Что с нами будет? Письма ученых о самом важном

Это письмо о компьютерном зрении и исследовании памяти

Привет!

В сегодняшнем письме речь пойдет о том, как в Европейском университете исследуют надписи на еврейских надгробиях с помощью компьютерного зрения. Участники программы дополнительного образования ПАНДАН — совместного проекта университета и «Яндекса» — Юлия Аматуни, Дмитрий Серебренников и Татьяна Ткачева рассказывают, зачем изучать надписи на еврейских надгробиях с помощью нейросетей и почему в эпитафиях второй половины XIX века, расшифрованных с помощью компьютерного зрения, возникают слова «сервис» и «уборка».

Зачем учить нейронные сети распознавать надписи на надгробиях?

Идея этого проекта возникла в разговоре с центром «Петербургская иудаика». В его распоряжении находился архив фотографий надгробий из коллекции художника Давида Гобермана. Он снимал еврейские кладбища преимущественно в Галиции и Буковине в 50–60-х годах XX века. На своих фотографиях Гоберман делал акцент на эстетической составляющей — его больше интересовал орнамент на надгробиях, а не сам текст эпитафии. Однако, несмотря на то, что снимки делались ради искусства, на большинстве из них надписи более или менее видны.

Язык, на котором пишут эпитафии, не похож на современный иврит — он очень клишированный и включает в себя элементы древнего, средневекового иврита и арамейского. Кроме того, в эпитафиях широко используются аббревиатуры. Мы пришли к выводу, что научить нейронную

сеть корректно распознавать такой текст — серьезная задача, в результате работы над которой можно создать новый инструмент для исследований и упростить процесс расшифровки надписей.

Для чего это нужно? Наша гипотеза заключается в том, что с помощью подобного инструмента будет возможно ставить более глубокие исследовательские вопросы и задачи. Например, если мы говорим о еврейских надгробиях, по ним можно отследить, когда и как происходила русификация поселений, потому что в какой-то момент на плитах начинают появляться надписи на русском языке. То есть благодаря таким инструментам исследователи смогут работать с разнообразными наборами данных.

Как технически происходила расшифровка надписей?

Мы разделили работу на два этапа: нейросети и автоматическое распознавание символов, OCR (optical character recognition). На первом этапе нашей задачей было получить из необработанной фотографии кусок текста с надгробия, максимально пригодный для автоматического распознавания.

Всего у нас было примерно 10 тысяч снимков — кроме коллекции Давида Гобермана фотографии для исследования нам предоставил центр «Сэфер». Сложность была в том, что данные оказались слишком разные. Гоберман подходил к вопросу как искусствовед и снимал только орнамент, а исследователи центра «Сэфер» работали более системно и фотографировали все надгробия подряд и так, чтобы их было видно полностью. Исходные фотографии были очень разного качества, нам приходилось учитывать, что:

- изображения могут быть размытыми/яркими/засвеченными/темными;
- на снимке может быть несколько надгробий;
- эпитафии могут быть нечитаемыми;
- надгробий на фото может не быть.

Для того чтобы распознать только ту часть, где есть надгробие с надписью, мы использовали две нейросети. Первая определяла, возможен или невозможен анализ этого изображения, есть ли на нем что-то, что может быть нам

интересно, и пригодного ли оно качества. А вторая выделяла на снимке текстовый фрагмент.

Чтобы улучшить результаты распознавания с помощью OCR, мы последовательно проверяли разные техники для улучшения качества изображения. В итоге наиболее эффективными оказались комбинации сразу нескольких инструментов.

Собственной системы OCR у нас не было, так как это очень сложная и дорогая технология. Поэтому, чтобы распознавать картинку и выводить с нее текст, мы использовали готовые инструменты от «Яндекса» и Google.

Какие ошибки допускали алгоритмы?

На этапе автоматического распознавания возникла следующая сложность: системы OCR, которые мы применяли, обучены на массивах данных с современными текстами на иврите. Из-за этого, когда мы распознавали с помощью них надписи на надгробиях и пытались их переводить, довольно часто алгоритмы искали в текстах тот смысл, которого там не было.

Как уже упоминалось выше, особенность эпитафий в том, что часто в них используются сокращения и символы, которые в современном языке, известном системе OCR, не используются. Есть хороший пример из латыни — сокращение STTL (Sit tibi terra levis, «пусть земля тебе будет пухом»). Широко встречающийся аналог в еврейских эпитафиях — «Да будет душа его завязана в узле жизни» в виде аббревиатуры на иврите. Алгоритм «Яндекса» или Google, который будет распознавать этот текст, скорее всего подумает, что это ошибка и предложит свой вариант, какое-то понятное для него слово с похожим набором букв.

В результате, когда мы расшифровывали надписи на надгробиях, нам попадались слова «сервис» и «уборка дешево». Такие ошибки мы в шутку называли «перевод с AliExpress». И в итоге пришли к выводу, что нам нужна своя OCR, которая будет заточена именно под нашу языковую модель и обучена на специфических текстах.

Что еще удалось выяснить в ходе

проекта?

Мы обработали весь объем имеющихся данных и в результате поняли, как подходить к решению подобных задач, как работать с такими данными и в какую сторону двигаться дальше, чтобы улучшить результаты. Если работа будет продолжаться, мы сможем распространить существующие технологии и наработки на другие языки и построить модели для распознавания текста на надгробиях и других похожих задач.

Мы убедились, что на качество распознавания влияет большое количество факторов, и продолжаем сотрудничать с полевыми исследователями. Например, на XXVII Ежегодной международной конференции по иудаике, где мы выступали с докладом, участники секции взяли наши результаты в оборот: планируется составить рекомендации по съемке надгробий, которые в перспективе позволят более успешно применять нейросети в работе с такими снимками.

Почему этот проект важен для будущего?

Наш проект связан с научным исследованием памяти, семей и родословных. В идеале мы бы хотели автоматизировать расшифровку текстов эпитафий и их перевод на русский язык, а также сделать базу, с помощью которой люди могли бы разбираться в генеалогии. С еврейскими надгробиями это сложно, потому что на них почти нет фамилий — в основном имена. Но, как правило, если человек знает, где и когда жил его дед или бабушка, а также имя его или ее отца, уже можно примерно сопоставить имеющуюся информацию и найти корни.

Кажется, что это очень важное социальное измерение, так как многих могил, которые изображены на фотографиях Давида Гобермана, уже нет. Кладбища теряются и исчезают, а подобная фиксация и работа над созданием баз данных может облегчить людям поиск родных и семейных историй.

Кроме того, наша работа показывает, как компьютерные науки, и в частности методы статистики, машинное зрение и алгоритмы для обработки естественных языков, помогают решать исследовательские задачи в гуманитарных науках.

Всё чаще встречаются междисциплинарные команды, в которых искусствоведы, историки и антропологи работают вместе с дата-сайентистами и математиками. Такое взаимодействие позволяет применять нейросети для автоматизации рутинных задач, производить наблюдения гораздо быстрее и выводить новые гипотезы на стыке нескольких научных областей.

На этом всё.

Хороших выходных
