УДК 004.912, 004.852, 004.822

СОВМЕСТНОЕ ИСПОЛЬЗОВАНИЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ И СЕМАНТИЧЕСКИХ СЕТЕЙ ДЛЯ РАЗРЕШЕНИЯ ЛЕКСИЧЕСКОЙ НЕОДНОЗНАЧНОСТИ В ТЕКСТЕ НА РУССКОМ ЯЗЫКЕ

Зубань Д.А. (Университет ИТМО), **Яркеев А.С.** (Университет ИТМО)
Научный руководитель – старший преподаватель, Клименков С.В. (Университет ИТМО**)**

В работе рассматривается актуальная задача обработки естественного языка – разрешение лексической неоднозначности. Рассматриваются современные подходы к решению данной задачи и указываются их недостатки в контексте ее решения для русского языка. Предлагается подход, основанный на использовании русскоязычной нейросетевой модели глубокого обучения ruBERT совместно с семантической сетью, построенной на основе русскоязычных тезаурусов – Викисловаря и РуТеза.

Процесс извлечения информации превращает неструктурированную информацию, содержащуюся в тексте, в структурированные данные, и является неотъемлемой частью разработки различных интеллектуальных систем, например, рекомендательных систем, интеллектуального поиска, чат-ботов. Большую роль при построении таких систем играет разрешение лексической неоднозначности слов, которые в зависимости от заданного контекста могут иметь несколько значений, потенциально не связанных между собой.
Широко используемые статические модели векторных представлений слов, такие как Word2vec и GloVe, не могут отразить эту динамическую семантическую природу. Как способ устранения этого ограничения, были созданы контекстуализированные вложения слов – динамические векторные представления, адаптирующиеся в зависимости от контекста. Так, завоевавшая популярность модель ELMo использует двунаправленную модель долгой краткосрочной памяти biLSTM для анализа контекста слов и последующего построения их векторных представлений.

Контекстуализированные вложения слов являются неотъемлемым компонентом наиболее эффективных в настоящий момент глубинных нейросетевых моделей для обработки естественного языка, таких как GPT-3 и BERT. Однако данные модели относятся к классу трансферного обучения и предназначены для решения широкого спектра задач по обработке естественного языка, вследствие чего в частном случае они требуют тонкой настройки и показывают худшие результаты, чем их модификации для конкретной практической задачи. Оценка подобных моделей производится с помощью тестовых заданий общей задачи General Language Understanding, представленных на сайте проекта-бенчмарка SuperGLUE. Так, на наборе данных WiC (word in context), предназначенном для тестирования систем разрешения задачи лексической неоднозначности, точность описанных выше англоязычных моделей не превосходит 79%. Несмотря на то, что SuperGLUE предоставляет наборы данных для тестирования языковых моделей на различных языках, русскоязычной версии в настоящий момент не существует.

Проект russianSuperGlue является адаптацией описанного выше бенчмарка для русского языка, разработанной компанией «Сбербанк», которая также представила предобученную модель ruBERT. RussianSuperGlue включает набор данных RUSSE, являющийся русскоязычной версией WiC. Наиболее эффективной траекторией разрешения лексической неоднозначности является дополнительное использование баз знаний, поскольку процесс обучения нейросети не может гарантировать, что модель учтет все существующие значения рассматриваемого слова. Однако, как сами признают авторы, на сегодняшний день удалость адаптировать только модели, не использующие тезаурусы и базы знаний, так как на русском языке отсутствует настолько же качественный источник данных, как англоязычная лексическая база WordNet.

В качестве одного из путей решения рассматриваемой проблемы авторы предлагают использовать нейросетевую модель ruBERT совместно с семантической сетью, построенной на основе русскоязычных тезаурусов Викисловарь и РуТез. Реализация данного подхода призвана восполнить пробел в существующих алгоритмах для решения задачи устранения лексической неоднозначности на русском языке.