УДК 004.89

ПОДХОДЫ К ОБЪЕДИНЕНИЮ МЕТОДОВ НА ОСНОВЕ ПРАВИЛ И МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ ТЕКСТОВ

Мамаев Н. К. (Университет ИТМО,ООО «ЦРТ-инновации»),
Лизунова И. А. (Университет ИТМО, ООО «ЦРТ-инновации»),
Маслюхин С. М. (ООО «ЦРТ-инновации»),
Ховричев М. А. (ООО «ЦРТ»)
Научный руководитель – к. т. н. Махныткина О. В.
(Университет ИТМО)

В данной работе рассматриваются методы автоматической обработки естественного языка, которые позволяют значительно ускорять процессы обслуживания в современном бизнесе. Одна из основных проблем в этой сфере – классификация запросов. Поскольку обеспечить достаточное количество данных для обучения классификатора на основе машинного обучения не всегда возможно, мы предлагаем два эвристических способа его объединения с классификатором на основе лингвистических правил, составляемых экспертами.

Работа выполняется при финансовой поддержке Минобрнауки России, Соглашение 14.575.21.0178 (Уникальный идентификатор проекта: RFMEFI57518X0178)

Методы автоматической обработки естественного языка широко используются для оптимизации процессов работы с клиентом в предприятиях, оказывающих услуги, а в отдельных случаях – и в производственных. Крупные предприятия выделяют значительные средства на сбор и обработку данных, необходимых для обеспечения высокоэффективной работы решений для обработки языка, основанных на машинном обучении. Однако в случаях, когда собираемых данных недостаточно, имеет смысл прибегнуть к менее автоматизированным подходам – основанным на лингвистических правилах, составленных вручную экспертами.

Немедленной проблемой, возникающей при переходе к автоматизации обработки запросов, является сепарация и переадресация запросов в зависимости от тематики, к которой они относятся – иными словами, классификация запросов. В этой работе мы рассматриваем автоматическую диалоговую систему, работающую с текстовыми запросами на естественном языке, одной из функций которой является классификация запросов по тематикам из предопределённого набора. Критическим ограничением, не позволившим использовать только решение на правилах или решение с применением машинного обучения, стало отсутствие достаточного количества примеров для одного подмножества тематик, и невозможностью обеспечить достаточное количество правил для другого. Таким образом, было предложено разработать гибридную систему-классификатор, агрегирующую предсказания от пары классификаторов.

Для классификатора, в основе которого лежат составленные экспертами правила, коэффициент уверенности для пары запрос-тематика (число, характеризующее количественную вероятность того, что запрос корректно отнесён к некоторой тематике), вычисляется на основе сопоставления запроса с набором регулярных выражений, соответствующих данной тематике, с использованием специальных формул. В классификаторе на основе машинного обучения, использующего свёрточную нейронную сеть, в качестве коэффициента уверенности берётся число с выхода последнего слоя нейронной сети.

Объединить классификаторы было предложено на уровне результатов сравнения, используя линейный подход – иными словами, суммировать линейно взвешенные коэффициенты уверенности от обоих классификаторов. Мы провели исследование, в котором сравнили два подхода к определению коэффициентов взвешивания:

1) наивный подбор коэффициентов линейным поиском,

2) разработка формулы для расчёта пары коэффициентов для каждого класса в зависимости от количества правил или обучающих примеров для этого класса. Формула характеризует оценку того, насколько правдоподобно каждая из моделей описывает некоторый класс.

Тестирование предложенного решения при разнообразных условиях (количество классов, правил и примеров) позволило прийти к выводу, что подход, использующий формулу, позволяет получить более высокое качество классификации (согласно метрике micro f-score) в ряде случаев, а в остальных – сравнимое с качеством, которое получается при использовании первого подхода.