УДК 004.89:004.93

Т.В. Ермоленко, В. В. Моховых

Донецкий национальный технический университет, г. Донецк
факультет компьютерных наук и технологий,
кафедра программного обеспечения интеллектуальных систем

ФОРМАЛИЗАЦИЯ ПРАВИЛ ВЫДЕЛЕНИЯ ПРЕДИКАТИВНОГО ЯДРА ПРЕДЛОЖЕНИЙ, ИСПОЛЬЗУЕМЫХ СИНТАКСИЧЕСКИМ ПАРСЕРОМ АНГЛИЙСКИХ ТЕКСТОВ

Аннотация

Ермоленко Т.В., Моховых В. В. Формализация правил выделения предикативного ядра предложений, используемых синтаксическим парсером английских текстов*. В статье сделан аналитический обзор существующих подходов к проведению автоматического синтаксического анализа текстов, рассмотрены минимальные структурные схемы простых английских предложений, в соответствии с ними разработаны правила выделения предикативного ядра.*

Ключевые слова: *автоматический синтаксический анализ, минимальные структурные схемы предложений, предикативное ядро.*

Введение. Выявление формальных структур естественного языка (ЕЯ), формализация языка в целом, построение компьютерной модели языка являются приоритетными направлениями информатики на протяжении последних десятилетий. Системы информационного поиска, диалоговые системы, инструментальные средства для машинного перевода и автореферирования, рубрикаторы и модули проверки правописания, так или иначе, проводят анализ ЕЯ-текстов. Таким образом, область применения систем автоматической обработки текстов достаточно разнообразна, а в виду большого роста объемов текстовой информации и сложной ее структурированности, анализ ЕЯ-текстов представляет собой очень актуальную проблему.

Самые большие возможности и высокое качество анализа текстов можно получить, проведя его полный лингвистический анализ. Лингвистический процессор системы, поддерживающей полный анализ ЕЯ-текста, содержит 3 основных компонента: морфологический анализ – построение морфологической интерпретации слов входного текста; синтаксический анализ – построение синтаксической структуры предложения; семантический анализ – построение семантического графа текста.

Построение достоверных синтаксических структур всех подряд предложений текста – очень важная и нужная ступень в автоматическом понимании текста. Описание сущностей входного текста, определение их

свойств и отношений между ними решается уже на уровне синтаксической модели, так как проявляются на уровне общей схемы, не зависящей от смысла высказываний, поэтому морфолого-синтаксические признаки и структуры привлекаются в качестве правил локального контекстного разбора. Таким образом, синтаксический анализ определяет качество работы лингвистического процессора в целом, что делает создание эффективного синтаксического компонента актуальной задачей.

Обзор существующих подходов к автоматическому синтаксическому анализу ЕЯ-текстов. Основная задача синтаксического анализа – используя морфологическую информацию о словоформах, построить синтаксическую структуру входного предложения. К началу работы синтаксического компонента лингвистического процессора весь текст представляется в виде последовательности характеристик к словоформам, т.о. алгоритм синтаксического анализа имеет дело не со словоформами, а с соответствующими характеристиками.

Общим подходом к проведению синтаксического анализа является его разбиение на несколько этапов [1, 2]: сегментация, частичное снятие омонимии, построение синтаксической структуры предложения.

При построении синтаксической структуры предложения выбор модели включает в себя выбор формальной грамматики как системы правил и ограничений, описывающих построение предложения из частей, и алгоритма разбора, который применяет данные правила к входному тексту.

Представления о бинарных синтаксических связях используются в двух известных моделях синтаксической структуры: графах зависимостей и графах непосредственных составляющих. В настоящее время эти две формы представления синтаксической структуры остаются основными, они используются в чистом виде или в смешанных формах, сочетающих в себе свойства обоих графов [3].

Существующие способы представления синтаксических структур имеют определенные недостатки: деревья подчинения не учитывают связей между словосочетаниями и синтаксически целостными группами слов, системы непосредственных составляющих игнорируют направленные связи и не позволяют описывать разрывные словосочетания. Кроме того, в этих представлениях члены предложения определяются на основе формальных признаков, а не по отношению к их семантическому содержанию. Поэтому ни одна из моделей не дает полного представления о синтаксической структуре предложения.

Для повышения качества синтаксического разбора наиболее оптимальным представляется использовать для формирования синтаксических моделей свойство предикативности, одной из важнейших характеристик простого предложения [4]. Предикат – центральная синтаксема в семантическом простом элементарном предложении, формирующая его

семантико-синтаксическую структуру. Предикатная модель наилучшим образом отражает смысл предложения, так как в предикатах указывается не только аргументная структура и количество актантов, но и их семантическое содержание.

К сожалению, на сегодняшний день количество публикаций, посвященных разработке правил распознавания синтаксических конструкций в рамках предикатной модели, крайне малочисленно. Данная статья посвящена вопросам формализации правил, позволяющих выделить предикативный минимум простых английских предложений. На основе этих правил в дальнейшем предполагается разработать синтаксический парсер английских текстов, позволяющий получать синтаксическую структуру предложений в виде предикатной модели и повысить качество дальнейшего семантического анализа.

Цель статьи – на основе анализа структурных схем, описывающих предикативный минимум английских предложений, разработать формальные правила, используемые синтаксическим парсером, для получения предикативного ядра предложений.

Постановка задачи. Правила распознавания предикативного минимума должны разрабатываться на основе информации о словах, полученной на этапе морфологического анализа. При этом каждой словоформе предложения приписывается соответствующий набор (наборы — в случае морфологической омонимии) морфологических характеристик. Таким образом, каждое предложение представимо в виде:

$$S = (s[1],..., s[i],..., s[N]),$$

где $s[i] = \{s[i][1],..., s[i][j],..., s[i][N]\}$ – вектор множеств интерпретаций словоформ, при этом каждое множество интерпретаций $s[i]$ является массивом пар (лемма, морфологические характеристики).

Результатом применения правил к предложениям является структура (*PRED, Subj*), описывающая предикативный минимум предложения, где *PRED* – ядро предиката, глагольная конструкция; *Subj* – грамматический субъект, являющийся левосторонним актантом предиката *PRED*.

Минимальные структурные схемы простых предложений английского языка. Множество простых предложений задается перечнем МСС, описывающих предикативный минимум предложения. В английском языке структура любого простого предложения характеризуется обязательной двучленностью: субъекта и предиката, которые составляют предикативное ядро предложения.

В большинстве предложений английского языка ядром грамматического предиката является глагольная конструкция. К показателям предикативности в английском языке относятся: личный глагол любого времени, залога и наклонения, спрягаемые формы глагола-связки to be, вспомогательные и модальные глаголы.

Итак, в МСС предложений английского языка входят формы слов, которые перечислены в таблице 1.

Таблица 1 – Формы слов, входящих в МСС

Форма слова	Сокращение
1. Признаки предикативности	
Непереходный личный глагол	Vi
Переходный личный глагол	Vt
Спрягаемая форма глагола-связки to be	be
Глаголы-связки, отличные от to be (to seem, to become)	Vb
Глагол действия, выступающий в роли глагола-связки	Vs
Вспомогательный глагол, выступающий в роли смыслового глагола	Vh
2. Имена и наречия	
Именная группа, представленная существительным в общем падеже либо местоимением	NP
Субстантив, выраженный существительным, прилагательным или причастием	subs
Адъективная группа, выраженная прилагательным	Adj
Наречная группа или предложная группа, способная сочетаться со связкой	AdvP

Исследователями английского синтаксиса были составлены списки структурных схем простого предложения, насчитывающие неодинаковое количество членов. Количество структурных схем, выделяемых разными авторами, колеблется от 3 до 39 [5].

Следует отметить, что при автоматическом выделении МСС целесообразно использовать прежде всего формальные критерии, придавая большое значение способам морфологического выражения элементов структурных схем. Так, согласно П. Робертсу [6], в зависимости от морфологической выраженности предикативного члена можно выделить 7 подтипов ядерных предложений английского языка (табл. 2).

Таблица 2 – Минимальные структуры предложений

№ п/п	Шаблон МСС	Пример
1	NP + Vi	John worked.
2	NP + Vt + NP	John paid the bill.
3	NP + be + subs	John is heroic (a hero).
4	NP + be + AdvP	John is in the room.
5	NP + Vb + subs	John became a hero (heroic).
6	NP + Vs + Adj	John felt sad.
7	NP + Vh + NP	John has a car.

Правила, позволяющие автоматически выделить предикатное ядро предложений. Обобщим сведения о подлежащем и сказуемом в английском языке.

Состав. Синтаксическим существительным (подлежащим) может выступать существительное или субстантивное словосочетание, местоимение, глагол в форме инфинитива или глагольная конструкция с инфинитивным ядром, глагол в форме герундия или глагольная конструкция с ядром-герундием, инфинитивное предикативное словосочетание, герундиальное предикативное словосочетание, составная разрывная конструкция, включающая слова there и it.

Сказуемым могут выступать одиночный глагол в простой или аналитической форме, глагол-связка to be с последующим именным членом, модальный глагол с последующей глагольной конструкцией, подчинительное словосочетание, сочинительное словосочетание.

Порядок следования: подлежащее предшествует сказуемому в повествовательных предложениях. Исключения: придаточные предложения условия и сравнения с ограниченным числом глаголов движения и для обособления наречий, особенно негативных (в обратном порядке стоят части составного глагольного сказуемого); усиление значения, смысловое выделение слов.

Правило согласования по числу: подлежащее и сказуемое имеют одинаковую характеристику числа. Исключения: омонимия формы числа существительного (например, The sheep was grazing – The sheep were grazing); омонимия в форме числа глагола во временах, отличных от настоящего, кроме глагола to be (The man spoke – The men spoke); использование в качестве подлежащего слов, означающих группу: audience, board, committee, company, crew, crowd, couple, family, government, group, guard, infantry, pair, Parliament, party, people, platoon и т.д. (The crowd was cheerful – The crowd were pushing); использование в качестве подлежащего местоимений none, all, who (All is well that ends well – All serve as non-adjuncts); использование в качестве подлежащего субстантивного словосочетания с адъюнктом-квантификатором множественности (числительным, кроме one) и ядром-существительным в форме множественного числа (обычно из семантической группы слов-наименований единиц измерения времени, расстояния или денежных сумм) ("Twenty years is a long time," muttered Soames.); использование сочинительного словосочетания в качестве подлежащего (Their happiness and pride was so great – Douglas and I are so happy together.)

Представим описанные правила в виде таблицы 3.

Таблица 3 – Правила выделения связей между главными членами предложения

Часть речи		Число		Время	Лицо		Часть речи допустимых слов-разделителей
PS_1	PS_2	N_1	N_2	T_1	F_1	F_2	
Vi, Vt, Vb, Vs	N	ед.	ед.	наст.			наречие, служебный глагол
		мн.	мн.				
		омонимия		прош./ буд.			
	Место-имение	ед.	ед.	наст.	F_2=3 л.		наречие, служебный глагол
		мн.	мн.				
		омонимия		прош./буд.			
		омонимия		любое	F_2=1, 2 л.		
be	N	ед.	ед.	наст./прош.			наречие
		мн.	мн.	наст./прош.			
		омонимия		буд.			
	Место-имение	ед.		любое	F_2=1 л.		наречие
		мн.		любое	любое		
		ед.		любое.	F_2=3 л.		

Предикатная модель предложения. Элементарное предложение определяется как строго монопредикативное, организующим центром которого является личный глагол, имеющий *n* актантов (не более семи). Актанты выступают в качестве семантических падежей и интерпретируются как «роли» в отношениях действия и состояния, которые выражаются предикатом. Следовательно, валентность глагола диктует обязательные позиции предложения. Для ее определения можно использовать данные синтагматической классификации глаголов современного английского языка, сопоставив каждому классу в соответствие определенный шаблон заполнения валентных гнезд.

Выводы. Рассмотрены минимальные структурные схемы простых английских предложений, использующие способы морфологического выражения их элементов. Это позволило разработать формальные правила для автоматического выделения предикативного ядра.

Предложена синтаксическая модель предложения в виде предикатной структуры, для формирования которой необходимо использовать лингвистические знания в виде словаря валентностей предикатов. Синтаксическая модель в таком виде позволит полностью выявлять как предикативные, так и синтагматические отношения, описывать не только аргументную структуру и количество актантов предиката, но также учитывать их семантическое содержание, используя синтагматическую классификацию глаголов, отражающую их

семантическую нагрузку. Предикатная модель – путь к пониманию текста, которое тесно связано с выявлением предикатных структур, характеризующих смысл предложений, а также – цепочек этих предикатных структур, которые опосредуют смысл текста.

Список литературы

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.

2. Автоматическая Обработка Текста – http://www.aot.ru/ technology.html.

3. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. - М: Наука, - 1985. - 144 с.

4. Ермоленко Т.В. Синтаксическая модель предложения русского языка на основе предикатных структур / Т.В. Ермоленко, А.С. Гайдамака // Искусственный интеллект. — 2012. — № 3. — С. 126–136.

5. С. Е. Кузьмина. Проблема выделения структурных схем простого предложения (на материале английского языка) // Вестник Челябинского государственного университета. — 2009. — № 10. — С. 53–57.

6. Roberts R. Transformation // Практикум по теоретической грамматике английского языка (на английском языке). - М: Наука, - 2010. С. 348-350.