САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

образовательная программа магистратуры "Прикладная и экспериментальная лингвистика"

РЕЦЕНЗИЯ

На выпускную квалификационную работу студентки Рощиной Наталии Юрьевны, выполненную на тему: «АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ПАРАФРАЗОВ В РУССКОЯЗЫЧНОМ КОРПУСЕ ТЕКСТОВ»

Данная работа посвящена оценке эффективности методов дистрибутивной семантики при распознавании парафразов в коротких русскоязычных текстах. Актуальность работы не вызывает сомнений, поскольку распознавание парафразов является одной из базовых задач автоматической обработки текстов. Модули распознавания парафразов используются в машинном переводе и при оценке его качества, в системах автоматического реферирования, диалоговых системах, информационном поиске, для выявления плагиата, при кластеризации и категоризации текстов.

Работа состоит из введения, трёх глав, заключения, библиографии, а также четырех приложений. В первой главе рассматривается парафраз как лингвистическое явление. Раскрывается содержание понятия «парафраз», приводится история его возникновения, существующие классификации парафразов, а также проведенный автором эмпирический анализ языковых средств парафразов в корпусе ParaPhraser.

Во второй главе автором работы осуществлен обзор методов автоматического распознавания парафразов. Рассмотрены т.н. «поверхностные» признаки, признаки, моделирующие глубинную структуру предложения, а также семантические модели с использованием векторного пространства. Также автор работы аргументирует выбор порогового подхода к принятию решения при автоматической классификации пар предложений на парафразы и не-парафразы.

В третьей главе описан проведенный автором эксперимент по автоматическому распознаванию парафразов. Эксперимент проходил в три этапа: 1) оценка близости проверяемых пар заголовков в векторном пространстве с помощью алгоритмов Word2Vec; 2) подбор оптимальных пороговых значений для каждой построенной модели; 3) анализ ошибок распознавания.

В качестве достоинств данной работы можно отметить ее многоаспектность и значительный объем. Исследование, выполненное Н.Ю. Рощиной, отличается самостоятельностью и применением актуальных методов решения поставленных задач. Автор работы демонстрирует хорошее понимание проблематики своего исследования, уверенно владеет материалом и терминологическим аппаратом.

Вместе с тем, в работе были выявлены некоторые недочеты. Ниже приведены критические замечания и вопросы к автору:

1. На стр. 6 постулируется новизна исследования, связанная «с тем, что нам удалось подтвердить целесообразность использования векторных представлений (embeddings) парафразов, генерируемых с помощью нейросетевых моделей Word2Vec». Однако существуют широко цитируемые научные работы, посвященные использованию распределенных представлений (эмбеддингов) в задаче определения парафразов, например:

1) Socher, R., Huang, E. H., Pennin, J., Manning, C. D., & Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In Advances in neural information processing systems (pp. 801-809).
2) Kenter, T., & De Rijke, M. (2015, October). Short text similarity with word embeddings. In Proceedings of the 24th ACM international on conference on information and knowledge management (pp. 1411-1420). ACM.

и другие. Может быть, автором имелась в виду новизна с точки зрения применения данных методов к русскоязычному материалу?

2. Ошибка в Таблице 1 на странице 21: неверно посчитан процент от общего числа пар заголовков для 14 пар с количеством трансформаций, равным 6.
3. Вызывает сомнение следующее утверждение автора работы: «Весьма популярным преобразованием на уровне синтаксиса оказалась **смена модели управления глагола** (55 вхождений) или существительного (18 вхождений). В качестве примера можно привести пару №90: *Грецию пригласили – Греции предложили.*» (стр. 26) Выделенное автором работы словосочетание, вероятно, должно предполагать изменение лишь модели управления, но не самого глагола (ср. пару: *ударить в стену – ударить [чем-либо] об стену*). Однако в иллюстрации, приведенной автором работы, меняется не только модель управления, но и сам глагол (*пригласили – предложили*). Можно ли в таком случае говорить об изменении модели управления отдельно (независимо?) от лексической замены? Существуют ли в рассмотренном автором материале примеры замены лишь модели управления глагола, когда сам глагол сохраняется?
4. В подразделе «2.1.1 Поверхностные признаки» данные признаки лишь перечисляются, но не приводится анализ их эффективности относительно друг друга и признаков иных типов.
5. Нумерация подразделов раздела 2.1 неверная: 2.1.1, 2.1.1 и 2.3. В главе 3 нумерация разделов начинается не с 3.1, как должно быть, а с 3.3.
6. В работе присутствуют стилистические погрешности. Например: «алгоритм наивного Байеса» (стр. 38) – это жаргонизм, ведь наивен не Байес, а допущение о независимости признаков, которые репрезентируют объект. Более терминологичное название алгоритма – «наивный байесовский классификатор». Ещё один пример: «Следующим шагом из обрабатываемого документа были удалены слова служебных частей речи» (стр. 39). В данном случае можно заменить *шагом… были удалены*

на *шагом… было удаление*.

7. В главе 3 описывается, в числе прочего, калибровка параметров семантической модели, а именно пороговых значений семантической близости пар заголовков. Пороговые значения должны отделять парафразы от не-парафразов. Однако здесь есть два замечания. Во-первых, поскольку происходит подбор (калибровка) параметров, не следует называть выборку, по которой эти пороговые значения сравниваются между собой с точки зрения точности классификации, тестовой выборкой. Скорее, это валидационная выборка (т.е. не test set, a validation set). Тестовая выборка не должна использоваться многократно для подбора оптимальных параметров, она используется лишь тогда, когда система уже «заморожена» и дальнейшие изменения в нее не вносятся. Для подбора пороговых значений можно было бы использовать десятикратную кросс-валидацию (10-fold cross-validation). Во-вторых, в связи с вышеизложенным, результаты эксперимента, проведенного автором, нельзя напрямую сравнивать с результатами участников дорожки по распознаванию парафразов AINL-2016, поскольку последние были получены на закрытой **тестовой** выборке, которая только после окончания соревнования стала частью открытого корпуса.
8. На сайте проекта ParaPhraser приведена статистика, согласно которой в корпусе уже 11 тысяч размеченных пар предложений (http://paraphraser.ru/scorer/stat/). Почему автором работы было использовано лишь около тысячи пар предложений? На большем объеме данных результаты могли бы оказаться лучше.
9. Представляется, что в рамках данной работы следовало бы использовать также более современные, чем *word2vec*, распределенные представления (*fasttext*, *doc2vec* и др.).

В целом, работа выполнена на высоком профессиональном уровне и представляет собой законченное, полноценное исследование. Магистерская диссертация соответствует основным требованиям, предъявляемым к квалификационным работами такого уровня, автор заслуживает присвоения квалификации магистра по специальности «Прикладная и математическая лингвистика». Рекомендуемая оценка ВКР: «отлично».

Рецензент:

Малафеев А.Ю.
Канд. филол. н.,
Доцент департамента прикладной лингвистики и иностранных языков
НИУ ВШЭ

«29» мая 2018 г.

ПОДПИСЬ А.Ю. Малафеева
ЗАВЕРЕНА НАЧАЛЬНИКОМ
ОТДЕЛА КАДРОВ
Н.А. ЕРМОЛИНОЙ

(подпись)