УДК: 004.822

Авторы:
И. А. Морозов, И. В. Редькина, О. А. Соболева Университет ИТМО, Санкт-Петербург
Научный руководитель:
Письмак Алексей Евгеньевич, Университет ИТМО, Санкт-Петербург

Тезис доклада:

Повышения качественных характеристик семантической сети

В докладе рассматривается правило повышения связности семантической сети за счет импорта связей из иноязычных источников, а также способ добавления транслингвальных данных (форм слов) из английского и немецкого викисловарей.

В викисловаре более миллиона статей на русском языке, которые имеют смысловые значения. Между ними есть связи, такие как: гипонимия, синонимия, антонимия и др. Структура словаря такова, что связи указывают на лексему, а не смысловые значения. Для повышения качественных характеристик семантической сети возможен способ восстановления связей с помощью данных из двух словарей. Но на текущий момент было всего 2 алгоритма для решения этой задачи, которые восстанавливают примерно 15 000 связей для словарного среза из более чем 400 тысяч словарных статей.

Правило основывается на поиске аналогичных связей из Wiktionary в RuThes. Особенностью тезауруса RuThes является наличие в имени концепта записанных в скобках уточнений, в результате чего имена концептов часто не совпадают с лексемами викисловаря, что затрудняет поиск одинаковых связей.

Разработанный подход основывается на дополнении данными смысловых значений из семантической сети данными из идентичных смысловых значений из иноязычных словарных срезов. Поиск эквивалентных смысловых узлов осуществляется на основе топологии: проверяются уже существующие связи по лемме смыслового узла и типу семантической связи. После нахождения нужного смыслового значения производится добавление словоформ этого понятия, которые являются переводами его лексемы на немецкий и английский языки, а также осуществляется добавление новых связей.

Так как Wiktionary редактируемый пользователями словарь, он может содержать ошибки. Поэтому не исключена погрешность совместимости семантических связей одних и тех же словарных узлов иноязычных словарей. В связи с чем был выбран словарный источник, где тип семантической связи имеет большую частоту совпадений. Стоит заметить, что итоговый артефакт, полученный в результате применения предлагаемого подхода, содержит небольшой процент погрешностей.