УДК 004.912

ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ ПРОЦЕДУРЫ СИНТАКСИЧЕСКОГО РАЗБОРА С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Р. Б. Рыбка, А. Г. Сбоев

Национальный исследовательский центр «Курчатовский институт»

И. И. Иванов

ФГУП НИИ «Восход»

Поступила в редакцию 14 апреля 2015 г.

Аннотация: *в статье предлагается модель процедуры синтаксического разбора предложения в формате Национального корпуса русского языка, основанная на нейросетевых алгоритмах. Представляются результаты сравнительного анализа точностей моделей нейронных сетей с различными топологиями для установления синтаксических отношений. Приведены результаты сравнения с другими системами разработанной модели процедуры синтаксического разбора, включающей комплекс отобранных нейросетевых алгоритмов в сочетании с экстрагированными на основе морфологических характеристик признаками потенциальных синтаксических отношений и инкрементальную схему разбора.*

Ключевые слова: *нейронные сети, анализ естественного языка, синтаксический анализ.*

Abstract: *in this paper we propose a model of procedure for syntactic parsing of a separate Russian language sentence. The model is based on the combination of the neural networks. Its purpose is to build a parse tree in the Russian National Corpus (RNC) format. RNC texts contain unambiguous morphological and syntactic markup and are used by us for training neural network models as a part of the procedure. We present the results of a comparison of different neural network algorithms that determine syntactic relations on the basis of methods of multiclass and binary classification. Estimates of accuracy of the developed model of procedure in comparison with the existing systems are also presented.*

Key words: *neural networks, natural language processing, syntactic analysis.*

В современных условиях рост интенсивности информационного обмена приводит к потребности создания автоматизированных систем обработки текстовой информации для аннотирования текстов, анализа контента бизнес-информации, сентимент-анализа, анализа эмотивности текста, выявления угроз в социальных сетях. Ключевым аспектом качества таких систем является способ установления отношений между словами в рамках отдельного предложения. С появлением развитых языковых корпусов появилась база для установления указанных выше отношений статистическими методами и методами искусственного интеллекта, преимуществом которых является меньшая трудоемкость в реализации (к примеру, система ЭТАП-3 [1] разрабатывается более 20 лет) и в чистом виде непереносимость на другие языки.

Сказанное выше актуализирует задачу разработки алгоритма установления синтаксических отношений между словами и формирования дерева синтаксического разбора предложения на основе данных из Национальных корпусов. Эффективное решение этой задачи методами корпусной лингвистики вызывает необходимость комплексного исследования вопросов: с одной стороны, достижимой точности, которую могут обеспечить языковые корпуса, а с другой – методов достижения этой точности.

Имеется ряд методов, которые используются для установления синтаксических отношений, в частности методы вероятностных грамматик (PCFG, LinkGrammar), методы искусственного интеллекта (SVM, SRN и RAAM). При этом существуют разные способы использования информации корпуса: в одном случае используются только грамматические признаки отдельных слов с добавлением признаков, характеризующих особенности написания слов, наличие разделителей, их мест в предложении и т.д., во втором случае используются также признаки индивидуальной словоформы из словаря корпуса (лексические).

В каждом случае применение той или иной комбинации методов и корпусной информации имеет свои недостатки, в частности: методы формальных

грамматик применяются в основном для случая языков с проективными связями.

Подход на основе нейросетевых моделей предпочтителен ввиду того, что нейронные сети обладают известными обобщающими свойствами. Использование этих свойств в комбинации с признаковым описанием слов и современными методами сжатия данных дает возможность сокращения размерности адресного пространства признаков задачи и построения методики в максимальной степени универсальной для различных языков.

Нашей целью являлось создание набора взаимосвязанных математических методов для численного моделирования процедуры синтаксического разбора на основе формата Национального корпуса русского языка (далее – НК) с обоснованием достижимой точности, которую могут обеспечить корпусные данные.

В литературе описаны различные подходы к формированию набора параметров и установления правил разбора на основе языковых корпусов: использование аппарата рекуррентных нейронных сетей RAAM [2], описание метода автоматического извлечения правил для снятия морфологической неоднозначности [3], метода экстракции признаков из текстов с использованием классификационных нейронных сетей свертки [4; 5].

В исследовании, проведенном нами ранее [6], обоснован набор параметров, включающий морфологические признаки, признаки больших букв слов, знаков после слов, удаленности между словами и признаков синтаксических отношений, установленных на основе морфологических характеристик (п_синто), который позволяет существенно снизить неоднозначность при установлении синтаксических отношений в предложении. Для экстракции признаков п_синто были сформированы модели на основе нейронных сетей PNN, SVM и MLP.

В данной работе представлены результаты исследования методов построения дерева синтаксического разбора с использованием выбранного ранее набора признаков.

1. Используемые средства и методы

1.1. Определение синтаксических отношений

Для определения синтаксических отношений исследуются методы на основе MLP, SGD [7], SVM со стратегией one-vs-all [8], РНД [9], ансамблей деревьев решений (RFC) [10] в комбинации с методами снижения размерности входного пространства (Nystroem) [11]. Под SGD понимается метод подбора параметров функции (1) с использованием (2) :

$$f(x) = w^T x + b \ , \qquad (1)$$

где x – входной объект; w – параметры модели; b – коэффициент.

$$E(w,b) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \alpha R(w), \qquad (2)$$

где y – желаемый выходной класс; L – функция потерь; R – параметр регуляризации (мера l2 или l1) ; α – положительный коэффициент (гиперпараметр).

В качестве L используются различные функции от двух аргументов y и $p = f(x)$:

1) $L(p,y) = \max(0,1-py) = \begin{cases} 1-py, py \le 1 \\ 0, py > 1 \end{cases}$,

2) $L(p,y) = \ln(1+\exp[-py])$,

3) $L(p,y) = \max(0,1-py)^2 = \begin{cases} 0,1 \le py \\ (1-py)^2, -1 \le py \le 1 \\ -4py, -1 \ge py \end{cases}$,

4) $L(p,y) = |y-p|$.

Для классификации синто и действий применяются следующие варианты решений:

1) создание последовательности классификационных нейросетевых моделей для определения синто или действий независимо (бинарная классификация). Последовательность формируется на основе количества примеров для каждого класса: от большего к меньшему;

2) группировка по нескольким синто с учетом количества примеров для них и точности их независимой классификации;

3) создание единой модели решающей задачу мультиклассовой классификации для всех действий и синто.

В 1-м и 2-м случаях каждый следующий классификатор обучается на объектах, свободных от примеров для тех действий/синто, которые были обучены ранее.

1.2. Подходы к построению дерева синтаксического разбора и формирования обучающих примеров для определения синтаксических отношений

Исследовались два подхода к построению синтаксического дерева разбора: на основе перебора всевозможных вариантов установления синто между словами в предложении и на основе инкрементальной схемы разбора Ковингтона [12], который применялся в других исследованиях применительно к текстам русского языка [13]. Первый из подходов основан на переборе всевозможных комбинаций слов предложения и определении синтаксических отношений в них. Набор примеров для обучения в этом случае (обучающая выборка) состоит из пар слов по всем предложениям НК, разбитыми на два класса:

1) образующие синто;

2) необразующие синто.

При этом количество примеров второго класса много больше первого.

Суть второго подхода заключается в построении модели переходов по трем спискам слов: b – правый список, состоящий из всех неразобранных слов, l – левый список, в который заносятся слова для нахождения связи между b[0] и l[0], и m – промежуточный список, который заполняется словами из списка l при условии не нахождения связи между b[0] и l[0]. При этом используются 4 класса действий: No-Arc – перенести l[0] в вершину m, Shift – перенести b[0] и все слова из m в l так, чтобы слово b[0] стало вершиной l[0], Right-arc – провести связи между b[0] и l[0], Left-arc – провести связи между l[0] и b[0]. При выполнении действий Right-arc и Left-arc слово l[0] переходит в m и становится его вершиной. Обучающая выборка при этом состоит из объектов, соответствующих вышеуказанным действиям схемы разбора Ковингтона с учетом типа синто.

1.3. Модель процедуры синтаксического разбора

В соответствии с выбранным подходом к формированию дерева синтаксического разбора строится модель процедуры синтаксического разбора, включающая следующие этапы:

1) установление потенциальных синтаксических отношений;

2) формирование дерева синтаксического разбора.

С использованием разработанной модели процедуры вычисляются точности:

1) определения синтаксической связи между словами без типа синто;

2) определения типа синто между словами;

3) определения корня (вершины) дерева синтаксического разбора;

4) построения структуры дерева разбора, т.е. с указанием связности слов в предложении без указания типов синто;

5) построения полной синтаксической структуры предложения.

2. Эксперименты

2.1. Оценка точности определения синтаксических отношений и выбора метода построения дерева разбора

2.1.1. Перебор всевозможных комбинаций слов предложения и определение синтаксических отношений в них

Для составления обучающей выборки нами использовались первые 500 текстов в составе НК, остальные 16 использовались для тестовой выборки. При разбиении объектов, образующих синтаксические отношения (650 тыс. примеров в обучающем множестве) от необразующих (10 млн примеров в обучающем множестве) рассматривались различные нейросетевые методы, из которых лучшие результаты показали: MLP (2 слоя: 22 и 22 нейрона) с точностью 83.45/88.3 % и MLP (1 слой: 50 нейронов) с точностью 82.1/89.12 %.

Для определения конкретного синтаксического отношения в рамках множества объектов (650 тыс. объектов в выборке), имеющих синто, были отобраны методы на основе:

• построения отдельных нейросетевых моделей, решающих задачу бинарной классификации для каждого синтаксического отношения;

• формирования групп синтаксических отношений на основе точностей их отдельного определения.

При построении моделей для бинарной классификации лучшие результаты продемонстрировали модели на основе MLP (2 слоя) и SVM с использованием методов обучения на основе SGD (табл. 1).

Таблица 1

Результаты определения синтаксических отношений с количеством образцовых примеров больше 200

Тип сети	MLP (2 слоя)	SGD (1)	SGD (2)	SGD (3)	SGD (4)
Средняя точность классификации (синто в % / не синто в %)	92\92.5	96\85	90\91	95\91	92\94

Для синтаксических отношений с малым количеством примеров (меньше 200) разработан специальный гибридный метод, объединяющий методы кластеризации и классификации PHД и PNN, суть которого заключается:

1) в предварительном разделении объектов на кластеры с выделением трех категорий кластеров: включающие объекты только 1-го класса, 2-го класса или обоих классов;

2) для объектов, относящихся к кластерам 2-й категории, проводится замена объекта на вектор центра масс данного кластера, рассчитанного при обучении;

3) далее преобразованное множество обрабатывается средствами сетей PNN.

Средняя точность определения синтаксических отношений с малым числом примеров при использовании гибридного метода составляет 99 %. Таким образом, общая точность определения синтаксических отношений в рамках выделенного множества синтаксически-значимых вариантов при использовании гибридного метода составляет 95.41/94.05 %.

При реализации метода формирования групп синтаксических отношений в результате исследования было создано 6 групп синто, а средняя точность определения синтаксических отношений при исполь-

зовании метода группировки синто и их классификации на базе сетей SGD и РНД в комбинации с PNN составляет 96.56/97.4 %.

Таким образом, общая точность установления синто после выделения синтаксически-значимого множества составляет 79.89 %.

2.1.2. Подход на основе инкрементальной схемы разбора

В данном подходе для определения действий и синтаксических отношений исследовались варианты решений на базе:

1) построения единой модели, выполняющей мульти-классовую классификацию;

2) решения на основе бинарной классификации действий, когда обучение каждого следующего классификатора проводилось на выборке, из которой исключены классы действий уже построенных ранее классификаторов.

Лучшие результаты показал алгоритм SVM с линейным ядром (табл. 2).

Из этого следует, что для реализации интеллектуального разборщика выбран подход на основе инкрементальной схемы разбора с классификатором на базе SVM с линейным ядром.

2.2. Модель процедуры синтаксического разбора

В соответствии с выбранным подходом к формированию дерева синтаксического разбора была реализована модель процедуры синтаксического разбора, в которой для установления признаков потенциальных синто используются сети SVM, MLP и PNN, а для определения действия и типа синтаксических отношения – модель, решающую задачу мультиклассовой классификации на базе SVM с линейным ядром (рисунок).

Т а б л и ц а 2

Определение действий в инкрементальной схеме разбора

Вариант решения	Алгоритм	Общая точность (%)
Мультиклассовый	SVM (Линейное ядро)	90/91
Мультиклассовый	Nystrom(100) + RFC	83/84
Мультиклассовый	Кодирование с использованием случайной бинарной нормировки + RFC	84/ 84
Бинарная классификация	SGD (различные функции сходимости)	87/88
Бинарная классификация	Комбинация: Nystrom(100) + RFC или SVC (1–4)	89/90

По результатам апробации разработанной модели процедуры синтаксического разбора на тестовых предложениях НК полученные точности построения структуры дерева разбора и построения полной синтаксической структуры предложения (табл. 3) превышают опубликованные в литературе данные по этим величинам для других систем.

Как показано в табл. 3, добавление словоформ к выбранной признаковой модели дает повышение точности определения синто на 10,03 % и формирования дерева синтаксического разбора на 20,86 %.

Заключение

Таким образом:

• на основе выбранного ранее набора параметров для установления синтаксических отношений разра-

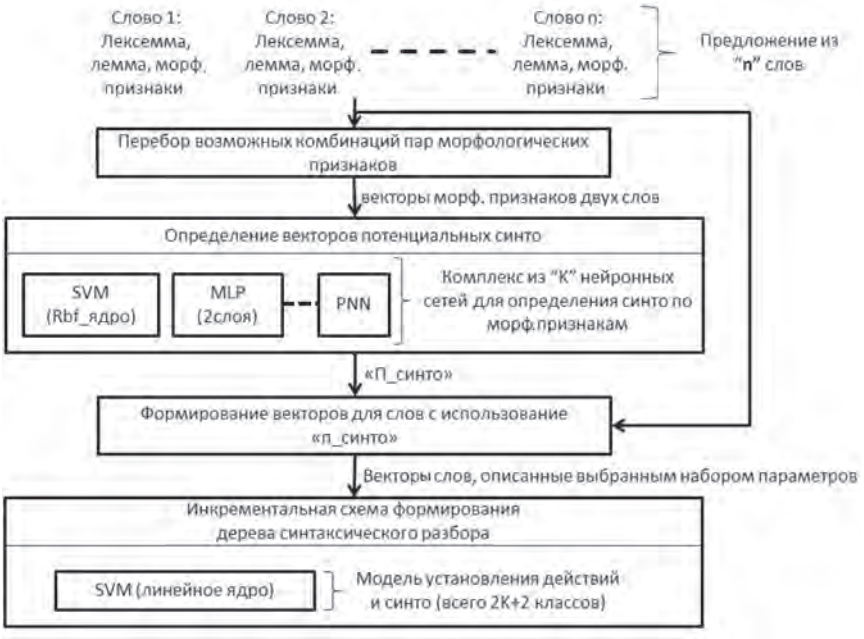

Рисунок. Схема алгоритма формирования дерева синтаксического разбора в соответствии с выбранными признаками и нейросетевыми моделями: К – количество синто в текстах НК

Таблица 3

Точности тестирования разработанной модели процедуры на предложениях НК в сравнении с литературными данными

Описание задачи	Точности определения синтаксической связи между словами без типа синто	Точности в определении корня дерева синтаксического разбора	Точности определения синтаксической связи между словами без типа синто	Точности построения полной синтаксической структуры предложения	Точности в построении структуры дерева
SVM + набор параметров без словоформ	85,81	82,23	79,33	14,05	29,47
SVM + набор параметров с добавлением словоформ	91,73	88,84	89,39	35,91	52,38
Литературные данные					
Лингвистический процессор ЭТАП-3 (ИППИ РАН) [14]	94,3		92,3	29,7	37,4
Инкрементальная схема разбора: признаковая модель дополненная данными из системы ЭТАП-3	93,5		88,6	26,1	35,2

ботана модель процедуры синтаксического разбора в формате Национального корпуса русского языка;

• апробирование модели на предложениях НК продемонстрировало точность формирования дерева разбора 35,91 %, его структуры 52,38 % и установления синтаксических отношений 89,3 %.

В перспективе планируется построение на основе разработанной модели практических процедур, пригодных для решения задач: определения эмотивности, тональности и отбора тематически схожих документов.

ЛИТЕРАТУРА

1. *Iomdin L.* ETAP Parser : State of the Art / L. Iomdin, V. Petrochenkov, V. Sizov, L. Tsinman // Компьютерная лингвистика и интеллектуальные технологии : материлы ежегод. Междунар. конф. «Диалог», 2012. – Т. 2, № 11 (18). – С. 119–131.

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

3. *Протопопова Е.* Автоматическое извлечение правил для снятия морфологической неоднозначности / Е. Протопопова, В. Бочаров ; Национальный Открытый Университет «ИНТУИТ». – 2012.

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

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

6. *Рыбка Р. Б.* Выбор параметров для выделения синтаксических отношений в предложениях русского языка / Р. Б. Рыбка, А. Г. Сбоев, И. И. Иванов // Вестник Воронеж гос. ун-та. – 2014. – № 2. – С. 117–124.

7. *Tong Zhang.* «ICML 2004» // Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms, 2004.

8. *Anderson Rocha, Siome Goldenstein.* Multiclass from Binary : Expanding One-vs-All, One-vs-One and ECOC-based Approaches // IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, 2013.

9. *Сбоев А. Г.* Алгоритм распараллеливания нейронной сети на основе растущего нейронного дерева / А. Г. Сбоев [и др.] // 11-я НПК : Современные информационные технологии в управлении и образовании. – М., 2012.

10. *Breiman L.* Random forest. – University of California, Berkeley, 2001.

11. *Sanjiv Kumar, Mehryar Mohri.* Ensemble Nystrom Method. – Courant Institute of Mathematical Sciences, New York, 2009.

12. *Joakim Nivre*. Incrementality in Deterministic Dependency Parsing // School of Mathematics and Systems Engineering, Vaxjo, Sweden, 2004.

13. *Sharov S.* The proper place of men and machines in language technology / S. Sharov, J. Nivre // Компьютерная лингвистика и интеллектуальные технологии : материалы ежегод. Междунар. конф. «Диалог». – 2011. – С. 657–670.

14. *Казенников А. О.* Сравнительный анализ статистических алгоритмов синтаксического анализа на основе деревьев зависимостей / А. О. Казенников // Компьютерная лингвистика и интеллектуальные технологии : материалы ежегод. Междунар. конф. «Диалог». – 2011. – № 9 (16). – С. 157–163.

Национальный исследовательский центр «Курчатовский институт»

Рыбка Р. Б., инженер-исследователь
E-mail: rybkarb@gmail.com
Тел.: 8-926-344-61-35

Сбоев А. Г., ведущий научный сотрудник, кандидат физико-математических наук
E-mail: sag111@mail.ru
Тел.: 8-926-253-72-17

ФГУП НИИ «Восход»

Иванов И. И., инженер-программист, магистрант
E-mail: honala@yandex.ru
Тел.: 8-905-754-96-51

National Research Center «Kurchatov Institute»

Rybka R. B., Research Engineer
E-mail: rybkarb@gmail.com
Tel.: 8-926-344-61-35

Sboev A. G., Candidate of Physical and Mathematical, Leading Researcher
E-mail: sag111@mail.ru
Tel.: 8-926-253-72-17

R&DI «Voskhod»
Ivanov I. I., Softwave Engineer, Undergraduate Student
E-mail: honala@yandex.ru
Tel.: 8-905-754-96-51