**УДК 004.89**

# Использование машинного обучения для автоматизированного расширения формальных грамматик, использующихся для классификации текстов в диалоговых системах

**Ионов Д.П.**, (Университет ИТМО; Dasha.AI), г. Санкт-Петербург
**Научный руководитель – Ульянцев В.И.**, к.т.н., научный сотрудник ФИТиП, Университет ИТМО, г. Санкт-Петербург

Данная работа посвящена разработке нового подхода для решения задачи классификации текстов, основанного на существующих решениях. Рассматриваются возможности улучшения решения, использующего формальные грамматики, с помощью решения, использующего машинное обучение.

**Введение.** Задача классификации текстов является одной из задач, которые решаются в системах, взаимодействующих с пользователем на естественном языке. Существует два основных подхода для решения этой задачи. Первый подход заключается в написании ряда правил, так называемой грамматики, по которой текст можно отнести к тому или иному классу. Этот подход обладает высокой точностью, так как специалист вручную пишет каждую грамматику, но он плохо масштабируем, потому что на каждое изменение в классе текста, либо на добавление нового класса текстов в структуру, специалист должен вручную изменить/составить грамматику. Второй подход заключается в использовании методов машинного обучения. В этом случае критерий принятия решения о выборе класса текста вычисляется автоматически из обучающей выборки. Однако для хорошей работы этого метода необходима хорошая обучающая выборка.

**Основная часть.** Написание формальной грамматики, которая хорошо классифицирует тексты является весьма трудоёмкой задачей и требует больших затрат времени от лингвиста. В частности, ему приходится записывать множество слов, которые имеют похожий смысл. В упрощении этого процесса и состоит основная идея нового подхода.

Будем автоматически расширять формальные грамматики, добавляя к некоторым словам их семантически близкие аналоги. Таким образом лингвисту не нужно будет перечислять большое количество похожих по смыслу слов, достаточно будет написать одно либо несколько слов, а остальные слова будут добавлены при расширении. Для нахождения семантически близких слов воспользуемся моделью векторного представления слов word2vec. Чтобы улучшить подборку семантически близких слов и убрать из неё шум, будем использовать различные методы ансамблирования моделей и контекстуализированные модели, обученные с использованием алгоритма ELMo. Также для лучшей работы подхода для каких-либо конкретных систем, можно дообучить модели на специальных наборах данных.

**Выводы.** Разработанный подход автоматического расширения грамматик позволяет уменьшить время необходимое для написания грамматики и упростить работу лингвистов. Также этот подход можно интегрировать в парсеры грамматик, чтобы не усложнять процесс парсинга необходимостью предварительной обработки грамматики.

Ионов Д.П. (автор) Подпись

Ульянцев В.И. (научный руководитель) Подпись