Секция «Программирование и компьютерные науки»

Автоматический перевод аббревиатур в фонемную запись в системе синтеза речи

Девбунова Вилиана Олеговна
Студент (бакалавр)
Московский государственный университет имени М.В.Ломоносова, Факультет вычислительной математики и кибернетики, Кафедра математических методов прогнозирования, Москва, Россия
E-mail: devbunova_vilya@mail.ru

Синтез речи — формирование речевого сигнала по печатному тексту. Самым частым приложением задачи синтеза речи является голосовой помощник. Если ассистент плохо выговаривает слова, имеет металлический голос, неестественную интонацию или игнорирует знаки препинания, его речь будет звучать странно. А значит, пользователи не захотят с ним общаться. Поэтому компании развивают технологии синтеза и ищут способы повысить качество.

В данной работе рассматривается задача автоматического перевода аббревиатур в фонемную запись в системе синтеза речи. Для решения этой задачи предложен новый метод, в основе которого лежит модификация пайплайна [1] синтеза с помощью блока обработки аббревиатур (см. Рис. 1.). Блок состоит из двух подзадач: выделения аббревиатур из входного текста и построения отдельной модели перевода отобранных слов в фонемы (grapheme-to-phoneme или G2P). Для решения каждой подзадачи исходной задачи предложен метод и программная реализация, проведены вычислительные эксперименты.

Первую подзадачу детекции аббревиатур в тексте можно отнести к более широкой задаче сегментации и разметки последовательности (sequence segmentation and labeling problem). К таким задачам относятся, например,

- определение частей речи слов во фразе (POS-taging) [1]
- нахождение именованных сущностей (Named Entity Recognition, NER) [2]
- этап обнаружения слов, нуждающихся в раскрытии [3]

В этих задачах, как и в нашей, входную последовательность из n слов необходимо отобразить в последовательность той же длины n из алфавита меток L. Для нашей подзадачи $L = \{0, 1\}$. Для решения данной подзадачи предлагается опробовать следующие архитектуры:

- детектор без контекста, по графемной записи
- детектор по эмбеддингам
- детектор по смеси эмбеддингов и графемной записи

Перейдем к рассмотрению второй подзадачи. Модели перевода слов в фонемы бывают пословные и переводящие несколько слов (обычно это фраза или предложение). Преимущество фразовых G2P заключается в том, что они получают на вход слова с контекстом, благодаря которому модель может разрешать проблему омографии [2] сама. К сожалению,

[1] Пайплайн — это последовательность обрабатывающих элементов (например, потоков или функций), которые выполняются один за другим, передавая данные с выхода предыдущего элемента на вход следующему.

[2] Омографы mdash; слова, которые совпадают в написании, но различаются в произношении. В русском языке чаще всего из-за различий в ударении.

качественных открытых датасетов для решения этой задачи нет, а ручное составление такого датасета очень трудоемкий и времязатратный процесс.

В общем случае для решения задачи G2P используют сложные архитектуры с миллионами параметров [5], [4], которые для обучения требуют больших тренировочных выборок. Но мы имеем частный случай применения этой задачи с ограниченным количеством данных, поэтому используя их специфику мы можем построить модель, основанную на правилах.

Одной из самых нетривиальных вещей в данной задаче является сбор данных. Для русского языка на настоящий момент еще не создано открытого корпуса с разметкой аббревиатур, поэтому для обучения и тестирования моделей собрали и системтизировали информацию по аббревиатурам с сайта https://ru.wiktionary.org/. Далее отобрали слова с меткой "Аббревиатура"и их произношение. Аналогично собрали и структурировали текст русской википедии – https://ru.wikipedia.org/ для извлечения предложений с аббревиатурами и составления соответствующих им последовательностей меток. Но таким образом мы смогли собрать только положительные примеры для задачи детекции. Добавление негативных примеров в набор данных производили методом случайного выбора слова в предложении и проверкой, что его начальная форма есть в словаре и не является аббревиатурой. Таким образом была создана автоматизированная система составления датасета для обучения модели.

В ходе работы был создан испытательный стенд для разработки блока обработки аббревиатур путем проверки гипотез, был собран первый в мире датасет русских аббревиатур. В данной работе предложено разбиение задачи на подзадачи и предложены методы для решения каждой из них.

В дальнейших исследованиях планируется составление правил чтения аббревиатур и написание G2P на их основе. Проверка гипотез с помощью созданного стенда. Сравнение полученных моделей с уже существующими.

Источники и литература

1) Conditional random fields: Probabilistic models for segmenting and labeling sequence data / Lafferty J., McCallum A. and Pereira F. //Proceedings of the 18th International Conference on Machine Learning, Williamstown, Massachusetts, 2001. — Pp. 282–289.
2) Named Entity Recognition with Bidirectional LSTM-CNNs / Jason P.C. Chiu, Eric Nichols – 2016
3) A Unified Transformer-based Framework for Duplex Text Normalization / Tuan Manh Lai, Yang Zhang, Evelina Bakhturina, Boris Ginsburg, Heng Ji – 2021
4) SoundChoice: Grapheme-to-Phoneme Models with Semantic Disambiguation / Artem Ploujnikov, Mirco Ravanelli – 2022
5) Transformer based Grapheme-to-Phoneme Conversion / Sevinj Yolchuyeva, Géza Németh, Bálint Gyires-Tóth – 2020

Иллюстрации

Рис. : 1. Упрощенная схема модифицированного пайплайна.