УДК 004.822

Метод извлечения лемм для смысловых узлов на разных языках

Редькина И.В. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Научный руководитель – старший преподаватель Клименков С.В.

(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В докладе рассматривается способ расширения семантической сети и добавления ей языкового разнообразия путём создания алгоритма, способного обрабатывать неструктурированную информацию в транслингвальной секции словаря Wiktionary.

Введение. Каждый день человек перерабатывает огромное количество различных текстов в попытке получить некоторую необходимую информацию, но чаще всего все эти тексты находятся в неструктурированном формате, что приводит к необходимости создания баз данных, содержащих понятия для конкретных предметных областей. Но такие базы знаний после их создания в какой-то момент становятся слишком большими, чтобы их возможно было постоянно поддерживать вручную, а также очень быстро меняются. И появилась необходимость в определении некоторого формата таких данных, чтобы им было удобно пользоваться, а также можно было бы реализовать алгоритм, способный поддерживать заполнение этой базы новыми данными и следить за их консистентностью. Кроме того, такие базы могут содержать в себе информацию не только, например, на русском языке, но и на других, и необходимо поддерживать связи между понятиями на разных языках, обозначающими один и тот же предмет. А также структура статей в открытых для редактирования словарях достаточно разнообразна на разных языках и содержит ошибки, что говорит о необходимости создания алгоритмов, способных обрабатывать такие статьи. Стоит отметить, что исследований в данной области с использованием семантических сетей и транслингвальной секции в словаре Wiktionary проводилось очень мало, а те, кто их изучали, ограничивались изучением только двух-трёх языков в один момент

Основная часть. Предлагается подход к обработке иерархической структуры открытого для редактирования словаря Wiktionary для распознавания расположения секции переводов на странице статьи для конкретного смыслового значения. Во-первых, была изучена структура кода для двадцати иностранных языков, доступных на сайте открытого словаря. Исследование кода для транслингвальной секции словаря показало, что все языки можно разделить на группы, имеющие сходную структуру, то есть для них можно составлять алгоритмы, умеющие обрабатывать сразу несколько языков. Во-вторых, было выбрано пять языков, для которых будут проводиться дальнейшие исследования. Далее происходила разработка и реализация алгоритмов на языке Java, способных находить на странице секцию переводов и создавать новые связи внутри семантической сети между смысловыми значениями, обозначающими одно и то же понятие на разных языках. И наконец, алгоритм был протестирован на выбранных группах языков, а полученные результаты изучены. Кроме того, планируется рассмотреть возможность создания классификатора и нейронной сети, чтобы впоследствии сравнить полученные разными алгоритмами результаты и выбрать лучший.

Выводы. В результате проведённой работы разработан алгоритм, способный извлекать смысловые понятия из секции переводов открытого словаря Wiktionary для выбранных групп языков, чтобы впоследствии расширять используемую семантическую сеть новыми словами и связями между ними, добавив ей языковое разнообразие. Готовая семантическая сеть может использоваться как базовое хранилище данных на разных языках, а также при

добавлении возможности корректно определять соответствие конкретных словоформ данная сеть может использоваться для переводов статей между языками. А полученный алгоритм распознавания данных внутри транслингвальной секции открытого словаря и построения связей между смысловыми понятиями будет использоваться для дальнейшего расширения семантической сети.

Редькина И.В. (автор) Подпись

Клименков С.В. (научный руководитель) Подпись