УДК 004.048

СЕМАНТИКО-СИНТАКСИЧЕСКИЙ АНАЛИЗ ПРЕДЛОЖЕНИЙ НА РУССКОМ ЯЗЫКЕ С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

Мельцов В.Ю., Нечаев А.А.

ФГБОУ ВО «Вятский государственный университет», Киров, e-mail: meltsov69@mail.ru

В настоящее время многие программные системы работают с данными, представленными в виде предложений и текстов на естественном языке. Автоматический семантико-синтаксический анализ предложений на русском языке крайне трудно осуществить из-за сложного, неоднозначного и свободного синтаксиса языка. Все существующие реализации обладают одним общим недостатком: качество разбора сильно зависимо от полноты и правильности базы онтологий. Задача создания полной базы онтологий для русского языка практически невыполнима. Было выдвинуто предположение, что оценки численных показателей, рассчитываемых на основе расположения слов в предложении и не опирающихся на выделенные вручную онтологии, достаточно для построения семантико-синтаксического дерева. Для слов предварительно должны быть помечены возможные морфологические характеристики и возможные значения. Оценка производится с помощью алгоритмов машинного обучения. Для проверки данной гипотезы создан прототип, определяющий принадлежность группы слов к некоторому узлу семантико-синтаксического дерева. Проведены эксперименты по частичному разбору первых двух глав романа М.А. Булгакова «Мастер и Маргарита». Полученные результаты позволяют оценить применимость описанного алгоритма при решении реальных задач, связанных с обработкой данных, представленных в виде предложений и текстов на русском языке.

Ключевые слова: естественный язык, семантико-синтаксический анализ, онтология, нейронные сети, модульная структура

SEMANTIC-AND-SYNTACTIC PARSING OF SENTENCES IN RUSSIAN USING NEURAL NEWTROKS

Meltsov V.Yu., Nechaev A.A.

Vyatka State University, Kirov, e-mail: meltsov69@mail.ru

Currently, there are many software systems using input data presented as texts and sentences in natural language. The complex, ambiguous and fluent syntax of Russian language causes very difficult implementation of automatic semantic-and-syntactic parser for sentences in Russian. All existing solutions have a common drawback: the output quality strongly depends on the completeness of an ontology base. However, it is impossible to create a complete ontology base for Russian language. It has been suggested that the evaluation of numeric characteristics, calculated from word positions in a sentence without using any human-developed ontology base is sufficient to build a semantic-and-syntax tree. Preliminary, words must be marked with possible meanings and morphological characteristics. Evaluation is performed with machine-learning algorithms. A prototype was developed to determine the accordance of a word group to a node of a semantic-and-syntax tree. First two chapters of M.A. Bulgakov's novel «Master and Margarita» have been particularly parsed using the prototype. The obtained results allow evaluate applicability of the described algorithm to solve real issues concerned with processing data presented as texts and sentences in Russian.

Keywords: natural language, semantic-and-syntactic parsing, ontology, neural networks, modular structure

В настоящее время многие программные системы работают с входными данными, представленными в виде текстов и предложений на естественном языке (ЕЯ) [1]. В большинстве случаев это обусловлено желанием сделать интерфейс взаимодействия с компьютером более удобным и понятным человеку. А некоторые предметные области однозначно предполагают представление обрабатываемых данных на ЕЯ: системы электронного документооборота, системы анализа тональности отзывов и сообщений в социальных сетях, машинный перевод и пр. [2].

Во многом возможности разбора и понимания естественного языка компьютером ограничиваются из-за применяемых алгоритмов разбора, основанных на использовании баз знаний для конкретной предметной области – онтологий [3]. Из-за сложности синтаксиса русского языка, порождающей огромное многообразие возможных словесных конструкций, не представляется возможным построить исчерпывающее множество онтологий. Это ведет к ошибкам, неточностям и неполноте разбора. Для выполнения автоматической обработки текстов на ЕЯ вычислительная машина должна работать с некоторым формализованным представлением его смысла и структуры – семантико-синтаксической моделью. Как правило, такой моделью является граф – либо это граф зависимостей, либо граф непосредственных составляющих, либо структура, совмещающая свойства обоих описаний [4, 5].

На текущий момент существует не такое большое количество решений постро-

ения подобных моделей для предложений на русском языке [6]. Рассмотрим наиболее известные из них.

Проект Автоматическая Обработка Текста (*АОТ)* – проект, начатый в 1998 г. как система автоматического перевода ДИАЛИНГ [7]. По результатам проведённых исследований было написано две диссертации и ряд научных статей. По мнению разработчиков проекта, наилучшими для разбора естественного языка являются консервативные методы, основанные на описании логики всех языковых механизмов, используемых человеком.

Проект ABBYY Compreno – технология анализа и разбора текстов на ЕЯ, способная проводить полный семантико-синтаксический анализ текста [8]. По результатам проведённых исследований было опубликовано более 20 научных статей. Основываясь на полученных в проекте результатах, широте применения выпущенных приложений и компонентов для решения реальных задач, а также научной ценности публикаций, можно считать на текущий момент ABBYY Compreno наиболее значимым достижением в области обработки ЕЯ (особенно, что касается текстов на русском языке). Наиболее серьезные недостатки, которые сложно устранить сторонним пользователям, использующим технологии на базе ABBYY Compreno, – очень высокое потребление вычислительных ресурсов, сильная зависимость качества разбора от качества базы онтологий и нестабильность выделения идиоматических выражений в текстах.

Отдельный интерес представляет работа [9] на тему «Исследование методов автоматического анализа текстов и разработка интегрированной системы семантико-синтаксического анализа», выполненная в Институте системного анализа РАН. Работа датируется 2015 годом и содержит большой объём аналитической информации о текущем состоянии и развитии теории алгоритмов для семантико-синтаксического анализа естественного языка на ЭВМ. В ней убедительно подтверждается, что основным недостатком всех известных масштабных проектов по анализу текстов является крайняя зависимость качества обработки от используемого набора онтологий (шаблонов, правил). Если набор неполон, то разбор ЕЯ завершается неверно, связи в семантико-синтаксическом дереве отсутствуют или создаются неправильно.

Для решения данной проблемы предлагается применять исключительно средства машинного обучения, не использующие для выполнения семантико-синтаксического разбора базу онтологий, которая зачастую составляется вручную и не проверяется на коллизии и другие ошибки и неточности. Можно предположить, что таких простых методик и численных характеристик, полученных на основе анализа расположения слов в предложении, будет достаточно для того, чтобы правильно распознавать принадлежность группы слов к одному узлу семантико-синтаксического дерева.

В качестве главного элемента для анализа предложения (слов в предложении) выбрано отношение «свойство», являющееся узлом будущего дерева разбора. Под отношением «свойство» понимается такая связь двух слов А и Б, при которой А является описанием одного из свойств объекта Б. Например, в предложении «Зелёные вагоны неистово мчались сквозь хвойный лес» есть два таких отношения – «Зелёные вагоны» и «Хвойный лес».

Для проведения экспериментов по анализу различных естественно-языковых текстов было разработано программное приложение, выполняющее поиск словосочетаний, между которыми установлено отношение «свойство». Несмотря на то, что семантико-синтаксическое дерево, описывающее предложение на русском языке, может содержать в себе узлы, относящиеся ко множеству различных типов отношений, помимо «свойство», эксперименты показали, что простая структура данного отношения позволяет значительно сократить временные и ресурсные затраты при разборе текстов. Кроме того, для такого отношения несложно найти противоречащие друг другу примеры, рассмотрение которых крайне важно на этапе семантической обработки. Результаты экспериментов подтвердили, что несмотря на необходимость учитывать и другие типы отношений между словами, определяющим является отношение «свойство».

В качестве алгоритмической основы разрабатываемого программного прототипа подсистемы семантико-синтаксического анализа предложений на русском языке был выбран аппарат искусственных нейронных сетей (НС) [10]. Тогда поставленная задача может быть рассмотрена как задача классификации точек в n-мерном пространстве, при которой необходимо обучение НС с учителем. На сегодняшний день существует множество различных типов нейронных сетей и моделей нейронов. С учётом специфики предметной области и анализа достоинств и недостатков известных НС, для дальнейших исследований были выбраны следующие типы – многослойный персептрон и LSTM-сеть. В качестве входных данных сети исполь-

зуется набор параметров, полученный для соответствующей пары исследуемых слов в предложении и представленный в виде вектора действительных чисел. Выходные данные – числовой вектор «01» или «10», являющееся ответом на вопрос: «Найдено ли искомое отношение в тексте?» («01» – не найдено, «10» – найдено).

Для любых двух слов в предложении эвристически был определён следующий набор числовых параметров, которые можно получить, не используя базы онтологий:

- расстояние между словами;
- количество знаков препинания между словами;
- длина групп слов, которые получаются путем разбиения промежутка между двумя исследуемыми словами знаками препинания.

При этом задачу, решаемую нейронной сетью, можно упростить, предварительно отбрасывая пары слов, не соответствующие эталонным значениям параметров (при поиске определённого словосочетания) и не согласованные по форме.

Для определения семантического смысла анализируемой пары слов будем каждое из них сопоставлять с каким-либо семантическим классом (СК). Считается, что два слова принадлежат к одному СК, если они имеют одинаковый смысл в заданном контексте. Соответствия слов семантическим классам и определённому контексту определяются по начальной форме слов и хранятся в базе данных. Начальная форма слова и его морфологические характеристики «вычисляются» внешней подсистемой на этапе морфологического анализа. Важно учитывать, что эта подсистема также не должна использовать базу онтологий, так как поступающие из неё данные могут быть неоднозначны из-за омонимии.

При расчете количества знаков препинания важно их ранжировать. Например, количество запятых и тире между двумя словами следует воспринимать по-разному. На основании эвристических правил знаки препинания были разделены на 3 класса. Таким образом, входной вектор имеет 5 полей (элементов). Итоговая структура входного вектора приведена в табл. 1.

Структура программного прототипа подсистемы семантико-синтаксического анализа предложений на русском языке приведена на рисунке. Модульный принцип структуры даёт возможность учитывать различные типы вводимых пользователем отношений для дальнейшего развития и повышения интеллектуальности системы символьной обработки.

Для каждого из типов отношений используется собственный распознаватель на основе специализированного модуля с «нейронным решателем». Каждая из нейронных сетей имеет своего учителя, а каждый учитель использует собственный генератор обучающей выборки. Также каждый распознаватель имеет доступ к модулю морфологического анализа, чтобы выделять необходимые параметры, например согласование слов по форме.

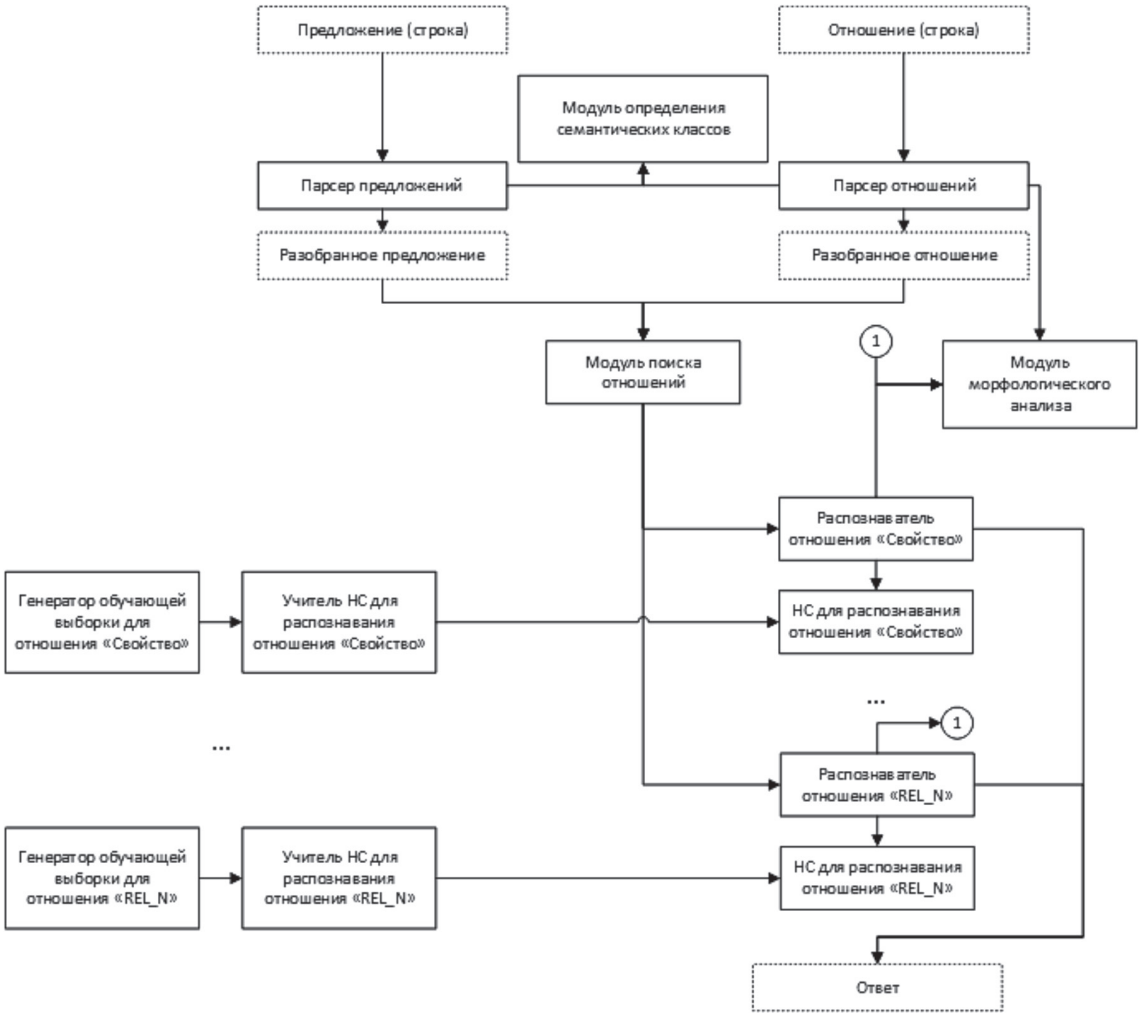
Пунктирными линиями на схеме отмечены данные и их внутренние представления. Предложение и искомое отношение в виде строк поступают на обработку в соответствующие парсеры (анализаторы). Во время разбора предложения слова помечаются согласно их семантическому классу. Парсер отношений также использует модуль морфологического анализа для определения типа отношения. После определения типа отношения данные передаются в соответствующий распознаватель, который и даёт искомый ответ.

Прототип был реализован на языке программирования Java 8. Модули искусственных нейронных сетей реализованы с применением фреймворка Deeplearning4j. Данный фреймворк высокопроизводителен, поддерживает LSTM-сети и имеет большое количество параметров для настройки как самих сетей, так и процесса их обучения.

Таблица 1

Значения ячеек входного вектора

№ п/п	Значение
1	Расстояние между словами
2	Длина групп слов, которые получаются путем разбиения промежутка между двумя исследуемыми словами знаками препинания
3	Количество знаков препинания класса 1 (,-)
4	Количество знаков препинания класса 2 (—:)
5	Количество знаков препинания класса 3 (.?!;)



Структура приложения

В качестве морфологического анализатора используется консольная программа свободного распространения Yandex MyStem [11], которая показала себя значительно лучше остальных решений по качеству анализа, особенно при работе с неизвестными словами.

Модуль определения семантических классов основан на использовании реляционной базы данных, содержащей слова в начальной форме, информацию об их принадлежности семантическим классам в определённом контексте. В качестве СУБД использована PostgreSQL 9.6. Заполнение таблиц было произведено на основе словаря синонимов Абрамова [12]. Незнакомые слова автоматически относятся к новым семантическим классам.

В качестве входных данных в экспериментах использовались первые две главы романа М.А. Булгакова «Мастер и Маргарита». Полученные данные содержали противоречия (менее 1% от всей выборки). Для того чтобы правильно провести обучение нейронной сети, из каждой пары строк противоречащих векторов был оставлен только один. В итоге была получена выборка размером около 1500 векторов. Эксперименты проводились для различных конфигураций указанных ранее нейронных сетей. В табл. 2 показаны рассмотренные варианты, в которых была получена наилучшая точность. Под точностью понимается отношение количества правильных ответов НС к общему количеству векторов в тестовой выборке.

По результатам экспериментов многослойный персептрон показал себя гораздо лучше LSTM-сети. Оптимальное количество эпох – 100. Оптимальная топология – один скрытый слой с 20 нейронами, передаточная функция – LEAKYRELU. Для выходного слоя используется SOFTMAX. Во всех случаях наилучшая точность приближается к 90%.

Таблица 2

Архитектуры НС, обеспечивших наилучшую точность

Тип НС	Размер	Количество эпох	Передаточная функция	Точность
MP	1x20	100	HARDSIGMOID	0,8767
MP	1x20	100	SOFTSIGN	0,8733
MP	1x20	100	LEAKYRELU	0,8817
MP	1x40	100	LEAKYRELU	0,8700
MP	1x20	200	LEAKYRELU	0,8700
LSTM	1x20	100	TANH	0,8650

Также были проведены эксперименты для определения временных затрат при обработке вышеуказанных текстовых файлов на русском языке. Получение ответа от модуля нейронной сети выполнялось в среднем за 0,1 мс. Полный поиск отношения «свойство» в предложении, включая морфологический разбор и обращения к БД, требовал в среднем 1,868 мс. При этом большую часть времени (1–1,5 мс) занимала работа программного модуля Yandex MyStem и обмен данными с ним. Необходимо заметить, что это время можно значительно уменьшить, посылая предложения на обработку большими группами.

Таким образом, по итогам первого этапа исследований можно сделать следующие выводы. Относительно небольшого количества данных, получаемых на основе анализа взаимного расположения слов в предложении (длина вектора для описания обнаруженной пары «взаимосвязанных» слов – пять вещественных чисел), и никоим образом не опирающегося на большие (иногда огромные) и трудоёмкие онтологии, достаточно, чтобы получить точность разбора текста на русском языке, близкую к 90%. Скорость принятия решения по одной паре слов очень высокая – до 10 тысяч проверок в секунду на подготовленных данных. При этом можно ожидать ещё лучшей точности при разборе предложений с учётом более сложных по структуре отношений, чем «свойство». Это обусловлено тем, что они будут порождать меньшее количество противоречивых данных, предоставляя при этом больше различных критериев для принятия решения.

Если точности в 90% недостаточно для качественной обработки текстов на естественном языке, например, в системах повышенной надёжности, то данную подсистему можно использовать на предварительном этапе анализа символьной информации подобной высокопроизводительной специализированной системы [13]. Такой подход значительно сократит как время предобработки текста на первом этапе, так и время «прецизионного» анализа текста в основном компоненте системы, основанном на базе проверенной и достоверной онтологии конкретной предметной области, отлаженной специалистами-экспертами в области инженерии знаний.

Список литературы

1. Евдокимова И.С. Естественно-языковые системы / И.С. Евдокимова – Улан-Удэ: Изд-во ВСГТУ, 2006. – 92 с.

2. Shipitsyna A.A., Meltsov V.Yu. An approach to designing intelligent test systems // European Conference on Innovations in Technical and Natural Sciences: 2nd International scientific conference. – 2014. – P. 28–33.

3. Dolzhenkova M.L., Meltsov V.Y., Strabykin D.A. Method Of Consequences Inference From New Facts In Case Of An Incomplete Knowledge Base // Indian Journal of Science and Technology. – 2016. – vol. 9, I. 39. – P. 100413.

4. Мельцов В.Ю., Чистяков Г.А. Эффективный метод построения оптимизированного дерева грамматического разбора формул темпоральной логики линейного времени / В.Ю. Мельцов, Г.А. Чистяков // Вестник Тамбовского государственного технического университета. – 2012. – Т. 18, № 4. – С. 813–820.

5. Мельцов В. Ю. Вывод следствий с построением схемы логического вывода / В.Ю. Мельцов, Д.А. Страбыкин // Фундаментальные исследования. – 2013. – № 11–8. – С. 1588–1593.

6. Ножов И.М. Морфологическая и синтаксическая обработка текста (модели и программы) [Электронный ресурс]. – Режим доступа: http://aot.ru/docs/Nozhov/chapter1.pdf (дата обращения: 11.09.17).

7. Сокирко А.В. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ): дис. ... канд. техн. наук. – Москва, 2001. – 120 с.

8. ABBYY Compreno [Электронный ресурс] / ABBYY: сайт. – URL: https://www.abbyy.com/ru-ru/infoextractor/compreno (дата обращения: 12.10.2017).

9. Шелманов А.О. Исследование методов автоматического анализа текстов и разработка интегрированной системы семантико-синтаксического анализа: дис. … канд. техн. наук. – ФГБУН «Институт системного анализа» РАН. – Москва, 2015. – 210 с.

10. Хайкин С. Нейронные сети: полный курс, 2-е издание. / Саймон Хайкин. – Пер. с англ. – М.: Издательский дом «Вильямс», 2006. – 1104 с.

11. MyStem [Электронный ресурс] / Технологии Яндекса: сайт. – URL: https://tech.yandex.ru/mystem/ (дата обращения: 11.10.17).

12. Абрамов Н.А. Словарь русских синонимов и сходных по смыслу выражений / Н.А. Абрамов – М.: Изд. Русские словари, 1999. – 431 с.

13. Meltsov V.Yu. High-performance systems of deductive inference: Monograph // Science Book Publishing House LLC. – Yelm, WA, USA, 2014. – 215 p.