УДК 004.912

ВЫБОР ПАРАМЕТРОВ ДЛЯ ВЫДЕЛЕНИЯ СИНТАКСИЧЕСКИХ ОТНОШЕНИЙ В ПРЕДЛОЖЕНИЯХ РУССКОГО ЯЗЫКА

А. Г. Сбоев*, Р. Б. Рыбка*, И. И. Иванов**

**НИЦ Курчатовский институт, **ФГУП НИИ «Восход»*

Поступила в редакцию 10.03.2014 г.

Аннотация. В статье предлагается оригинальный подход к формированию на основе Национального Корпуса Русского Языка набора параметров, позволяющих снизить неопределенность установления синтаксических отношений между словами в предложении с использованием самоорганизационных и классификационных нейронных сетей. Представлена оценка неопределенности при установлении синтаксических отношений и при построении на их основе деревьев синтаксического разбора.

Ключевые слова: нейронные сети, анализ естественного языка, синтаксический анализ, синтаксические отношения, синтаксическая неоднозначность.

Annotation. An original approach to forming set of parameters based on National Russian Language Corpus that reduce uncertainty of establishing syntactic relations between words in a sentence using self-organization and classification of neural networks was proposed in this article. Estimations of uncertainty in determination of syntactic relations with the parse trees based on them were presented.

Keywords: set of parameters, natural language processing, syntactic parsing, syntactic relations, neural network.

1. ВВЕДЕНИЕ

Перенос многих сторон человеческой активности в виртуальную среду, развитие технологий Big Data, предназначенных для обработки больших объемов текстовой и медийной информации, стимулируют рост интереса к развитию средств анализа масс-медиа, в частности к системам аннотирования текстов, анализа контента бизнес информации, сентимент-анализа, анализа эмотивности текста, выявления угроз в социальных сетях. Качество работы упомянутых средств напрямую зависит от систем анализа естественных языков. Основным затруднением для реализации подобных систем для русского языка является отсутствие качественных синтаксических парсеров, не говоря уже о семантических [1]. Это связано с языковыми особенностями, в первую очередь со слабой упорядоченностью слов в предложении, порождающей большой набор вариантов разбора фразы. Существует два основных подхода к определению синтаксического отношения во фразе: лексический и грамматический. Первый основан на выделении связей между словами, второй – на отношениях, возникающих между наборами морфологических признаков соответствующих слов выражения (главного и зависимого). Недостатком использования лексического метода является чрезмерная разреженность адресуемого пространства слов языка, требующая применения эвристических методов сглаживания для установления отношений [2, 3] и ограничивающая использование универсальных интеллектуальных методов для задачи синтаксического разбора, в частности нейросетевых. Поэтому в данной работе рассматривается второй подход. Результатом синтаксического анализа является дерево разбора, содержащее синтаксические связи. В процессе обработки текстового выражения для построения дерева синтаксического разбора возникают неоднозначности (омонимии) различного типа. Это связано как с многозначностью входных

данных (одна и та же словоформа может быть получена от различных нормальных форм), так и с неоднозначностью синтаксических связей для слов предложения.

Нашей целью являлось выделение комбинации признаков (включая морфологических), снижающей неоднозначность в определении синтаксического отношения между двумя словами в предложении. Выделение такой комбинации необходимо для реализации эффективной процедуры определения синтаксических отношений в рамках проведения синтаксического анализа выражения нейросетевыми и вероятностными методами.

В литературе описаны различные подходы к формированию набора параметров и установления правил разбора на основе языковых корпусов [4, 5, 6]. В частности, в работе [5] используется аппарат рекуррентных нейронных сетей RAAM, в работе [6] описывается метод автоматического извлечения правил для снятия морфологической неоднозначности. Наиболее близкий к данной работе подход излагается в работе [4], где рассматривается экстракция признаков из текстов с использованием классификационных нейронных сетей свертки [7, 8]. Экстракция признаков происходит на основе выделения объектов, описанных признаками низкого уровня. Поэтому необходимыми этапами являлись: составление списка экстрагируемых признаков и формирование обучающих выборок. В случае, когда необходимые признаки экстрагированы, дальнейшая их систематизация может осуществляться разными способами, как с применением нейронных сетей, так и методами минимизации энтропии [9] или скрытых марковских моделей (СММ).

2. ИСПОЛЬЗУЕМЫЕ СРЕДСТВА И МЕТОДЫ

2.1. Национальный корпус русского языка

Для формирования набора значимых признаков необходимыми условиями являются наличие разобранного (тегированного) набора образцовых предложений, а также комплекса программных средств анализа этого набора. В нашей работе мы основывались на “Национальном корпусе русского языка” (НК) [10] – представительном собрании текстов в электронной форме, содержащем информацию о свойствах входящих в него текстов (разметку), что позволяет применять его для научных исследований лексики и грамматики языка. В составе корпуса существует синтаксически размеченный корпус (“СинТаГРус”), который содержит тексты, снабжённые морфосинтаксической разметкой, при этом каждому слову текста сопоставляется одна морфологическая структура, а каждому предложению ставится в соответствие одна синтаксическая структура.

За основу для графематического и морфологического анализа взят свободно распространяемый в исходных кодах пакет библиотек группы АОТ [11], ориентированный на работу с русским языком.

В качестве базового графематического анализатора был выбран модуль “ГРАФАН” АОТ. Некоторые графематические дескрипторы (а именно: признаки заглавной буквы слова и знака препинания, непосредственно следующего за словом) использовались для создания вектора дополнительных признаков слова (см. ниже).

В данной работе рассматривается 4 набора признаков слов выражения (табл. 1), использующихся для определения синтаксических отношений.

Таблица 1

Принадлежность различных параметров наборам

Признаки\ № набора	**1**	**2**	**3**	**4**
морфологические признаки	+	+	+	+
дополнительные		+	+	+
смещение главного слова относительно зависимого			+	+
потенциальные синтаксические отношения между парой слов			+	+
потенциальные синтаксические отношения от слов пары к другим словам выражения				+

Морфологические признаки включают в себя: часть речи, одушевленность, род, число, падеж, степень сравнения, краткость, репрезентацию, наклонение, вид, время, лицо, залог.

К дополнительным признакам отнесены признаки большой буквы и знака после слов в предложении, входящих в состав рассматриваемой пары.

Под смещением главного слова относительно зависимого понимается разница в позициях главного и зависимого слов в предложении.

Под потенциальными синтаксическими отношениями (п_синто) понимаются все синтаксические отношения рассматриваемой пары слов, установленные только на основе морфологических характеристик слов.

Рассматриваемый набор потенциальных синтаксических отношений слов пары к другим словам выражения включает все п_синто между главным словом рассматриваемой пары и другими словами предложения, а также все п_синто между зависимым словом рассматриваемой пары и другими словами предложения.

Для оценки неоднозначности в определении синтаксических отношений разработан ряд методов (см. главы 2.2., 2.3, 2.4, 2.5).

2.2. Метод оценки неоднозначности в определении синтаксического отношения по выделенным наборам признаков

Для проведения исследований по оценке влияния выбранных наборов признаков на определение синтаксических отношений был создан комплекс программных алгоритмов для работы с НК, который включает:

1) процедуру обхода по закодированным текстам корпуса с целью составления относящихся к синтаксическим отношениям НК списков пар слов по различным наборам признаков, описанным в п. 1, 2, 3, 4, таблицы 1;

2) процедуру расчета среднего количества неоднозначных связей для всех слов всех предложений;

3) процедуру расчета среднего количества неоднозначных связей для слов предложения, имеющих неоднозначность;

4) процедуру расчета доли однозначно определяемых синтаксических отношений к общему числу выявленных синтаксических отношений;

5) процедуру расчета вероятности правильного определения конечных элементов в дереве синтаксического разбора.

После выполнения описанных выше процедур расчетов необходимо выбрать набор параметров, для которого:

- Доля однозначно определяемых синтаксических отношений к общему числу выявленных синтаксических отношений максимальна;
- Вероятность правильного определения конечных элементов в дереве синтаксического разбора максимальна.

2.3. Экстракция признаков потенциальных синтаксических отношений на основе морфологических характеристик слов предложения

Для экстракции признаков потенциальных синтаксических отношений применяются нейронные сети типа «Многослойный персептрон» (MLP) [12]. Обучающие множества составляются на основе списков пар слов, составленных для первого набора параметров с использованием процедуры (1) главы 2.2, по следующему алгоритму:

1) Выбираются все пары слов, участвующие в образовании синтаксических отношений в НК.

2) Формируются обучающие выборки для каждого синтаксического отношения в НК. Объекты каждой выборки – это зашифрованные вектора морфологических признаков главного и зависимого слов, образующих синтаксические отношения, которые принадлежат к одному из двух классов: к 1-му относятся те пары, которые принадлежат к рассматриваемому синтаксическому отношению, ко 2-му те, которые не принадлежат.

Далее проводится обучение нейронных сетей для каждого синтаксического отношения и определяется точность их распознавания.

2.4. Установление синтаксических отношений на основе признаков потенциальных синтаксических отношений

Получение набора параметров, выделяющих синтаксические отношения с минимальной неоднозначностью, даёт возможность реализовать на их основе процедуру установления синтаксических отношений и оценить её точность. Для этого формируется обучающее множество для рассматриваемого синтаксического отношения, с использованием списков, полученных в результате выполнения процедуры (1) главы 2.2.

Для установления синтаксических отношений используются подходы:

1) С использованием нейронной сети типа MLP.

2) В этом случае проводится обучение нейронной сети MLP [12] по составленному ранее обучающему множеству объектов и определяется точность распознавания конкретного синтаксического отношения.

3) С использованием комбинации нейронной сети «Растущее Нейронное Дерево» (РНД) [13] и MLP.

РНД относится к классу сетей с динамической архитектурой, так как в процессе конкурентного самоорганизационного обучения нейронное дерево имеет свойство расти в параметрическом пространстве. Алгоритм решает задачу кластеризации. Критерий близости в РНД определяется иерархически структурой дерева: нейроны обучаются на основе конкуренции тогда и только тогда, когда они принадлежат ветвям, имеющим общую вершину. Победивший нейрон находится в соответствии с мерой близости – тот нейрон, который ближе всего находится к текущему примеру, является победившим и может обучаться (передвигаться по направлению к текущему примеру). Целью использования алгоритма РНД является разбиение множества объектов на кластеры, и затем исследование полученных кластеров на типаж классов, содержащихся в них объектов.

В рассматриваемой задаче возникают кластеры двух типов:

1) кластеры, включающие в себя объекты одного класса;

2) кластеры, включающие в себя объекты разных классов.

Для кластеров второго типа определение синтаксического отношения объектов производится с помощью нейронной сети MLP. Альтернативно используются два подхода с выделением всех объектов, не принадлежащих кластерам 1-го типа, в одно (1ый подход) или несколько (2ой подход) обучающих множеств и обучение одной (1ый подход) или нескольких (2ой подход) нейросетевых моделей.

По оценке точности распознавания на основе заранее выбранного набора параметров выбирается способ установления синтаксического отношения.

2.5. Метод оценки процедуры синтаксического разбора на основе списков объектов, составленных с использованием НК, и процедуры нормализации

Получение списков в результате выполнения процедуры (1) главы 2.2. даёт возможность провести оценку неоднозначности системы синтаксического разбора. Методика оценки включает в себя процедуру построения деревьев синтаксического разбора и их нормализации.

После определения синтаксических отношений между словами выражения может возникнуть синтаксическая неоднозначность (см. рис. 1 а).

Для оценки количества возможных синтаксических разборов одной морфологической структуры предложения была реализована соответствующая процедура нормализации полученных разборов. За основу при составлении процедуры взяты свойства синтаксических деревьев, представленных в НК:

- вершина дерева синтаксического разбора только одна,
- для всех слов предложения кроме вершины существует только одна входная связь,
- все слова предложения входят в дерево разбора.

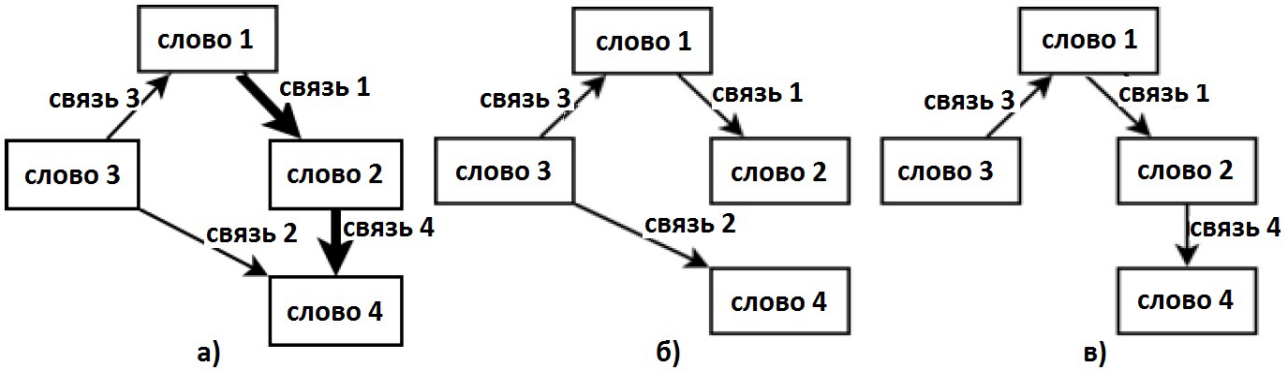

Рис. 1. Примеры деревьев синтаксического разбора без нормализации (а) и после нормализации (б, в).

В результате выполнения процедуры нормализации на основе определенных синтаксических отношений деревья синтаксического разбора выглядят следующим образом (см. рис. 1 б, в).

После определения синтаксических отношений для каждого из предложений НК считается количество однозначных и многозначных вариантов разборов для выбранного набора параметров.

3. ЭКСПЕРИМЕНТЫ

3.1. Выделение набора признаков для определения синтаксических отношений с минимальной неоднозначностью

Для всех предложений с определенной морфологической структурой, содержащихся в СинТагРус, было проведено сравнение наборов признаков для определения синтаксических отношений на основе полученных множеств образцовых примеров и программных алгоритмов, описанных в главе 2.2. (табл. 2)

Набор признаков №4 является наиболее предпочтительным для использования в рамках процедуры установления синтаксических отношений.

3.2. Экстракция признаков потенциальных синтаксических отношений с использованием нейронных сетей

На основе подхода к экстракции признаков потенциальных синтаксических отношений по морфологическим характеристикам слов, описанного в главе 2.2, были созданы и обучены нейронные сети типа MLP для всех типов синтаксических отношений НК. При обучении нейронных сетей использовались наборы примеров морфологических признаков пары слов. Каждый набор состоит из 13 тысяч различных пар, описанных векторами закодированных морфологических характеристик слов (30 значений).

Средняя точность распознавания для экстракции признака потенциального синтаксического отношения равняется 0.4 %.

Таблица 2

Сравнение наборов признаков для определения синтаксических отношений

Набор признаков	Среднее количество неоднозначных связей для слов выражения	Доля однозначно определяемых синтаксических отношений к общему числу выявленных синтаксических отношений	Вероятность правильного определения конечных элементов в дереве
1	102,29	58,48	65,21
2	56,28	78,9	79,04
3	8,84	85,72	90,36
4	1,43	98,91	99,15

Таким образом, можно использовать обученные нейросети для определения потенциальных синтаксических отношений в рамках системы синтаксического разбора предложения в формате НК.

3.3. Установление синтаксических отношений с использованием выбранного набора параметров

На примере предикативного синтаксического отношения было проведено исследование точности установления синтаксических отношений по процедуре описанной в главе 2.3. Обучающее множество включало порядка 582 тысяч объектов, описанных 265 признаками. С применением нейронной сети MLP точность распознавания равнялась 2,5 % для всего обучающего множества.

После выполнения кластеризации с использованием нейронной сети РНД множество разбилось на 1269 кластеров, из которых 514 относятся к кластерам 1-го типа (238 тыс. объектов) (см. главу 2.3). Объекты, относящиеся к кластерам 2-го типа (344 тысячи объектов), были классифицированы с использованием нейронной сети типа MLP:

1) При создании и обучении сети с использованием множества, включающего все объекты, относящиеся к кластерам 2-го типа, средняя ошибка расчета или мера неоднозначности равнялась 2,3 %. Общая средняя ошибка расчета для всего множества объектов равняется 1,35 %.

2) При создании 11 обучающих множеств, включающих примерно по 30 тысяч объектов, и обучении 11 нейронных сетей средняя ошибка расчета равнялась 0,96 %. Общая средняя ошибка расчета для всего множества объектов равняется 0,56 %.

3.4. Оценка процедуры синтаксического разбора на основе составленных с использованием НК списков объектов и процедуры нормализации

С использованием метода, описанного в главе 2.4, была проведена оценка неоднозначности при построении деревьев синтаксического разбора на основе 4-го набора параметров, отобранного по результатам исследования главы 3.1. Результат представлен в табл. 3.

4. ЗАКЛЮЧЕНИЕ

Предложен набор признаков для пар слов анализируемого выражения естественного языка, включающий морфологические признаки рассматриваемой пары слов, их дополнительные признаки, потенциальные синтаксические отношения, полученные на основе набора морфологических признаков рассматриваемой пары, а также потенциальные исходящие синтаксические отношения от главного и зависимого слов к другим словам выражения.

В рамках исследования были созданы методики:

- выбора набора параметров для установления синтаксических отношений с минимальной неоднозначностью,
- экстракции признака потенциальных синтаксических отношений на базе морфологических характеристик двух слов,
- установления синтаксического отношения с применением нейронных сетей типа многослойный персептрон и растущее нейронное дерево,
- оценки неоднозначности при построении деревьев синтаксического разбора с использованием выбранного набора параметров.

Таблица 3

Результат оценки работы процедуры синтаксического разбора с использованием выбранного набора параметров

Количество предложений в НК	Процент однозначно разобранных предложений	Процент неоднозначно разобранных предложений	Среднее количество деревьев синтаксического разбора для неоднозначно разобранных предложений
42977	79,9	20,1	21.3

В результате исследования показано, что при использовании комбинации нейронных сетей РНД и MLP средняя ошибка расчета при установлении синтаксических отношений составляет 0,56 %.

СПИСОК ЛИТЕРАТУРЫ

1. *Ермаков С.А., Ермакова Л.М.* Методы оценки эмоциональной окраски текста, Вестник Пермского университета. – Вып. 1(9). – 2012.

2. *Jan Hajič, Pavel Krbec, Pavel Květoň, Karel Oliva.* Serial combination of rules and statistics: a case study in Czech tagging, 2001. P. 268–275.

3. *Сокирко А.В., Толдова С.Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп), http://www.aot.ru/docs/RusCorporaHMM.htm, 2005.

4. *Ronan Collobert, Jason Weston, L´eon Bottou, Michael Karlen, Koray Kavukcuoglu and Pavel Kuks.* Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research 12, 2011. – P. 2493–2537.

5. *Wong Chun Kit.* Recursive Auto-Associative Memory as Connectionist Language Processing Model-Training Improvements via Hybrid Neural-Genetic Schemata, City University of Hong Kong, 2004.

6. *Протопопова Е., Бочаров В.* Автоматическое извлечение правил для снятия морфологической неоднозначности. Сборник трудов. Национальный Открытый Университет «ИНТУИТ», 2012.

7. *Yann LeCun, Koray Kavukcuoglu and Clement Farabet.* Convolutional Networks and Applications in Vision. Computer Science Department, Courant Institute of Mathematical Sciences, New York University, 2010.

8. *Саймон Хайкин.* Нейронные сети. Полный курс. – 2006. – 330 с.

9. *Randford M.Neal, Geofrey E. Hinton.* A view of the EM algorithm that justifies incremental, sparse, and other variants, 1999.

10. Национальный корпус русского языка – http://ruscorpora.ru/

11. Группа автоматической обработки текста (АОТ) – http://aot.ru/

12. *Осовский С.* Нейронные сети для обработки информации. Перевод с польского И.Д. Рудинского. – М. : Финансы и статистика 2002. – С. 50–55.

13. *Сбоев А.Г., Кукин К.А, Рыбка Р.Б., Сбоев А.А.* «Алгоритм распараллеливания нейронной сети на основе растущего нейронного дерева», 11-ая научно-практическая конференция: «Современные информационные технологии в управлении и образовании». Сборник научных трудов, 2012.

Сбоев Александр Георгиевич – ведущий научный сотрудник, НИЦ Курчатовский институт, к.ф.-м.н. Тел.: 8-926-253-72-17.
E-mail: sag111@mail.ru

Рыбка Роман Борисович – НИЦ Курчатовский институт инженер-исследователь, аспирант. Тел.: 8-926-344-61-35.
E-mail: rybkarb@gmail.com

Иванов Игорь Игоревич – инженер-программист, ФГУП НИИ «Восход», магистрант. Тел.: 8-905-754-96-51. E-mail: honala@yandex.ru

Sboev Alexandr Georgievich – National Research Center «Kurchatov Institute», leading Researcher. Phone: 8-926-253-72-17.
E-mail: sag111@mail.ru

Rybka Roman Borisovich – National Research Center «Kurchatov Institute», research engineer, PhD Student. Phone: 8-926-344-61-35.
E-mail: rybkarb@gmail.com

Ivanov Igor Igorevich – engineer, R&D Institute "Voskhod". Phone: 8-905-754-9651.
E-mail: honala@yandex.ru