НОВОСТНЫЕ КЛАСТЕРЫ КАК ИСТОЧНИК ПОЛУЧЕНИЯ ШАБЛОНОВ В ЗАДАЧЕ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ТЕКСТОВ

Котельников Д. С.

Московский государственный университет им. М. В. Ломоносова,
Ф-т вычислительной математики и кибернетики, каф. Алгоритмических языков,
Московская обл., г. Королев, ул. Мичурина, д. 21, кв. 631
Тел.: (+7 916) 410-87-82, E-mail: dimonnot@mail.ru

Задача извлечения информации из текстов состоит в выделении из неструктурированной информации на естественном языке структурированной информации. Стандартными подзадачами данной задачи являются извлечение совокупности упоминаемых в тексте сущностей, отношений между этими сущностями, ситуаций, в которых участвовали эти сущности.

Существующие методы извлечения информации можно разделить на два принципиально различных класса: методы, основанные на знаниях, и методы машинного обучения. В методах, основанных на знаниях, шаблоны выделения событий задаются экспертами. Недостатком этого подхода является высокая трудоемкость создания системы и сложность её адаптации для извлечения новых событий. При применении методов машинного обучения используется коллекция документов, предварительно размеченная человеком. Создание такой коллекции обучения также является трудоемкой задачей.

В работе исследуются методы пополнения и обобщения шаблонов, извлекающих информацию из текста, за счет нахождения в новостном кластере (кластере похожих новостей) нескольких близких по содержанию предложений, в которых хотя бы в одном удалось обнаружить извлекаемое событие. Для оценки качества методов были проведены эксперименты по извлечению информации о фактах получения кредитов из новостных документов.

В качестве базового инструмента извлечения информации из текста использовалась программа RCO Fact Extractor [1], кластеризация новостных сообщений производится новостным кластеризатором НИВЦ МГУ[2].

Литература

1. *Киселев С. Л., Ермаков А. Е., Плешко В. В.* Поиск фактов в тексте естественного языка на основе сетевых описаний. // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог’2004. – Москва, Наука, 2004. – С. 282-285.
2. *Лукашевич Н.В., Добров Б.В.* Автоматическое аннотирование новостных кластеров на основе тематического представления. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15).– М.: РГГУ, 2009. 620 с.