

АНАЛИЗ СВОЙСТВ ПРЕДСТАВЛЕНИЙ СЛОВНЫХ СЕТЕЙ В ЗАДАЧЕ ИНДЕКСИРОВАННОГО ПОИСКА ПО РЕЧЕВЫМ ДАННЫМ

Петров О. Е. (ООО “ЦРТ”), Кабаров В. И. (Университет ИТМО)
Научный руководитель – д.т.н. Матвеев Ю. Н.
(Университет ИТМО)

Введение. Индексированный поиск является эффективным методом поиска информации в больших объемах данных. Он основан на специальных структурах данных — поисковых индексах, позволяющих быстро искать нужную информацию. Основная задача индексированного поиска заключается в том, чтобы обеспечить быстрый доступ к данным и оптимизировать процесс поиска.

Индексированный поиск используется в таких областях, как поисковые системы, базы данных, файловые системы, анализ данных, а также в программных приложениях для быстрого доступа к информации. В качестве документов могут выступать как текстовые данные, так и медиаданные — аудио или видео. Одним из примеров является речевая аналитика — процесс анализа и извлечения информации из аудиозаписей разговоров, записанных в контакт-центрах или офисах продаж для контроля качества обслуживания.

Основная часть. Для построения поисковых индексов по речевым данным в рамках работы использовалась система автоматического распознавания речи. В процессе работы лингвистического декодера порождается множество гипотез, объединяющихся в одну структуру — словную сеть, которая представляет собой направленный словный граф без циклов. Использование индексов, построенных на основе словных сетей, позволило повысить полноту результата за счет возможности обнаружения слов, распознанных с низким уровнем достоверности и не попавших в лучший результат распознавания [1]. Количество параллельных состояний и переходов из каждого состояния словной сети может быть произвольным. Такая нерегулярная структура плохо поддается индексации, поэтому использовались алгоритмы дополнительной обработки, позволяющие строить представления словных сетей с регулярной структурой. На основе словных сетей также могут строиться и векторные представления, учитывающие не только контекст, но и акустическую близость слов [2].

В рамках данной работы рассмотрены различные представления словных сетей и проведена оценка точности поиска на различных тестовых выборках, включающих в себя как телефонные данные, так и данные, записанные на удаленный микрофон в сложных акустических условиях. Рассмотрены такие структуры, как сети спутывания [3], основанные на алгоритме декодирования с минимизацией байесовского риска для построения сетей спутывания с минимальным значением пословной ошибки (Word Error Rate, WER) [4], а также расширенные сети спутывания, основанные на двойном проходе алгоритма прямого-обратного хода по словной сети с объединением узлов с одинаковыми временными метками.

Выводы. В рамках работы создан прототип, реализующий алгоритм поиска по индексу на основе различных представлений словных сетей, а также были получены показатели точности работы системы, подтверждающие возможность ее использования в задачах речевой аналитики для аудио записей из контакт-центров и офисов продаж.

Список использованных источников:

1. Petrov O. The architecture of a system for full-text search by speech data based on a global search index // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. – 2021. – № 5(21). – С. 791–794.
2. Gurunath Shivakumar P., Georgiou P., Narayanan S. Confusion2Vec 2.0: Enriching ambiguous spoken language representations with subwords // PLoS ONE 17(3): e0264488 – 2022.

3. Mangu, L., Brill, E., Stolcke, A. Finding consensus in speech recognition: word error minimization and other applications of confusion networks // *Computer Speech & Language* 14.4 – 2000. – C. 373–400.
4. Xu, H., Povey, D., Mangu, L., Zhu, J. Minimum bayes risk decoding and system combination based on a recursion for edit distance // *Computer Speech & Language* 25.4 – 2011. – C. 802–828.