МЕТОДЫ УЛУЧШЕНИЯ ТОМИТА-ПАРСЕРА ДЛЯ АНАЛИЗА УЧЕБНЫХ ТЕКСТОВ И СТАТЕЙ

Огородник Р. В., Серебряная Л. В.

УО «Белорусский государственный университет информатики и радиоэлектроники», Минск, Республика Беларусь, aharodnik@gmail.com

Для изучения учебного материала необходимо применять средства автоматизации изучения, в частности анализаторы текстов, чтобы иметь возможность сократить объём прочитанного материала для изучения какой-либо конкретной части учебного материала.

В качестве инструмента для анализа часто используются GLR-парсеры, одним из самых популярных является Томита-парсер. Несмотря на развитые правила и систему добавления грамматик Томита-парсер оставляет множество идей и улучшений для увеличения возможностей целевого анализа текста.

В результате исследования были предложены следующие методы улучшения Томита-парсера:

- слияние синонимичных сущностей,
- слияние сущностей представленных местоимениями,
- исключение «мусорных» стоп-слов.

Слияние синонимичных сущностей осуществляется с помощью определения словаря синонимов. С использованием словаря синонимом, перед тем, как использовать текст в анализаторе проводится замена дополнительных синонимичных сущностей на основную сущность. Таким образом, использование синонимов позволяет расширить сеть фактов и данных для какого-либо понятия, делает граф слов более свзяным.

Слияние сущностей, представленных местоимениями, преследует ту же цель, что и слияние синонимичных сущностей. Процесс слияния более сложный и требует на момент слияния определения грамматической роли всех слов. Личные местоимения, относящиеся к сущностям сливаются с сущностями именами существительными, которые имеют те же грамматические свойства, что и местоимение, которые заменяется найденным существительным. Во время поиска соответствующее местоимение сливается с подходящим по грамматическим свойствам существительным, которое является ближайшим по тексту непосредственно перед местоимением.

Исключение мусорных слов производиться, чтобы избавить текст от лишней информации, которая никак не помогает в извлечении фактов, замусоривает выдачу, делает граф менее связным и определяет высокую степень соседства и связанности для сущностей, на самом деле лежащих на большем расстоянии друг от друга.

Таким образом, эти методы для улучшения Томита-парсера помогают производить более целевой анализ текста

Литература