УДК 004.8

Извлечение фактов из неструктурированного текста для языков с сложной морфологией

Будков С. А., Университет ИТМО, г. Санкт-Петербург
Бурая К. И., Университет ИТМО, г. Санкт-Петербург
Научный руководитель – Фильченков А. А., к.ф.-м.н., доц. ФИТиП Университета ИТМО

Введение

Извлечение фактов из сети интернет является актуальной задачей сегодня. Одной из целей является создание системы, которая могла бы извлекать высококачественные факты из интернета, ограничиваясь только их количеством в сети, и не завися от вмешательства человека. В любом естественном языке, в том числе русском, используются различные конструкции для обозначения схожей информации, что усложняет семантический анализ текста. Возникает задача структурирования текста для обеспечения возможности его обработки методами машинного обучения, которую можно решить использованием онтологий – формализации знаний с использованием концептуальной схемы. Одним из перспективных подходов для извлечения фактов из неструктурированного текста для дальнейшего построения онтологий является подход «нескончаемого обучения». Его основным преимуществом является отсутствие необходимости предварительной разметки данных и вмешательства человека. Для обработки русского и других балто-славянских языков следует учитывать их сложную морфологию.

Цель работы

Целью данной работы является улучшение точности работы ранее адаптированного алгоритма CPL-RUS, подбор наилучших начальных параметров для работы алгоритма, проведение экспериментов на текстовом корпусе, составленном из всех русскоязычных статей ресурса Wikipedia, без учета структурных особенностей данного ресурса.

Базовые положения исследования

В ходе данной работы были исследованы два подхода для фильтрации извлеченных сущностей и шаблонов, были проведены эксперименты для выявления наилучшей стратегии. Для повышения точности работы ранее адаптированного алгоритма была введена дополнительная фильтрация по минимальному количеству совместных вхождений шаблона или сущности вместе со словом категории, а также проведены эксперименты для выявления влияния ограничения на точность работы алгоритма. При анализе результатов, полученных ранее адаптированным алгоритмом, было обнаружено, что некоторые шаблоны, извлекаемые на различных итерациях, являются сложными шаблонами, т. е. содержат более общий шаблон и сущность, относящуюся к категории. Такие шаблоны чаще извлекали ложные факты, что ухудшало точность работы алгоритма, поэтому был добавлен метод выделения сложных шаблонов и их замены на более общие. Чтобы избежать возможные потери части сущностей при замене сложных шаблонов, были добавлены контейнеры сущностей. Была добавлена обработка сложных сущностей, т. е. пар прилагательное и существительное или существительное и прилагательное. Обработка таких сущностей позволяет повысить абстрактность шаблонов, убрав из них лишние слова, описывающие конкретные сущности, а также может быть использована для дальнейшего анализа текста. Были проведены исследования на текстовом корпусе, составленном из всех статей русскоязычной Википедии, без учета структурных особенностей ресурса.

Практические результаты

В рамках данной работы были произведены улучшения алгоритма и проведены новые эксперименты на текстовом корпусе, содержащем более 1.3 миллиона страниц. Адаптированный ранее алгоритм позволял извлекать факты из неструктурированного русского текста с точностью от 16% до 100% и с средней точностью 61%. После внедрения

разработанных улучшений точность извлекаемых фактов составила от 43% до 96%, с средней точностью 69%. Полнота работы алгоритма с внедренными улучшениями составила от 67% до 86% в зависимости от категории, полнота для ранее адаптированного алгоритма не измерялась. Данные результаты позволяют применять алгоритм уже сегодня, но также показывают необходимость дальнейших исследований: разработка улучшений для стабилизации точности извлекаемых фактов для категорий, разработка более гибких методов фильтрации для извлекаемых фактов.

Список литературы

1. Carlson A. et al. Coupled semi-supervised learning for information extraction //Proceedings of the third ACM international conference on Web search and data mining. – ACM, 2010. – С. 101-110.
2. Buraya K. et al. Toward Never Ending Language Learning for Morphologically Rich Languages //Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing. – Association for Computational Linguistics, 2017.