УДК 004.89

Настройка автоматической проверки правописания и ранжирования документов для построения диалоговой системы на естественном языке

Н. К. Мамаев
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – И. А. Черных
(Университет ИТМО, г. Санкт-Петербург)

Исследования выполнены за счет стартового финансирования университета ИТМО в рамках НИР № 618278 «Синтез эмоциональной речи на основе генеративных состязательных сетей»

Разработка автоматических диалоговых систем сегодня пользуется большим спросом за счёт потенциальной экономии времени работников служб поддержки, а также возникающей возможности масштабирования работы подобных служб. В частности, этим требованиям удовлетворяют целеориентированные автоматические диалоговые системы, помогающие клиенту достичь совершить какое-то действие: найти подходящий продукт, совершить бронирование номера в отеле или получить помощь в форме ответа на вопрос. Среди подходов к построению целеориентированных диалоговых систем выделяют, во-первых, т.н. pipeline-подход, делящий процедуру обработки запроса на четыре этапа – понимание запроса, обновление состояния диалога, выбор действия и генерацию ответа [1]. Также активно развиваются end-to-end подходы, которые осуществляются с помощью генеративных нейросетей и подразумевает монолитность системы в некотором общем смысле.

End-to-end подход, получивший распространение благодаря росту вычислительных мощностей и объёмов данных, является более актуальным, поскольку позволяет достичь высокой эффективности при выдаче ответов на запросы без необходимости вручную обрабатывать большие массивы текстов и составлять громоздкие сценарии.

В качестве примера end-to-end диалоговой системы можно привести, например, нейросетевую архитектуру R-Net, осуществляющую генерацию ответа на вопрос по контексту. Опираясь на текстовый отрывок, данная система «учится» выбирать из него промежуток, подходящий в качестве ответа [2]. Механизм векторного представления, используемый в этой системе, не является устойчивым к шуму, и в связи с этим целесообразно производить предварительное исправление ошибок написания в вопросе. Наконец, при подключении к данной системе компонента, ранжирующего документы по смысловой близости к вопросу, получим систему, аналогичную DrQA, способную выполнять предыдущую задачу на множестве документов [3].

Мы исследовали способы увеличить точность генерации ответов вопросно-ответной системы архитектуры DrQA с помощью сравнительного анализа результатов работы различных инструментов **автоматического исправления ошибок правописания** и **ранжирования документов**. Среди инструментов автоматического исправления ошибок правописания мы исследовали алгоритм, основанный на нечётком поиске с использованием расстояния редактирования Damerau-Levenshtein; алгоритм, основанный на подходе noisy channel [4]; а также ПО LanguageTool 3.3, основанное на детектирующих паттернах. Вычисление метрик проводилось путём сравнения типов слов в автоматически исправленном документе с типами слов во вручную исправленном документе. На основании результатов можно сделать следующий вывод: все три инструмента продемонстрировали схожую эффективность работы, однако алгоритм #3 позволяет производить обучение и этим выгодно отличается от алгоритма #2 и инструмента #1. Результаты анализа можно видеть в Таблице 1.

Мы также исследовали способы увеличить точность генерации ответов, анализируя подходы к ранжированию документов, использующие векторные представления TF-IDF и word2vec. При сопоставлении переформулировок одних и тех же вопросов показатель

точности достиг 0.303 для ранжировщика на основе TF-IDF и 0.206 для ранжировщика на основе word2vec (режим – skipgram, размерность векторных представлений – 100, активационная функция последнего слоя – иерархический софтмакс, размер обучающей выборки – около 6.1 млн. слов). Несмотря на то, что точность работы первого ранжировщика на использованных данных выше, второй ранжировщик выгодно отличается тем, что модель word2vec потенциально может быть оптимизирована за счёт увеличения обучающей выборки, использования при обучении данных узкой тематики и оптимизации параметров обучения.

Таблица 1 - Результаты анализа эффективности методов исправления ошибок правописания

	Точность	**Полнота**	**F-мера**
#1: LanguageTool 3.3	0.232	0.491	0.315
#2: алгоритм с исп. нечёткого поиска	0.232	0.491	0.315
#3: алгоритм с исп. подхода noisy channel	0.211	0.548	0.305

1. H. Chen, X. Liu, D. Yin, J. Tang. A Survey on Dialogue Systems: Recent Advances and New Frontiers // Режим доступа: https://arxiv.org/abs/1711.01731 (дата доступа: 19.01.2019)
2. Natural Language Computing Group, Microsoft Research Asia. R-Net: Machine Reading Comprehension With Self-Matching Networks // Режим доступа: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/05/r-net.pdf (дата доступа: 20.01.2019)
3. D. Chen, A. Fisch, J. Weston, A. Bordes. Reading Wikipedia to Answer Open-Domain Questions // Режим доступа: https://arxiv.org/pdf/1704.00051 (дата доступа: 20.01.2019)
4. E. Brill, R. C. Moore. An Improved Error Model for Noisy Channel Spelling Correction // Режим доступа: http://www.aclweb.org/anthology/P00-1037 (дата доступа: 19.01.2019)