

Отзыв на магистерскую диссертацию Бодровой Анастасии Александровны

Разрешение кореференции методом кластеризации

Разрешение кореференции является одной из ключевых подзадач обработки текста и заключается в объединении текстовых упоминаний, относящихся к одной сущности. Текстовые упоминания могут быть выражены именами собственными, именной группой или местоимениями. К настоящему времени разработано множество систем, позволяющих решать задачу кореференции. Чаще всего, они используют подходы машинного обучения на основе различных лингвистических и концептуальных признаков. Ключевыми типами моделей являются попарная модель (mention-pair model), модель ранжирования (ranking model) и сущностная модель (entity-based model). Современные исследования направлены на улучшение результатов методов машинного обучения "без учителя" и сущностные модели.

В работе Бодровой А. рассматривается применение алгоритма кластеризации для разрешения кореференции на русскоязычных новостных текстах. В качестве объектов кластеризации рассматриваются имена собственные, которые относятся к персонам. В работе предложен алгоритм, включающий в себя два этапа:

1. Извлечение упоминаний. Для извлечения именованных сущностей был использован свободно-доступный инструмент Томита-парсер, использующий формализм контекстно-свободных грамматик и газеттиры для извлечения фактов. Для извлечения имен собственных была написана контекстно-свободная грамматика.

2. Кластеризация. Для объединения извлечённых имён в кластеры, использовалась аггломеративная кластеризация. Суть алгоритма в следующем: изначально каждое упоминание находится в собственном одноэлементном кластере, затем на каждом шаге два наиболее подходящих кластера объединяются. Объединение кластеров происходит на уровне сущностей: любая пара упоминаний между объединяемыми кластерами не должна содержать противоречий.

Эксперименты проводились на новостных текстах, размеченных для соревнования Dialogue Evaluation factRuEval-2016. В работе приведено сравнение полученных результатов с результатами участников соревнования.

Своей работой автор продемонстрировала, что она в достаточной мере креативна, умеет работать с российскими и зарубежными источниками информации, владеет методами программирования и знакома с приемами обработки и анализа данных. Полагаю, что работа Бодровой А.А. может быть оценена на "отлично".



Научный руководитель

канд. физ-мат наук,

доцент Графеева Н.Г.