УДК 004.822

ИНДЕКСАЦИЯ ДЛЯ МОРФОЛОГИЧЕСКОГО И СИНТАКСИЧЕСКОГО АНАЛИЗА ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ НА ОСНОВЕ СЕМАНТИЧЕСКОЙ СЕТИ

Шибаев С.С. (Университет ИТМО)

Научный руководитель – старший преподаватель факультета программной инженерии и компьютерной техники, Клименков С.В.

(Университет ИТМО)

Введение. Эффективное решение задачи морфологического и синтаксического анализа текста на естественном языке в настоящее время является актуальной задачей для разных сфер жизни человечества. Для создания программного обеспечения, решающего данную задачу, в качестве базы знаний для хранения словоформ естественного языка может быть использована семантическая сеть [1].

Семантическая сеть — структура данных, состоящая из узлов, соответствующих понятиям, и связей, указывающих на взаимосвязи между узлами [2]. При помощи семантической сети появляется возможность оперировать смыслами, а не словами при работе с текстами на естественном языке [3]. Сама по себе семантическая сеть не является инструментом анализа текста, но представляет интерес как база для построения таких инструментов. Благодаря вышеописанным свойствам семантической сети в ней можно эффективно хранить информацию о словообразовании, словоизменении и определении синтаксических признаков отдельных слов [4].

Сейчас для решения задачи морфологического и синтаксического анализа текста на естественном языке в основном применяются методы машинного обучения, что не является эквивалентным способом решения поставленной задачи с помощью создания точной и однозначной базы знаний, что позволит решать данную задачу менее ресурсоемко и точнее, однако, зачастую медленнее [2]. Для увеличения производительности, а именно для уменьшения времени обработки анализа естественного языка, необходимо создать индексацию, которая позволит быстрее производить итерации по словоформам.

Основная часть. В качестве источника данных для решения задач морфологического и синтаксического анализа текста, а также задач анализа словоизменения на русском языке, был выбран грамматический словарь русского языка А.А. Зализняка, содержащий около 100000 словарных статей. Каждая статья содержит информации о словообразовании, словоизменении и определении синтаксических признаков отдельных слов. Отдельная словарная статья содержит главное слово, несущее основную смысловую нагрузку статьи, и группу словоформ, получающуюся из главного слова путем словоизменения (спряжения или склонения) [1].

Основным этапом является именно создание эффективной индексации для представленных словарных статей на основе синтаксических признаков словоформы (падеж, число, спряжение, склонение и т.д.). Необходимо было произвести сравнение созданных решений на базе индексации на разных синтаксических признаках словоформы и выбрать решение, обеспечивающие наименьшее время обработки.

Выводы. Проведен анализ семантической сети как базы знаний и проведена апробация использования индексации для уменьшения времени поиска в семантической сети.

Список использованных источников:

1. Д.Е.Шуклин Структура семантической нейронной сети, реализующей морфологический и синтаксический разбор текста // Кибернетика и системный анализ. Киев. Изд-во Института кибернетики НАН Украины, 2001. - № 5. С. 172-179.
2. John F. Sowa. Semantic Networks. [Электронный ресурс]. Режим доступа: https://www.jfsowa.com/pubs/semnet.htm (дата обращения: 6.02.2023).

3. Письмак А.Е., Харитонова А.Е., Цопа Е.А., Клименков С.В. Метод автоматического формирования семантической сети из слабоструктурированных источников // Программные продукты и системы - 2016. - № 3. - С. 74-78
4. Клименков С.В., Цопа Е.А., Жмылёв С.А., Покид А.В., Ткешелашвили Н.М. Метод быстрого поиска узлов семантической сети по точному совпадению словоформы. Известия высших учебных заведений. Приборостроение. 2017. Т. 60. № 10. С. 932-939. [Тип: Статья, Год: 2017]