УДК 004.896

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ МАССИВОВ ТЕКСТОВ И ЕЕ СТРУКТУРИЗАЦИЯ С ПОМОЩЬЮ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Ходорченко М.А. (Университет ИТМО)
Научный руководитель – к.т.н., доцент ФЦТ Бутаков Н.А.
(Университет ИТМО)

Для построения качественных вопросно-ответных систем необходимо иметь структурированную информацию по интересующей области знаний. Автоматическое построение онтологий и соответствующих графов знаний в высокой степени увеличивает скорость адаптации к новой информации. Использование тематических моделей для решения данной задачи обосновано из гибкостью и способностью извлекать кластера связанных между собой сущностей.

На сегодняшний день построение онтологий зачастую осуществляется с помощью привлечения экспертов, т.е. в ручном режиме. Основные исследования концентрируются на наборе задач, включающем в себя заполнение графов знаний и добавление пропущенных связей в существующие онтологии. Задача построения самих онтологий в автоматическом или полуавтоматическом режиме сложна и малоизучена. Среди работ можно отметить использование обучение с подкреплением для перевода текста в форму графа и обратно, а также вариационные автоэнкодеры. К недостаткам существующих подходов можно отнести то, что они все же больше направлены на формирование графов знаний, а не онтологий.

Для эффективного выделения кластеров сущностей из текстов, а также триплетов субъект-отношение-объект из текстов необходимо знание статистических характеристик сущностей, которые могут быть сильно разнесены по корпусу. Для этого можно использовать тематические модели, которые обладают рядом достоинств - выделение связанного набора терминов, представляющего концепт онтологии и узлы-сущности для графа знаний; гибкость и устойчивость модели, обоснованная тем, что для обучения нет необходимости в метках классов, хотя есть модификации и с использованием меток-сидов; поддержка модальностей (время). Предлагаемый способ состоит из использования автоматической настройки для получения необходимых параметров тематической модели, с помощью обучения на конечной задаче, а именно проверки качества на подмножестве вопросов вопросно-ответной системы. Получаемые в таком случае темы представляют собой значимые подмножества сущностей и связей, которые могут быть при необходимости интерпретированы. Составление графов знаний, т.е. извлечение триплетов возможно через подсчет количества связей между темами в рамках сегментов текстов.

Автоматическое построение онтологий и соответствующих графов знаний имеет важное практическое приложение - позволяет производить структуризацию данных с целью построения вопросно-ответных систем способных эффективно обрабатывать запросы пользователей. Полученные результаты позволяют сделать вывод о том, что задача может быть решена с использованием тематического моделирования при правильном выборе функции качества и способа настройки параметров.

Ходорченко М.А. (автор)

Бутаков Н.А. (научный руководитель)