оценки является отношение улучшений к остальным оценкам в процентах.

Для получения этих цифр используется одна из вспомогательных программ системы ЭТАП tgtproc, которая позволяет автоматически сравнивать наборы структур в формате tgt. и находить различия между ними. Кроме того, для нужд оценки экспериментальных данных было написано три маленьких программы на perl: для выделения изменений структуры заданного типа, сведения изменений в таблицы для удобства оценки и подсчета числа изменений в процентах. С их помощью можно запоминать оценку каждого отдельного изменения внутри предложения и автоматически переносить эту оценку на результат сравнения любого нового разбора тех же данных, если оцененное изменение повторяется. Полученные результаты сведены в Таблицу 2.

Тип изменения	всего	улучшений	ухудшений	нейтральных	без оценки
TOPNODE:	19	12 (63%)	6 (31.5%)	1	0
LINKS:	399	186 (46.6%)	5 (1.2%)	7	201
KSNAMES:	148	125 (84.4%)	18 (12%)	5	0
FEATURES:	174	151 (86.7%)	21 (12%)	2	0

Таблица 2: Достигнутое на момент написания статьи соотношение улучшений структуры к внесенным ошибкам и нейтральным изменениям при сравнении улучшенного набора структур тестовых файлов с базовым.

Кардинальную важность имеют изменения имен лексем (KSNAME), которые в большинстве случаев связаны с изменением части речи и прямо определяют изменения морфосинтаксических признаков соответствующих слов (FEATURES). С ними опосредованно связаны изменения вершин дерева (TOPNODE), которые в свою очередь ока-

зывают огромное влияние на общий рисунок дерева и число изменений отдельных связей.

Тип оставшейся ошибки	всего	% от общего количества слов, связей и вершин
KSNAMES:	87	2,18%
FEATURES:	91	2,28%
TOPNODE:	22	12,10%
LINKS:	781	19,64%

Таблица 3: Оставшееся на момент написания число ошибок улучшенного набора структур 181 тестового предложения в сравнении с эталонным.

Большинство случаев несовпадения лексем между улучшенным и эталонным наборами связаны со служебными словами, которые не имеют семантической разметки в корпусе. Особо частыми были систематические исправления наречия UP на предлог в сочетании "UP TO" и автоматически выбираемого прилагательного THOSE там, где на самом деле имеется форма множественного числа существительного THAT.

Большинство случаев смены вершины произошли в предложениях типа "..." said jury. Они вызваны тем, что в эталонном наборе глагольные группы типа "said jury" оформлены как подчиненные вершине вводные фразы, а в автоматически построенных структурах глагол на месте said сам оказывался вершиной.

4.2.1. Влияние различных данных.

Поскольку корпус содержит разметку различного типа и разной достоверности, программа преобразования исходных данных имеет несколько параметров настройки. Они позволяют включить в получаемые tgt-файлы или игнорировать разные типы данных: морфологическую разметку и отдельно разметку предлогов и наречий (метка IN), выделение зафиксированных в корпусе групп и результаты их автономного разбора, а также выделение сегментов в кавычках и скобках с разбором. Данные были преобразованы с несколькими вариантами настроек, чтобы подобрать оптимальные параметры исходных данных. Также для всех файлов были получены базовые разборы с помощью синтаксического анализатора ЭТАП без какой-либо дополнительной информации.

В результате сравнения разных вариантов исходных данных оказалось, что даже не совсем надежная статистическая частеречная и морфологическая разметка улучшает результат, включая служебные части речи. Исключением стали только систематически неразличаемые в корпусе предлоги и наречия, многие из которых обозначены одной меткой "IN". Использование приемов сегментации с предварительным разбором фрагментов также дает положительный эффект.

Тип изменения	всего	улучшений	ухудшений	нейтральных	без оценки
TOPNODE:	19	15 (78.9%)	3 (15.7%)	1	0
KSNAMES:	148	117 (79%)	26 (17.5%)	5	0
FEATURES:	174	104 (59.7%)	23 (13.2%)	47	0

Таблица 4: Соотношение улучшений структуры к внесенным ошибкам и нейтральным изменениям при сравнении улучшенного набора структур тестовых файлов с базовым с использованием сегментации.

Тип изменения	всего	улучшений	ухудшений	нейтральных	без оценки
TOPNODE:	22	15 (68.1%)	5 (22.7%)	2	0
KSNAMES:	147	112 (76.1%)	29 (19.7%)	6	0
FEATURES:	174	103 (59.1%)	23 (13.2%)	48	0

Таблица 5: Соотношение улучшений структуры к внесенным ошибкам и нейтральным изменениям при сравнении улучшенного набора структур тестовых файлов с базовым без сегментации.

В варианте без сегментации растет число ошибочных и нейтральных замен. По KSNAME разница 2.9%, По TOPNODE разница около 10%. В контрольном файле разница больше. При этом подсчете использовались не все известные из корпуса фрагменты. Из-за ненастроенности правил выбора вершины в именных группах на анализ фрагментов предложения и ошибочного

предпочтения омонимов-глаголов пришлось удалить как вероятнее всего ошибочно истолкованные анализатором все словосочетания, в составе которых были обнаружены глаголы в финитной форме. Типичным примером таких ошибок стало ранее упомянутое "to miss Grant".

4.3 Заключение

Описанное в статье преобразование корпуса в набор синтаксических структур является автоматическим процессом, опирающимся на словари и правила, которые постоянно улучшаются. В частности, на момент написания статьи проводится проверка правильности соответствий концептов UNL и Wordnet, которые определяют выбор лексем КС для слов с семантической разметкой и далее получаемую синтаксическую структуру. Одновременно происходит улучшение собственно синтаксического анализатора и его словарей. Это значит, что число ошибок в получаемом автоматическом разборе будет сокращаться, а общий достигнутый уровень правильности структур может быть повышен еще на несколько процентов.

5. Список литературы

[1] И.А. Мельчук Опыт теории лингвистических моделей класса «Смысл Текст». Москва, Наука, 1974

[2] И.М. Богуславский, Д.Р. Валеев, Л.Л. Иомдин, В.Г. Сизов. Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов // Труды Международной конференции «Корпусная лингвистика – 2008». СПб.: Санкт-Петербургский государственный университет, 2008. ISBN 978-5-288-04769-5. С. 56-74

[3] В.Г. Сизов, Д.Р. Валеев, Л.Г. Крейдлин Сравнение качества работы синтаксического анализатора системы ЭТАП и статистического синтаксического анализатора MaltParser на материале текстов из корпуса СинТагРус // Сборник трудов 31-ой Конференции молодых ученых и специалистов ИППИ РАН «Информационные технологии и системы (ИТиС'08)». Геленджик, 2008 (27 сентября -04 октября), с. 219-224

[4] George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. (1994). Using a Semantic Concordance for Sense Identification. In: Proceedings of ARPA Human Language Technology Workshop

[5] George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. (1993). A Semantic Concordance. In: Proceedings of the 3 DARPA Workshop on Human Language Technology

[6] Brill, Eric (1993). Transformation-Based Error-Driven Parsing. In Proceedings of the Third International Workshop on Parsing Technologies. Tilburg, The Netherlands

[7] Brill, Eric (1995). Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. In Computational Linguistics, vol. 21(4):543–565

[8] www.undl.org

[9] http://ixa.si.ehu.es/Ixa/resources/selprefs

[10] И.М. Богуславский, Н.В. Григорьев, С.А. Григорьева,Л.Л. Иомдин, Л.Г. Крейдлин, Н.Е. Фрид. 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