

МЕТОДЫ КОРРЕКЦИИ ПОВЕДЕНИЯ ЯЗЫКОВЫХ МОДЕЛЕЙ В ЗАДАЧЕ АБСТРАКТИВНОГО РЕФЕРИРОВАНИЯ

Чернышев Даниил Иванович¹

1) МГУ им. М.В.Ломоносова, НИВЦ, e-mail: chdanorbis@yandex.ru

УДК: 004.912

Аннотация

Доклад посвящен проблеме неустойчивости процесса порождения рефератов нейросетевыми языковыми моделями абстрактного реферирования. В докладе будут рассмотрены причины поведенческих аномалий в работе лучших систем абстрактного аннотирования, а также представлены новые методы коррекции поведения моделей, позволяющие повысить надежность существующих систем.

Ключевые слова:

языковое моделирование, нейронные сети, абстрактное реферирование

Содержание

Несмотря на универсальность текущего поколения языковых моделей, качество автоматического абстрактного реферирования до сих пор остается неустойчивым к контексту. Большие языковые модели, такие как ChatGPT, часто искажают факты в порождаемых рефератах, а на бенчмарках по абстрактному реферированию [1] проигрывают более ресурсоэффективным языковым моделям прошлого поколения BRIO-BART [2]. В то же время качество работы существующих систем автоматического абстрактного аннотирования сильно зависит от стилистических и структурных особенностей предметной области, что ограничивает возможность их практического применения.

Для объяснения феномена неустойчивости процесса автоматического порождения рефератов был проведен детальный анализ лучших предобученных нейросетевых моделей абстрактного реферирования. Эксперименты проводились на 5 предметных областях: новости, литература, научные статьи, патенты и руководства.

Анализовались внешние (степень абстрактности, устойчивость процесса генерации) и внутренние (влияние отдельных слоев на процесс распространения информации в сети) особенности моделей в условиях ограничений на обучающую выборку (few-shot tuning). В результате удалось установить явную закономерность между стратегией предобучения и особенностями работы моделей, обуславливающими эффективность их адаптации на документы различных предметных областей. Была выдвинута гипотеза, что современные предобученные модели абстрактного аннотирования имеют некорректное представление о статистическом позиционном распределении важной информации в документах, что в итоге приводит к снижению эффективности механизма перефразирования.

Для проверки гипотезы был проведен эксперимент по коррекции поведения модели Pegasus [3] путем дообучения на задаче ранжирования фрагментов в соответствии с информационной центральностью. В результате была получена новая модель абстрактного реферирования, Pegasus-SP, превосходящая предшествующие аналоги по качеству реферирования в режимах без обучения (zero-shot) и с частичным обучением (few-shot).

Также для коррекции поведения без дополнительного обучения и архитектурных модификаций (plug-and-play) на основе обнаруженной закономерности был разработан новый метод контроля моделей архитектуры encoder-decoder, Biased Encoder Mixture (BEM). Метод BEM позволяет интегрировать внешние системы отбора контента для коррекции распространения информации в архитектуре encoder-decoder, а также влиять на стилистику порождаемого текста вспомогательными кодировщиками. Тестирование на 4 наборах данных показало, что BEM превосходит аналоги по точности и полноте коррекции, достигая 10% относительного роста качества.

Литература

1. Goyal T., Li J. J., Durrett G. News summarization and evaluation in the era of gpt-3 //arXiv preprint arXiv:2209.12356. – 2022.
2. Liu Y. et al. BRIO: Bringing Order to Abstractive Summarization //Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2022. – С. 2890-2903
3. Zhang J. et al. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization //International Conference on Machine Learning. – PMLR, 2020. – С. 11328-11339.