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 представлена модель образования краткого изложения текста на языке, отличном от текста документа. Для этого используется сокращение текста выбором предложений с последующим машинным переводом; при отборе предложений учитывается не только их содержание, но и оценка качества перевода. Исследуется зависимость качества сокращения от качества перевода. Перевод и сокращение осуществляются специально спроектированными для этих целей нейронными сетями. При этом базовая модель исследовалась на малом числе наборов данных; в этой работе идет дальнейшее рассмотрение переносимости этой модели на другие данные и внесение коррективов для улучшения модели в будущем.

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

Введение

Данное исследование посвящено задаче аннотирования, т.е. краткого изложения текстов. Задача машинного аннотирования возникла в связи с развитием крупных хранилищ документов (в данном исследовании - статей), которые требуется представить в удобном для быстрой оценки виде. При решении задач подобного рода можно выделить два подхода: абстрактное и экстрактивное изложение. В первом случае аннотация является полностью синтетической, в то время как второй подход отбирает предложения из исходного текста. В данной работе рассматривается в основном экстрактивное аннотирование, т.к. оно проще в реализации и в целом на настоящий момент показывает лучшие результаты [2].

Необходимость делать сокращения текстов на других языках и развитие технологий машинного перевода подтолкнуло создание моделей, реализующих межъязыковое аннотирование текстов (англ. Cross-Language Automatic Text Summarization). Наиболее простым решением проблемы является последовательное применение двух техник. Такие модели называются LateTrans и EarlyTrans - в первом случае сначала идет изложение на языке оригинала, а затем перевод, во втором - наоборот. Обе концепции показали себя не лучшим образом в связи с несовершенством обеих технологий: неидеальный выход первой модели еще сильнее искажался второй. Wan и др.[4] предложили идею усовершенстованной LateTrans модели: при аннотировании на языке оригинала учитывались не только информативность предложения, но и предсказание качества перевода. Помимо этого, Wan и др. [5] реализовал систему, создающую изложения-кандидаты, полученные, разными способами, и отбирающую лучшие из них. Pontes и др.[3] использовали кластеризацию и сжатие исходных предложений для получения более информативных предложений для отбора.

Помимо совершенствования систем в целом, ведутся дальнейшие исследования в моноязыковом аннотировании, [6][2]. Можно также отметить прогресс в, к примеру, задаче векторизации предложений [7]. Модульная архитектура позволяет использовать эти наработки в содании более эффективных CLATS-моделей.

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

В данной работе предлагается развить идею Wan[4] в приложении к сокращению с переводом с английского языка на русский, используя более совершенные составляющие: в качестве базовой модели используется SummaRunner2016[2], для перевода - openNMT, описанная в [1]. 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.

Для обучения SummaRunner используется исходная выборка - CNN/DailyMail corpus, а для обучения openNMT - параллельный корпус OPUS. Кроме того, имеются данные на русском языке для оценки качества итогового изложения.

Постановка задачи

Литература

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[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] 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.

[5] Xiaojun Wan, Fuli Luo, Xue Sun, Songfang Huang, and Jin ge Yao. Cross-language document summarization via extraction and ranking of multiple summaries. *Knowledge and Information Systems*, jan 2018.

[6] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[7] Chi Zhang, Shagan Sah, Thang Nguyen, Dheeraj Peri, Alexander Loui, Carl Salvaggio, and Raymond Ptucha. Semantic sentence embeddings for paraphrasing and text summarization. In *GlobalSIP*, pages 705–709. IEEE, 2017.