Cross-Language Document Extractive Summarization with Neural Sequence Model*

Захаров П.С, Пискун М.Г., Сельницкий И.С., Кваша П.А., Дьячков Е.А, Петров Е.Д.

Московский физико-технический институт

В данной работе исследуется зависимость качества сокращения текста от качества перевода. Текст представляет собой краткое содержание документа на языке, отличном от языка написания документа. Для его получения используется сокращение документа выбором предложений с последующим машинным переводом; при отборе предложений учитывается не только их содержание, но и оценка качества перевода. Также в работе исследуются условия, при которых возможен перенос обучения на набор данных на другом языке.

Ключевые слова: *Сокращение текстов, машинный перевод, нейронные сети.*

1. Введение

Cross-language document summarization опредляется как задача составления краткого изложения текста на языке, отличном от языка текста документа(Wan et al., 2010). В наше время огромный объем информации доступен в Интернете, однако большинство людей не может с ним ознакомится из-за того, что информация представлена на незнакомом читателю языке.

Существует не так много исследований, посвященных этой задаче. Наиболее простыми простыми являются EarlyTrans и LateTrans. Суть первой заключается в том, что сначала переводится текст на другой язык а потом производится сокращение. Вторая же действует противоположенным образом: сначала сокращение, потом перевод. Позже Wan [6] усоверешнствовал вторую модель, став учитывать не только инфрмативность, но и качество перевода, также Wan [5] предложил одновременно ранжировать предложения на исходном языке и на языке, на который надо перевести. Linhares Pontes [3] и его группа предложили кроме того, чтобы использовать инофрмацию с двух языков, еще и сжимать предложения.

В этом исследовании мы фокусируемся на сокращении английского текста с последующим переводом на русский язык. Входными параметроми являются наборы текстов на англйисклм языке а выходными - краткое изложение этих документов на русском языке. Чтобы получить краткое изложение на русском языке необходимо произвести машинный перевод с английского на русский. Однако качество машинного перевода далеко не идеально: результаты перевода содержат множество ошибок и шумов. Это затрудняет задачу краткого изложения текста на разных языках, поскольку ошибки и шумы, вызванные машинным переводом, оказывают большое негативное влияние на оценку предложения или краткое изложение на целевом языке.

В данной работе предлагается развить идею Wan [6] в приложении к сокращению с переводом с английского языка на русский. Планирутся использовать имеющуюся модель сокращения текстов SummaRuNNer, предложенную Nallapati [2], стоит отметь существующее улучшение этой модели [4], в котором оценивалось связанность пердложения с предыдущими и следующими за ним приидложениями. Обучаться модель будет на CNN/DailyMail corpus. Для перовода воспользуемся фреймворком openNMT, описанным

Научный руководитель: Стрижов В. В. Задачу поставил: Хританков А. С. Консультант: Романов А. В.

в [1]. Для данной модели необходим параллельный корпус OPUS. Ставится задача решить проблемы модели сокращения текста, связанные с переносом на другую выбору документов, а также определить необходимость дополнительной предобработки текстов и границы применимости модели.

Литература

[1] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation. *CoRR*, abs/1701.02810, 2017.

[2] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. *CoRR*, abs/1611.04230, 2016.

[3] Elvys Linhares Pontes, Stéphane Huet, Juan-Manuel Torres-Moreno, and Andréa Carneiro Linhares. Cross-language text summarization using sentence and multi-sentence compression. In *Natural Language Processing and Information Systems*, pages 467–479. Springer International Publishing, 2018.

[4] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Furu Wei, Jun Ma, and Maarten de Rijke. Leveraging contextual sentence relations for extractive summarization using a neural attention model. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 95–104, New York, NY, USA, 2017. ACM.

[5] Xiaojun Wan. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1546–1555, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[6] Xiaojun Wan, Huiying Li, and Jianguo Xiao. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 917–926, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.