Cross-Language Document Extractive Summarization with Neural Sequence Model*

Захаров П.С., 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: *Аннотирование текстов, машинный перевод, нейронные сети.*

1. Введение

Межъязыковое реферирование текстов (англ. Cross-Language Automatic Text 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.

Необходимость делать сокращение текстов на других языках и развитие технологий машинного перевода подтолкнуло к созданию моделей, реализующщих межъязыковое реферирование текстов. Наиболее простыми решениями являются EarlyTrans и LateTrans, суть которых заключается в последовательном применении двух инструментов - машинного перевода и моноязыкового реферирования, однако они отличаются поледовательностью применения этих инструментов: первая - сначала переводит текст на другой язык а потом производит сокращение, вторая - наоборот. Обе концепции были слишком неточны, так как неточный результат первой модели еще сильнее искажался второй. 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 исходные предложения.

В этом исследовании мы фокусируемся на сокращении английского текста с последующим переводом на русский язык. Входными параметроми являются наборы текстов на англйсклм 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 работе предлагается развить идею Wan [6], предполагающую предсказание качества машинного паеревода для исходных предложений с целью дальнейшего отбора предложений, в приложении к сокращению с переводом с английского языка на русский. Планирутся использовать имеющуюся модель сокращения

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

текстов SummaRuNNer, предложенную Nallapati [2], стоит отметь существующее улучшение этой модели [4], в котором оценивалось связанность пердложения с предыдущими и следующими за ним приидложениями. Для обучения нейросети будет использоваться минимизация логистической функции потерь. Для перевода воспользуемся фреймворком openNMT, описанным в [1], для обучения которой используется минимизация среднеквадратичной ошибки и коэффициента Пирсона. Для конечной оценки качества используется метрика ROUGE. Ставится задача решить проблемы модели сокращения текста, связанные с переносом на другую выборку документов, а также определить необходимость дополнительной предобработки текстов и границы применимости модели.

Для обучения SummaRunner используется исходный датасет - CNN/DailyMail corpus, а для обучения openNMT - параллельный корпус OPUS.

2. Постановка задачи

В основе реализуемой модели лежит объединение модели моноязыкового аннотирования и модели предсказания качества машинного перевода. В следующих двух подразделах по отдельности поставлены задачи для каждой из двух моделей, в третьем описано их объединение.

2.1 Извлечение предложений

В этой работе мы рассматриваем извлечение предложений как проблему классификации последовательностей, в которой для каждого предложение принимается решение включать его в вывод модели, учитывая уже выбранные предложения. Используется рекуррентная нейронная сеть в качестве основоного блока классификатора последовательности. Введем следующие обозначения: $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$ - выборка (предложение в документе и бинарное целевое значение для этого предложения), где $\mathbf{X} = \{\mathbf{X}_i\}, \mathbf{X}_i \in \mathbb{R}^{N_i \times n}$ - набор предложений, $i \in \{1, .M\}$. При этом $\mathbf{X}_i = [\mathbf{x}_{ij}] \in \mathbb{R}^{N_i \times n}$ - предложения, состоящие из векторных представлений слов длиной n. Нейронная сеть задается двумя гейтами $\mathbf{u}_j$ и $\mathbf{r}_j$ следующими формулами:

$$\mathbf{u_j} = \sigma\left(\mathbf{W}_{ux}\mathbf{x}_j + \mathbf{W}_{uh}\mathbf{h}_{j-1} + \mathbf{b}_j\right) \quad (1)$$

$$\mathbf{r}_j = \sigma\left(\mathbf{W}_{rx}\mathbf{x}_j + \mathbf{W}_{rh}\mathbf{h}_{j-1} + \mathbf{b}_r\right) \quad (2)$$

$$\mathbf{h}'_j = tanh\left(\mathbf{W}_{hx}\mathbf{x}_j + \mathbf{W}_{hh}\left(\mathbf{r}_j \odot \mathbf{h}_{j-1}\right) + \mathbf{b}_j\right) \quad (3)$$

$$\mathbf{h}_j = \left(1 - \mathbf{u}_j\right) \odot \mathbf{h}'_j + \mathbf{u}_j \odot \mathbf{h}_{j-1} \quad (4)$$

где $\mathbf{W}$ и $\mathbf{b}$ глобальные параметры рекурентной нейронной сети, $\mathbf{h}_j$ вещественный вектор скрытого сотсояния в момент времени j, $\mathbf{x}_j$ соответсвенно входной вектор, $\odot$ - произведение Адамара.

Модель состоит из двухслойной двунаправленной рекурентной нейронной сети (RNN). Первый уровень RNN работает на уровне слова - $\mathbf{x}_j = \mathbf{x}_{ij}, j \in \{1..N_i\}$ и последовательно вычисляет представления скрытого слоя в каждой позиции слова на основе текущего вхождения и предыдущего состояния скрытого слоя. Мы также используем другую RNN на уровне слов - $\mathbf{x}_j = \mathbf{x}_{iN_i-j}, j \in \{1..N_i\}$, которая идет назад от последнего слова к первому, и мы относим пару прямых и обратных RNN как двунаправленную RNN.

Модель также состоит из второго слоя двунаправленной RNN, которая работает на уровне предложения и принимает в качестве входных данных объединенные скрытые состояния двунаправленной RNN уровня слова. Для прямой RNN:

$$\mathbf{x}_j = \frac{1}{N_j}\sum_{k=1}^{N_j}\left[\mathbf{h}_k^f, \mathbf{h}_k^b\right], \quad (5)$$

Для обратной:

$$\mathbf{x}_j = \frac{1}{N_{M-j}} \sum_{k=1}^{N_{M-j}} \left[\mathbf{h}^f_{M-k}, \mathbf{h}^b_{M-k} \right], \tag{6}$$

где $\mathbf{h}_j = k^f$ и$\mathbf{h}_j = k^b$- скрытое состояние нейронов относящееся к j-ому преложению прямой и обратной RNN уровня слов. $\mathbf{N}_j$ - количество слов в предложении. Квадратные скобки означают конкатенацию векторов.

Представление документа d формируется следующим образом:

$$\mathbf{d} = tanh \left(W_d \frac{1}{M} \sum_{j=1}^{M} \left[\mathbf{h}^f_j, \mathbf{h}^b_j \right] + \mathbf{b}_j \right), \tag{7}$$

где M количество предложений а $\mathbf{h}^f_j$ и $\mathbf{h}^b_j$ - скрытые состояния прямой и обратной цепочек RNN на уровне предложений

Для классификации каждое предложение подается на вход логистическому слою, который принимает решение, будет ли предложение принадлежать аннотации:

$$P\left(y_j = 1 | \mathbf{h}_j, \mathbf{s}_j, \mathbf{d}\right) = \sigma\left(\mathbf{W}_c \mathbf{h}_j + \mathbf{h}^T_j \mathbf{W}_s \mathbf{d} - \mathbf{h}^T_j \mathbf{W}_r tanh\left(\mathbf{s}_j\right) + \mathbf{W}_{ap} \mathbf{p}^a_j + \mathbf{W}_{rp} \mathbf{p}^r_j + \mathbf{b}\right) \tag{8}$$

где $\mathbf{y}_j$ - бинарное значение, показывающее, будет ли j-ое предложение принадлежать аннотации, $\mathbf{s}_j$ - динамическое представление аннотации на j-ом шаге, $\mathbf{p}^a_j$ и $\mathbf{p}^r_j$ - абсолютные и относительные положения в документе, $\mathbf{h}_j$ скрытое сотсояние второй нейронной сети, работающей на уровне предложений, ($\mathbf{W}_c \mathbf{h}$ отвечает за информативность j-ого предложения, $\mathbf{h}^T_j \mathbf{W}_s \mathbf{d}$ отвечает за значимость предложения по отношению к документу, $\mathbf{h}^T_j \mathbf{W}_r tanh\left(\mathbf{s}\right)$ фиксирует избыточность предложения относительно текущего состояния аннотации, $\mathbf{W}_{ap}$ и $\mathbf{W}_{rp}$ коэффициенты важности абосютного и относительного положения предложения в документе.

Ставится задача минимизовать логистическую функцию правдоподобия:

$$\begin{gathered} l\left(\mathbf{W}, \mathbf{b}\right) = -\sum_{k=1}^{D} \sum_{j=1}^{M_k} \left(y^k_j log P\left(y^k_j = 1 | \mathbf{h}^k_j, \mathbf{s}^k_j, \mathbf{d}_k\right)\right) + \\ + \left(1 - y^k_j\right) log\left(1 - P\left(y^k_j = 1 | \mathbf{h}^k_j, \mathbf{s}^k_j, \mathbf{d}_k\right)\right) \rightarrow min \end{gathered} \tag{9}$$

Полученное мягкое предсказание в дальнейшем используется для формирования конечного прогноза.

2.2 Предсказание качества машинного перевода

Пусть $\mathfrak{D} = (\mathbf{X}, \mathbf{y}) \subset \mathbb{R}^{m \times M} \times \mathbb{R}^M$ - выборка (объекты и целевой вектор), $\mathbf{X} = [\mathbf{x}_i] \in \mathbb{R}^m$ - объекты (предложения). В этом разделе мы будем пользоваться ε-SVR методом (Vapnik 1995) для задачи прогнозирования качества машинного перевода предложения. Требуется найти такую гладкую функцию f, которая аппроксиммирует отношение между точками $(\mathbf{x}_i, y_i)$:

$$\underset{\mathbf{w}, b, \xi, \xi^*}{min} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{n} \xi_i + C \sum_{i=1}^{n} \xi^*_i \tag{10}$$

при условии, что

$$\begin{aligned} \mathbf{w}^T\mathbf{f}(\mathbf{V}_i) + b - y_i &\leqslant \varepsilon + \xi_i \\ y_i - \mathbf{w}^T\mathbf{f}(\mathbf{V}_i) - b &\leqslant \varepsilon + \xi_i^* \\ \varepsilon, \xi_i, \xi_i^* \geqslant 0, i = 1, .., M & \end{aligned} \tag{11}$$

Константа C>0 является параметром, определяющий компромисс между гладкостью f и величины, до которой допускаются отклонения, превышающие ε.

Для оценки результатов предсказания используются две метрики

$$\begin{aligned} \mathbf{MSE} &= \frac{1}{M}\sum_{i=1}^{M}(\hat{y}_i - y_i)^2 \quad \text{- среднеквадратичная ошибка} \\ \rho &= \frac{\sum_{i=1}^{M}(y_i - \overline{y})\left(\hat{y}_i - \overline{\hat{y}}\right)}{M s_y s_{\hat{y}}} \quad \text{- коэффициент Пирсона,} \end{aligned}$$

где $\overline{y}$ и $\overline{\hat{y}}$ - средние значения точного результата и предсказанных предложений соответственно , а s_y и $s_{\hat{y}}$ - их среднеквадратичные отклонения.

После обучения выданные оценки качества нормируются максимальным значением в документе: $\tilde{y_i} = \frac{\hat{y_i}}{\underset{i}{max}\hat{y_i}}$

2.3 Предсказание качества машинного перевода

После получения оценки информативности каждого предложения $y_{INF,i}$ и качества перевода $y_{MT,i}$ и (см.стр 2, 3), принимающих значения $y_{INF,i}, y_{MT,i} \in [0, 1]$, итоговая оценка вычисляется следующим образом:

$$y_{overall,i} = (1 - \lambda)\, y_{INF,i} + \lambda y_{MT,i} \tag{12}$$

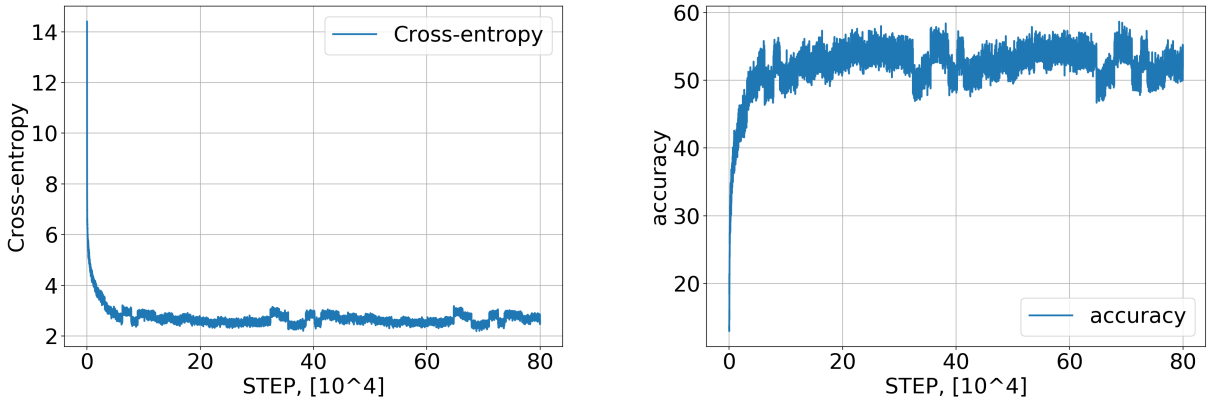
где $\lambda \in [0, 1]$ - глобальный параметр контролирующий важность одной оценки - качество перевода, относительно другогой - иформативность. После этого отбирается несколько предложений с наивысшими оценками. Количество отбираемых предлжений также настраиваемый параметр, зависящий от метрики ROUGE. После этого происходит перевод, описание см. в [1].

3. Вычислительный эксперимент

3.1 OpenNMT

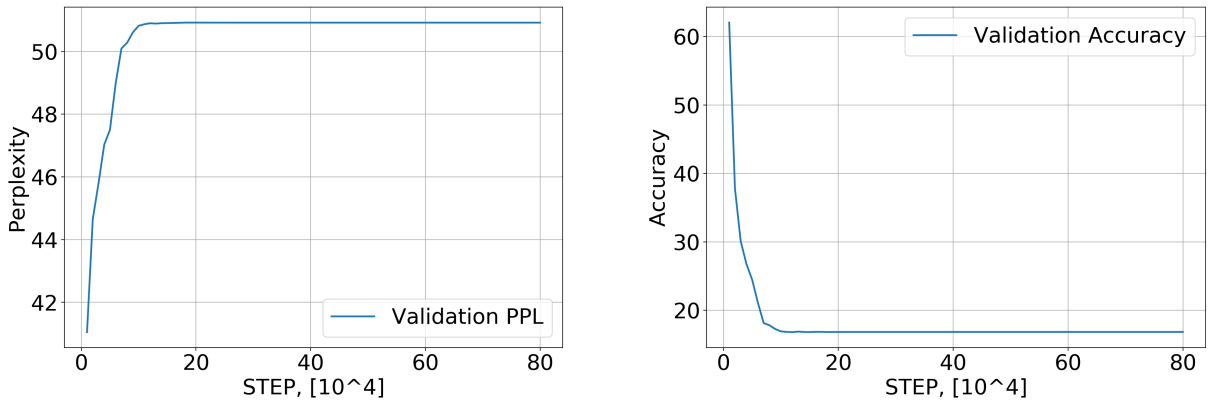
Обучение OpenNMT происходило на корпусе OpenSubtitles2018[1] - 20728084 объектов. Размерность скрытого состояния LSTM - 500. Сделано 800000 итераций, размер пачки данных - 64, размерность данных для валидации 64

[1] `http://opus.nlpl.eu/OpenSubtitles2018.php`



(a) Кросс-энтропия (b) Точность

Рис. 1. Обучение OpenNMT



(a) Perplexity (b) Точность

Рис. 2. Валидация OpenNMT

Видно, что на валидации точность и perplexity выходят на постоянный уровень: 51 и 15 соответственно Также на обучении кросс-энтропия и точность выходят практически на постоянный уровень и колеблются около него: 2.2 и 55

3.1 SummaRuNNer

Для обучения SummaRuNNer использовался датасет CNN/DailyMail - 193983 объекта. Размерность скрытого состояния - 200, количество эпох - 20. Размер словаря - 153824, размер пачки данных - 32, размерность данных для валидации 32

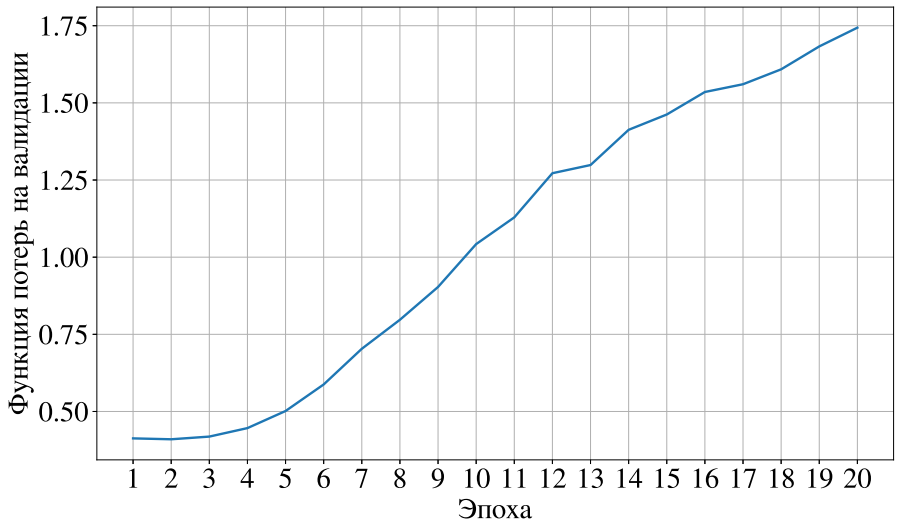


Рис. 3. Функция потерь на валидации в зависимости от числа прошедших эпох

Кроме того, нейронная сеть была обучена на 1/8, 1/16 и 1/32 обучающей выборки; были сравнены результаты теста по метрике ROUGE:

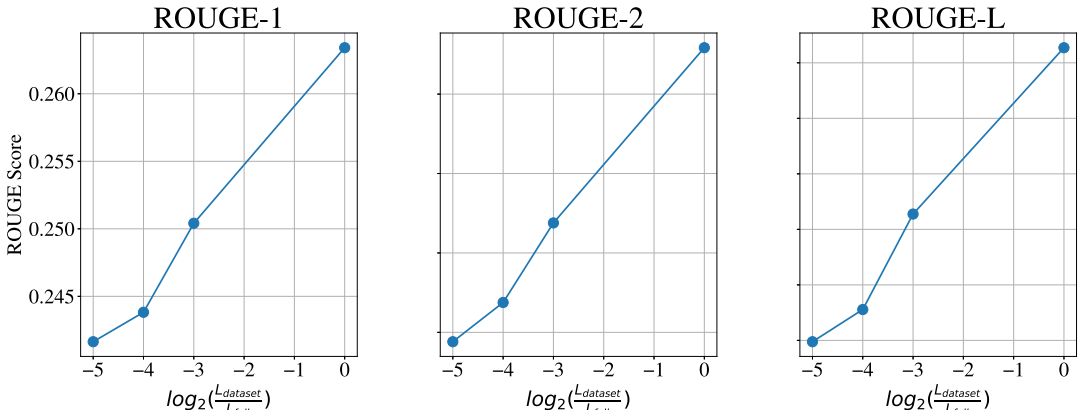


Рис. 4. Метрика качества извлечения в зависимости от размера выборки

Видно, что можно безпрепятственно сократить число эпох до 5, SummaRuNNer сохраняет всегда только лучшую модель и во всех эксперминтах эта модель получалось на 2-5 эпохах, дальнейшее увеличения числа итераций не приводит к получению более качественной модели. Так же можно сократить размер выборки тем самым значительно сократив время обучения

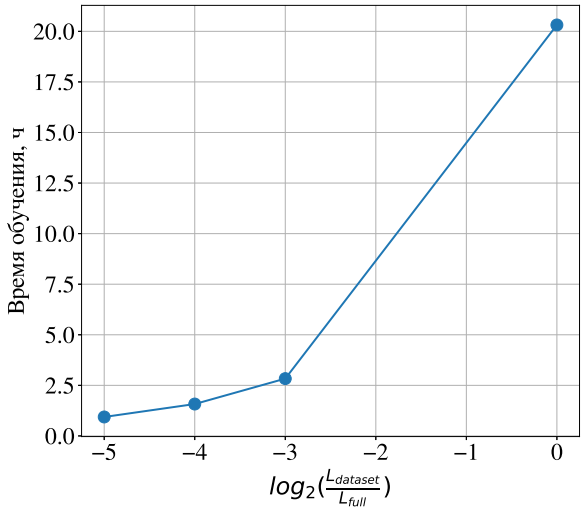


Рис. 5. Время обучения модели на различных размерах выборки

Ниже произведено сравнение используемой модели с разными опциями и некоторыми аналогичными моделями.

Таблица 1. Обучение SummaRuNNer на различных объемах выборки

Модель	ROUGE-1 Recall	ROUGE-2 Recall	ROUGE-L Recall
Использ. реализация	0.26341	0.11792	0.14068
Реализация в [2]	0.262	0.108	0.144
CNN-RNN^2	0.258	0.113	0.138
Hierarch. Attention	0.26	0.114	0.138

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.