УДК 004.85

Разработка алгоритма расширения базовых онтологий для морфологически богатого языка

Кустарева А. Д., Университет ИТМО, Санкт-Петербург

Научный руководитель Сметанников И. Б., Университет ИТМО, Санкт-Петербург

Введение

Морфологически богатые языки имеют сложную структуру: многообразие словоформ и грамматических конструкций позволяет конструировать сложные предложения, использующие широкий словарный запас и не имеющие четкой структуры. Это создает сложности при решении задач, связанных с автоматической обработкой текстов на таких языках. Онтологии — инструмент, который позволяет формально описать структуру языка. Их использование позволяет снизить сложность целого класса задач, связанных с обработкой текстов, благодаря организации информации в структурированной форме.

Цель работы

Целью данной работы является разработка алгоритма расширения базовых онтологий для русского языка на основе корпуса текстов для обучения.

Базовые положения исследования

В общем случае онтологии — инструмент для формализации некоторой системы с помощью концептуальной схемы. Основная цель использования онтологий в снижении сложности работы с некоторым материалом путем организации информации в структурированной форме. Обычно под онтологией понимается структура из множества объектов заданной области и формальных связей и ограничений между этими объектами.

Применительно к языку, области — это категории, объединяющие схожие по смыслу и использованию слова, объекты области — обозначенные ранее слова, а связи между ними — текстовые паттерны, которые используются в языке для создания связей между категорией и объектом в ней или между двумя объектами одной категории.

Задача расширения базовой онтологии состоит в следующем. Входными данными алгоритма является базовая онтология (множество слов, принадлежащих заданной категории, и языковые выражения, которые могут связывать слова из категории с самой категорией). Выходом алгоритма является расширенное множество слов, принадлежащих категории, и расширенный список выражений языка, использующихся для связи категории и её объектов.

В ходе исследования был реализован следующий алгоритм. На каждом его шаге есть доверительное множество объектов — это слова, которые алгоритм считает принадлежащими заданной категории, и доверительное множество паттернов — это выражения, которые алгоритм считает можно использовать для связи категории и её объектов. На первом шаге эти множества есть исходное множество слов и исходное множество выражений соответственно. На каждом следующем шаге алгоритм пытается их расширить. Расширение можно поделить на два основных этапа:

1. Извлечение новых объектов. Алгоритм последовательно анализирует тексты из тренировочного корпуса. Выделяются конструкции, в которых употребляется связка из списка доверительных паттернов и в качестве аргумента она содержит категорию. Тогда второй аргумент связки добавляется в множество слов-претендентов на попадание в доверительный список на следующем шаге.

2. Извлечение новых паттернов. Аналогично предыдущему пункту анализируются тексты из тренировочного множества. Выделяются конструкции, в которых присутствуют категория и слово из доверительного множества объектов. Связка между этими словами добавляется в множество связок-претендентов на попадание в доверительное множество паттернов на следующем шаге.

После завершения процедуры расширения происходит анализ претендентов на попадание в соответствующие доверительные списки. Оценка релевантности найденных объектов и паттернов базируется на вычислении относительной частоты их встречаемости. Алгоритм завершает работу после прохождения заранее указанного количества итераций.

Результаты

В ходе работы были рассмотрены различные методы решения поставленной задачи. Был реализован описанный выше алгоритм, показывающий наиболее успешные результаты. Рассматриваются возможные улучшения текущего алгоритма благодаря использованию другой метрики для оценки релевантности претендентов или другой стратегии их отсечения.