УДК 004.891.3

# Выбор моделей представления синтаксической структуры предложений и инструментальных средств синтаксического анализа текста на русском языке

**Р. Р. Вафин**

vafrus74@gmail.com

ФГБОУ ВО «Уфимский государственный авиационный технический университет» (УГАТУ)

**Аннотация.** Одним из источников получения медицинских знаний для систем поддержки принятия решений является профессиональная медицинская литература. Автоматизированный анализ затруднен вследствие слабой структурированности информации на естественном языке. Основная проблема анализа встречается на стадии синтаксического анализа предложений и интерпретации результатов. На данный момент существуют различные модели представления синтаксической структуры предложения, которые в разной степени подходят для разных языков. Приводится сравнительный анализ моделей представления синтаксической структуры предложений и инструментальных средств синтаксического анализа текста на русском языке.

**Ключевые слова:** СППВР; медицина; обработка естественного языка; синтаксический анализ; грамматика зависимостей; грамматика составляющих.

## ВВЕДЕНИЕ

В настоящее время в медицине активно применяются информационные технологии для увеличения качества предоставления медицинских услуг. Одной из технологий является применения систем поддержки принятия врачебных решений (СППВР). СППВР помогают врачам анализировать. Одна из основных проблем современных систем поддержки принятия врачебных решений является отсутствие достаточной по объему базы знаний для принятия решений [1, 2].

Одним из основных источников получения медицинских знаний являются клинические рекомендации [3]. Автоматизированный анализ данного источника крайне сложен, вследствие неструктурированной и неформальной природы информации на естественном языке [4, 5]. Для извлечения знаний требуются методы интеллектуального анализа текста.

Для анализа текста последовательно применяют следующие типы анализа: лексический; морфологический; синтаксический; семантический [6]. Лексический анализ это выделение последовательности слов и знаков пунктуации из непрерывной последовательности входных символов текста. Морфологический анализ это – сопоставление отдельных слов и словоформ в словаре и выяснением грамматических характеристик слов [7]. Лексический анализ русского языка не вызывает проблем. Морфологический анализ сталкивается с проблемой омонимов. Но в контексте формальных медицинских текстов данная проблема менее значима.

Основная проблема состоит в синтаксическом анализе, так как часть информации не имеет жестко заданной структуры, а является обычным текстом, смысл которого человек может легко понять, но автоматизированные системы не могут точно выявить нужные знания из такого неструктурированного текста [6, 8].

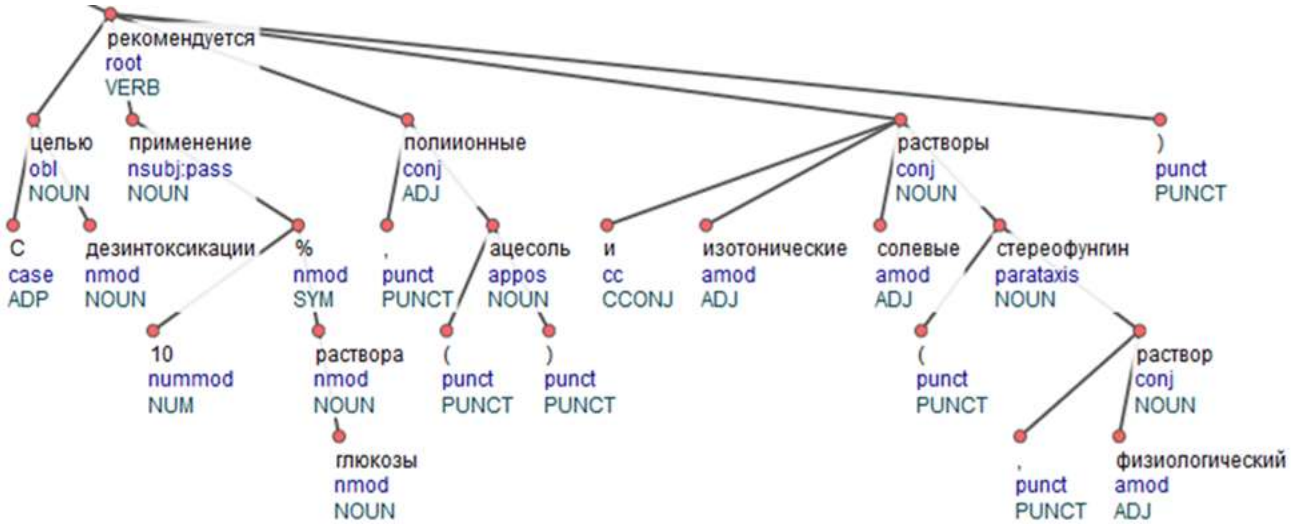


**Рис. 1.** Пример дерева зависимостей

Для синтаксического анализа были разработаны различные методы и соответствующие программные реализации. Цель данной статьи проанализировать методы синтаксического анализа предложений и выявить наиболее подходящий метод для анализа узкоспециализированных медицинских текстов (клинических рекомендаций) и выбрать программные инструменты для осуществления выбранного метода.

В общем понимании синтаксический анализ – это процесс сопоставления линейной последовательности лексем естественного или формального языка с его формальной грамматикой. Результатом работы синтаксического парсера является граф, узлами которого выступают слова в предложении. Если слова каким-либо образом связаны, то соответствующие вершины графа связаны дугами.

Грамматика составляющих основана на постулате, согласно которому всякая сложная грамматическая единица складывается из двух более простых и не пересекающихся единиц, называемых ее непосредственными составляющими [6]. Составляющая, включающая более одного слова, называется, а слово, соответствующее корневому узлу в дереве зависимостей, описывающем группу, вершиной группы. Представление синтаксической структуры предложения в виде иерархии непосредственных составляющих используется в различных вариантах в формальных моделях языка, в частности в генеративной лингвистике Н. Хомского.

Другой способ представления синтаксической структуры предложения состоит в том, чтобы выяснить зависимости между словами в предложении. Данная модель называется грамматика зависимостей. В грамматике зависимостей порядок слов в предложении не важен. Пример дерева зависимостей показан на рис. 1.

Одна из проблем грамматики составляющих заключается в снятии неоднозначностей (синтаксической омонимии) [6]. В русском языке существует относительно свободный порядок слов в предложении, поэтому одному и тому же предложению будут соответствовать несколько синтаксических деревьев. Для таких случаев использование подхода к анализу на основе грамматики составляющих вызывает значительные трудности. В связи с этим, значительно продуктивнее использовать подход на основе грамматики зависимостей.

В настоящее время существует большое число программных средств для анализа естественного языка. В табл. 1 приведены наиболее популярные библиотеки, которые поддерживают синтаксический анализ предложений русского языка.

Из данной таблицы видно, что наиболее подходящим являются UDPipe. Томита-парсер и NLTK работают на основе базы правил, которые необходимо заполнять вручную, либо с помощью методов машинного обучения [8, 12]. В библиотеке UDPipe есть готовая обученная модель на основе корпуса русского языка SynTagRus.

Таблица 1

**Сравнительная таблица программных пакетов для анализа текста**

| *Библиотека* | *Построение дерева зависимостей* | *Готовые модели для русского языка* | *Выделение ключевых компонентов и шаблонов* | *Лицензия* | *Использование стороннего морфологического анализатора* |
|---|---|---|---|---|---|
| SyntaxNet [9] | + | - | - | Apache License 2.0 | + |
| Stanford CoreNLP [10] | + | - | - | GPL v2 | + |
| UDPipe [11] | + | + | + | Mozilla Public License 2.0 | + |
| Томита-парсер [12] | + | - | + | Mozilla Public License 2.0 | + |
| NLTK[8] | + | - | + | Apache License 2.0 | + |

Так же UDPipe поддерживает сторонние морфологические анализаторы и имеет API интерфейс для использования в модулях, реализованных на языке python.

Рассмотренные модели синтаксических структур могут применяться для различных языков. Деверья составляющих подходят для языков со строгим порядком слов, например для английского языка. Для русского языка более эффективным будет применение деревьев зависимостей, ввиду более свободного порядка слов. Для анализа и представления синтаксической структуры предложения в выбранной модели предлагается использовать программный пакет и набор обученных моделей UDPipe.

**СПИСОК ЛИТЕРАТУРЫ**

1. **Атьков, О.Ю.** Система поддержки принятия врачебных решений. //Врач и информационные технологии. – 2013. – № 6. – С. 67–75. [O. U. At'kov “Medical Decision Support System,” in Vrach i informatsionnye tekhnologii, 2013, vol. 6, pp. 67-75.]

2. **ГОСТ Р 56034-2014** Клинические рекомендации (протоколы лечения). Общие положения М.: Стандартинформ, 2015. [Clinical recommendations (Protocols for patient's cure), General regulations Federal standard R 56034-2014, Moscow, Standatrinform, 2015.]

3. **Виноградов А. Н.** Перспективные направления исследований в области клинического моделирования, управления и принятия решений //Врач и информационные технологии. – 2014. – №. 5. [A. N. Vinogradov, “Promising areas of research in the field of clinical modeling, management and decision making,” in Vrach i informatsionnye tekhnologii, 2014, vol. 5.]

4. **Кобринский Б. А.** Системы поддержки принятия решений в здравоохранении и обучении //Врач и информационные технологии. – 2010. – №.2. [B. A. Kobrinskii “Decision Support Systems in Health and Education,” in Vrach i informatsionnye tekhnologii, 2010, vol. 2.]

5. **Ермакова Л. М.** Методы извлечения информации из текста //Вестник Пермского университета. Серия: Математика. Механика. Информатика. – 2012. – №. 1. – С. 77–84. [L. M. Ermakova “Methods for extracting information from text,” in Vestnik Permskogo universiteta. Seriya: Matematika. Mekhanika. Informatika, 2012, vol. 1, pp. 77-84.]

6) **Батура Т.В., Чаринцева М.В.** Основы обработки текстовой информации: учебное пособие / Институт систем информатики им. А.П. Ершова СО РАН. Новосибирск, 2016. [T.V. Batura and M.V. Charintseva “Text Processing Basics: Tutorial,”, A. P. Ershov Institute of Informatics Systems, Novosibirsk, 2016.]

7. **Korobov M.** Morphological analyzer and generator for Russian and Ukrainian languages //International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2015. – С. 320–332. [M. Korobov “Morphological analyzer and generator for Russian and Ukrainian languages,” Analysis of Images, Social Networks and Texts, 2015, pp. 320-332.]

8. **Bird S., Klein E., Loper E.** Natural language processing with Python: analyzing text with the natural language toolkit. – O'Reilly Media, Inc., 2009. [S. Bird, E. Klein, E. Loper “Natural Language Processing with Python,” O’Reilly Media, Inc. Sebastopol, CA, USA, 2009.]

9. SyntaxNet: A TensorFlow toolkit for deep learning powered natural language understanding. [Электронный ресурс]. URL: https://github.com/tensorflow/models/tree/master/research/syntaxnet (дата обращения: 20.01.2020). [SyntaxNet: A TensorFlow toolkit for deep learning powered natural language understanding. [Online]. Available: https://github.com/tensorflow/models/tree/master/research/syntaxnet]

10. **Chen D., Manning C. D.** A fast and accurate dependency parser using neural networks //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 740-750. [D.Chen, C. D.Manning “A fast and accurate dependency parser using neural networks” in Proceedings of the 2014 conference on empirical methods in natural language processing, 2014, pp. 740-750.]

11. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. [Электронный ресурс]. URL: http://ufal.mff.cuni.cz/udpipe (дата обращения: 20.01.2020). [Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe [Online]. Available: http://ufal.mff.cuni.cz/udpipe]

12. Томита-парсер [Электронный ресурс]. URL: https://yandex.ru/dev/tomita/ (дата обращения: 20.01.2020). [Tomita-parser [Online]. Available: https:// yandex.ru/dev/tomita/]

**ОБ АВТОРЕ**

**ВАФИН Руслан Рустамович**, Дипл. бакалавра информатика и вычислительная техника (УГАТУ, 2018), магистрант 2-го курса факультета ИРТ, УГАТУ.

**METADATA**

**Title:** Analysis of models of representation of syntactic structure of proposals and instruments for syntactic analysis of Russian texts

**Author:** R. R. Vafin

**Affiliation:**

Ufa State Aviation Technical University (UGATU), Russia.

**Email:** vafrus74@gmail.com

**Language:** Russian.

**Source:** Molodezhnyj Vestnik UGATU (scientific journal of Ufa State Aviation Technical University), no. 1 (22), pp. 30-33, 2020. ISSN 2225-9309 (Print).

**Abstract:** One of the sources of obtaining medical knowledge for decision support systems is professional medical literature. Automated analysis is difficult due to the weak structure of information in natural language. The main problem of analysis is encountered at the stage of parsing sentences and interpreting the results. At the moment, there are various models for representing the syntactic structure of sentences, which are to varying degrees suitable for different languages. A comparative analysis of presentation models of the syntactic structure of sentences and tools for parsing Russian text is given.

**Key words:** Medical DSS; medicine; natural language processing; parsing; dependency grammar; phrase structure grammar.

**About author:**

**VAFIN, Ruslan Rustamovich**, Master Student UGATU, Bachelor of Computer Science and Engineering (UGATU, 2018).