Интеллектуальная обработка текстовой информации для автоматической генерации заключения врача при цитологическом исследовании материала щитовидной железы

Дюльдин Евгений Владимирович
Факультет кибернетики и информационной безопасности, Москва
E-mail: Zhecos1@yandex.ru
Соавторы: Дюльдин Е.В., Боброва Е.В., Моисеенко О.И., Шифман Б.М.

В медицине с каждым годом становится все более явным, что классические подходы в обработке текстов уже не всегда способны достичь желаемых результатов. С тем, чтобы поддержать принятие решений при решении сложных задач в медицине и обеспечить разработку новых умных систем диагностики и лечения пациентов, все чаще применяются современные технологии обработки текста и глубокого обучения.

Одной из таких областей, где новые подходы особенно полезны, является генерация врачебных заключений и классификация по Bethesda, которые очень востребованы в цитологических исследованиях щитовидной железы. Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) предлагает шкалу классификации с шестью диагностическими категориями, каждая из которых имеет соответствующий риск злокачественного заболевания.

Целью данной работы является использование цифровых технологий и разработка умных систем диагностики для поддержки принятия решений в медицине, способных превзойти классические подходы и обеспечить более точную диагностику и лечение пациентов.

Исходные данные были представлены, как набор пар:

X = {описание; заключение}, где внутри заключения предположительно находилась метка Bethesda, общее число составляло 17000 примеров.

Предобработка данных состояла из удаления специальных символов в каждом из текстов, так как заполнитель пробелов и пустых пространств включал символ ‘%ro’. Следующий шаг разделение общего пространства примеров на три подкатегории:

1) Данные которые не нуждаются в дальнейшей очистке
2) Данные которые необходимо разделить на несколько значащих блоков
3) Данные которые после выделения метки Bethesda необходимо дополнительно обработать регулярными выражениями

После разделения данных на три подкатегории общее число примеров изменилось (Таблица 1).

Увеличение общего числа данных было вызвано тем, что часть заключений включало в себя несколько блоков описаний, что позволило отображать один пример {описания; заключения} в несколько примеров: {описание_i; заключение_i}.

Для решения задач классификации на основе меток Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) использовались подходы на основе технологий трансформеров, как BERT [1]. В сравнении были применены методы и архитектуры, описанные в работе [2], RoBERTa естественное продолжение идеи encoder архитектур. В качестве базовой модели, которая являлась отправной точкой в данной работе были применены методы на основе последовательных сетей, как LSTM [3], в цикл обучения были внесены незначительные модификации Batch Normalization [4] для обеспечения более быстрой сходимости в задаче классификации. Сравнение базовых подходов на основе методов LSTM [5] и полученных в ходе работы представлены в (Таблица 2).

Далее были применены методы доменной адаптации, которые позволяют, использовать слабосвязанные наборы данных с исходным набором примеров, что позволяет получить больший объём выборки, во время доменной адаптации проводилось дообучение исходной модели для минимизации сдвига от исходной области обучения и сохранения знаний о лингвистике языка. Последним этапом был сбор данных из сходных тематик и примешивание текущих текстов, как к дообучающим выборкам, так и для первичной тренировки, так как в этом случае мы хотим обучать модель разбираться в новой для модели доменной области, где мы независимо получаем прирост качества в классификационных задачах и генерационных задачах. Далее проведем полноценную аугментацию данных, для увеличения разнообразия примеров в корпусе текстов.

Аугментация является способом увеличить размер обучающей выборки, тем самым повысив качество модели машинного обучения. В задаче генерации медицинских заключений по их описанию аугментация является важным шагом, поскольку классы Bethesda не сбалансированы по своей частоте. В полученном нами датасете было очень много примеров, когда описание отсутствовало, и заключение присутствовало, т.о. можно было решить обратную задачу: генерировать описание по заключению. Для этого были рассмотрены языковые модели семейства encoder-decoder. Природа медицинских текстов довольно постоянна: используются повторяющиеся формулировки для конкретных случаев. Для уточнения информации при генерации описаний можно использовать н-граммные признаки-счетчики, или обучать несколько моделей под каждый класс Bethesda, указанный в заключении.

Адаптацию модели под конкретную метку Bethesda можно рассматривать как локальный TAPT (task adaptive pretraining task), в рамках которого модели настраивается на тематику подобласти, связанной с определенной меткой Bethesda.

Признаки-счетчики можно конкатенировать к скрытому состоянию из T5, полученный вектор подавать на вход линейному слою и софтмакс-функции для выбора следующего токена.

Информация, извлекаемая из признаков счетчиков может восприниматься как излишняя, если заранее провести TAPT для категорий Bethesda. Однако, внутри каждой категории также присутствует определенная вариативности внутри текстов, которая может уточняться с помощью нграммных признаков.

Подбор необходимых н-грамм для признаков можно проводить на небольших encoder-decoder моделях, в то время как самую методологию отрабатывать на модели побольше.

Полученные таким образом аугментации заметно качественнее, чем те, которые можно получить обучая одну модель для всего датасета заключений, или же не используя признаки-счетчики.

Источники и литература

1) Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.
2) Liu Y. et al. Roberta: A robustly optimized bert pretraining approach //arXiv preprint arXiv:1907.11692. – 2019.'
3) Staudemeyer R. C., Morris E. R. Understanding LSTM–a tutorial into long short-term memory recurrent neural networks //arXiv preprint arXiv:1909.09586. – 2019.
4) Ioffe S., Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift //International conference on machine learning. – pmlr, 2015. – С. 448-456.

5) Zhou C. et al. A C-LSTM neural network for text classification //arXiv preprint arXiv:1511.08630. – 2015.

Иллюстрации

Class (index)	Number of class entries	Bethesda entry percentage
1	3,409	12.43
2	17,799	69.42
3	1,05	3.83
4	2,16	7.88
5	1,076	3.92
6	952	3.47

Рис. : 1. Число данных и меток Bethesda

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
RNN + LSTM	80.7	81.5	79.9	80.7
CNN + 1D Conv	86.4	85.9	87.2	86.5
Hybrid Model	89.2	89.8	88.6	89.2

Рис. : 2. Сравнение моделей