УДК 004.822

Постобработка форматирования pdf-документов, обработанных для использования в семантической сети

Черный К.В. (Университет ИТМО)
Научный руководитель – старший преподаватель Клименков С.В.
(Университет ИТМО)

Аннотация.
Доклад посвящен проблеме распознавания форматирования текста, распознанного из pdf-документа. Решение основано на учете статистики расположения косвенных признаков в документе. Результат исследования может использоваться в парсерах натурального языка для увеличения доли сохраненного контекста.

Введение.
Существует семантическая сеть, которая может быть использована в различных областях исследований. Сеть может быть использована, например, для перевода текстов с сохранением контекста, но для этого исходный документ необходимо считать. Текущее решение считывания документа позволяет распознать текст буквально, но не все элементы форматирования(такие как списки) восстанавливаются. Это приводит к проблеме потери контекста, которую должно решить мое исследование.

Основная часть.
В разработанном решении рассматривается алгоритм восстановления форматирования pdf-документов на основе статистики по косвенным признакам расположения текста и распознанных элементов форматирования, таких как отступы, цвет и размер шрифта. На вход мы получаем набор лексем, к которым прилагаются позиция лексемы в документе(ее координаты, отступы, тип, цвет, стиль, шрифт). Модуль выявляет закономерности, например, текст в абзаце имеет один и тот же отступ и размер шрифта, между абзацами есть расстояние. В списках по госту имеется нумерация, отступ от края. Проблемой являются вложенные списки, которые локально могут иметь различное форматирование или стартовый символ, но они отделены от основного текста или заголовка и мы на основе статистики можем сделать вывод о том, что набор лексем является частью списка.

Выводы.
Результат исследование предлагается использовать в семантической сети для улучшения качественного распознавания текста, в том числе pdf-документов.

Черный К.В. (автор) Подпись

Клименков С.В. (научный руководитель) Подпись