

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 представлена модель реферирования текстов 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 реферирования, т.е. краткого изложения текстов. Задача машинного реферирования 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: обобщение и извлечение. В первом случае сокращенный текст генерируется синтетически, в то время как второй подход отбирает предложения из исходного текста. В данной работе рассматривается в основном извлечение, т.к. оно проще в реализации и на настоящий момент показывает лучшие результаты [1].

Необходимость делать сокращения текстов на других языках и развитие технологий машинного перевода подтолкнуло создание моделей, реализующих межъязыковое реферирование текстов (англ. Cross-Language Automatic Text Summarization). Наиболее простым решением проблемы является последовательное применение двух инструментов - моноязыкового реферирования и машинного перевода. Такие модели бывают двух типов - LateTrans и EarlyTrans - в первом случае сначала идет сокращение 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.[2] предложили идею усовершенствованной LateTrans модели: при реферировании на языке оригинала учитывались не только информативность предложения, но и предсказание качества перевода. Помимо этого, Wan и др. [3] реализовал систему, создающую изложения-кандидаты, полученные, разными способами, и отбирающую лучшие из них. Pontes и др.[4] использовали кластеризацию и сжатие исходных предложений для получения более информативных предложений для отбора.

Помимо совершенствования систем в целом, ведутся дальнейшие исследования в моноязыковом реферировании, [1][5]. Можно также отметить работы по векторизации предложений [6]. Модульная архитектура позволяет использовать эти наработки в создании более эффективных CLATS-моделей.

В данной работе предлагается развить идею Wan [2] в приложении к сокращению с переводом с английского языка на русский. Данная архитектура предполагает предсказание

качества машинного перевода для исходных предложений и учет этих предсказаний при отборе кандидатов из предложений. После совершенного с помощью SummaRunner2016 [1] извлечения предложений на английском языке полученные сокращенные тексты переводятся. В базовой модели для перевода используется пакет openNMT, описанный в [7]. Для обучения нейронной сети отбора предложений используется logloss (термин на русском?), для обучения модели предсказания - MSE и коэффициент Пирсона, для конечной оценки качества - 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 исходная выборка - CNN/DailyMail corpus, а для обучения openNMT - параллельный корпус OPUS. Кроме того, имеются данные на русском языке для оценки качества итогового изложения. (будет уточнено позднее, когда появится более подробная информация)

Постановка задачи

В основе реализуемой модели лежит объединение модели моноязыкового аннотирования и модели предсказания качества машинного перевода. В следующих 2 подразделах по отдельности поставлены задачи для каждой из двух моделей, в третьем описано их объединение. Нужно что-то сказать про гипотезу порождения данных. При созвоне что-то говорилось про i.i.d, но ведь это не так! мы для того и используем RNN, чтобы использовать зависимости между словами, между предложениями. Или я что-то не понимаю?

Извлечение

Для реферирования используется трехслойная двухсторонняя рекуррентная нейронная сеть. Пусть $\mathfrak{D} = (\mathfrak{V}, \mathbf{Y})$ - выборка (предложения в документе и бинарный целевой вектор), где $\mathfrak{V} = \{\mathbf{V}_i\}$, $\mathbf{V}_i \in \mathbb{R}^{N_i \times n}$ - набор предложений, $i \in \{1..M\}$. При этом $\mathbf{V}_i = [\mathbf{v}_{ij}] \in \mathbb{R}^{N_i \times n}$ - предложения, состоящие из векторных представлений слов длиной n . Слои первых двух слоев нейронной сети состоят из нейронов, которые описываются двумя гейтами - как правильно? \mathbf{u}_j и \mathbf{r}_j по следующим формулам:

$$\begin{aligned} \mathbf{u}_j &= \sigma(\mathbf{W}_{ux}\mathbf{x}_j + \mathbf{W}_{uh}\mathbf{h}_{j-1} + \mathbf{b}_j) \\ \mathbf{r}_j &= \sigma(\mathbf{W}_{rx}\mathbf{x}_j + \mathbf{W}_{rh}\mathbf{h}_{j-1} + \mathbf{b}_r) \\ \mathbf{h}'_j &= \tanh(\mathbf{W}_{hx}\mathbf{x}_j + \mathbf{W}_{hh}(\mathbf{r}_j \odot \mathbf{h}_{j-1}) + \mathbf{b}_j) \\ \mathbf{h}_j &= (1 - \mathbf{u}_j) \odot \mathbf{h}'_j + \mathbf{u}_j \odot \mathbf{h}_{j-1} \end{aligned} \quad (1)$$

На первом слое строится две цепочки нейронов для каждого предложения в тексте. Для одной цепочки $\mathbf{x}_j = \mathbf{v}_{ij}$, $j \in \{1..N_i\}$, для другой $\mathbf{x}_j = \mathbf{v}_{iN_i-j}$, $j \in \{1..N_i\}$ - во второй цепочке слова подаются в обратном порядке. Эту цепочку будем называть обратной, а первую - прямой. Здесь N_i - количество слов в i -ом предложении

На втором слое строятся такие же цепочки нейронов, для прямой:

$$\mathbf{x}_j = \frac{1}{N_j} \sum_{k=1}^{N_j} [\mathbf{h}_j^f, \mathbf{h}_j^b], \quad (2)$$

где \mathbf{h}_j^f - скрытое состояние нейронов прямой цепочки для j -ого предложения, а \mathbf{h}_j^b - обратной. Квадратные скобки означают конкатенацию векторов. Для обратной цепочки:

$$\mathbf{x}_j = \frac{1}{N_{M-j}} \sum_{k=1}^{N_{M-j}} [\mathbf{h}_{M-j}^f, \mathbf{h}_{M-j}^b], \quad (3)$$

где M - число предложений в документе Представление документа \mathbf{d} формируется следующим образом:

$$\mathbf{d} = \tanh \left(W_d \frac{1}{M} \sum_{j=1}^M [\mathbf{h}_j^f, \mathbf{h}_j^b] + \mathbf{b}_j \right), \quad (4)$$

где \mathbf{h}_j^f и \mathbf{h}_j^b - скрытые состояния прямой и обратной цепочек на втором слое.

Для классификации используется логистический слой:

$$P(y_j = 1 | \mathbf{h}_j, \mathbf{s}_j, \mathbf{d}) = \sigma \left(\mathbf{W}_c \mathbf{h}_j + \mathbf{h}_j^T \mathbf{W}_s \mathbf{d} - \mathbf{h}_j^T \mathbf{W}_r \tanh(\mathbf{s}_j) + \mathbf{W}_{ap} \mathbf{p}_j^a + \mathbf{W}_{rp} \mathbf{p}_j^r + \mathbf{b} \right) \quad (5)$$

Здесь \mathbf{s}_j - динамическое представление аннотации на j -ом шаге, а \mathbf{p}_j^a и \mathbf{p}_j^r - абсолютные и относительные положения в документе. Члены, обозначенные \mathbf{W} и \mathbf{b} с индексами, являются параметрами модели. Представление аннотации определяется следующим образом:

$$\mathbf{s}_j = \sum_{i=1}^{j-1} \mathbf{h}_i P(y_i = 1 | \mathbf{h}_i, \mathbf{s}_i, \mathbf{d}) \quad (6)$$

Ставится задача минимизовать логистическую функцию правдоподобия:

$$\begin{aligned} l(\mathbf{W}, \mathbf{b}) = & - \sum_{k=1}^D \sum_{j=1}^{M_k} (y_j^k \log P(y_j^k = 1 | \mathbf{h}_j^k, \mathbf{s}_j^k, \mathbf{d}_k) + \\ & + (1 - y_j^k) \log (1 - P(y_j^k = 1 | \mathbf{h}_j^k, \mathbf{s}_j^k, \mathbf{d}_k))) \rightarrow \min \end{aligned} \quad (7)$$

Полученное мягкое предсказание в дальнейшем используется для формирования конечного прогноза.

Предсказание качества машинного перевода

Пусть $\mathfrak{D} = (\mathfrak{V}, \mathbf{Y}) \subset \mathbb{R}^{m \times M} \times \mathbb{R}^M$ - выборка (объекты и целевой вектор), $\mathfrak{V} = [\mathbf{V}_i] \in \mathbb{R}^m$ - объекты (предложения). Подчеркнем, что ввиду решения другой задачи в этом подразделе представление предложений отличается - здесь они сами являются объектами, в то время как в предыдущем объектами были слова предложений.

Для предсказания качества машинного перевода используется ϵ -SVR метод. Требуется найти гладкую функцию \mathbf{f} такую, что:

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i + C \sum_{i=1}^n \xi_i^* \quad (8)$$

при условии, что

$$\begin{aligned} \mathbf{w}^T \mathbf{f}(\mathbf{V}_i) + b - y_i & \leq \epsilon + \xi_i \\ y_i - \mathbf{w}^T \mathbf{f}(\mathbf{V}_i) - b & \leq \epsilon + \xi_i^* \\ \epsilon, \xi_i, \xi_i^* & \geq 0, i = 1, \dots, M \end{aligned} \quad (9)$$

Метриками качества являются

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

где \bar{y}, \bar{y} - средние предсказанных и данных значений соответственно, а $s_y, s_{\hat{y}}$ - их средне-квадратичные отклонения.

После обучения выданные оценки качества нормализуются максимальным значением в документе: $\tilde{y}_i = \frac{y_i}{\max_i y_i}$

Получение итогового результата

На выходе моделей для каждого предложения V_i получены результаты $y_{MT,i}, y_{ES,i} \in [0, 1]$. Итоговое предсказание строится по правилу

$$y_{final,i} = y_{MT,i} (1 - \lambda) + y_{ES,i} \lambda, \quad (10)$$

где λ - эмпирически подбираемый параметр, отражающий важность информативности предложения по сравнению с предполагаемым качеством перевода. К примеру, при $\lambda = 1$ качество перевода вообще не учитывается. После этого происходит перевод, описание см. в [8][7]. Итоговое качество реферирования оценивается по метрике ROUGE (Какую именно? скорее всего, будет выбрано несколько - допишется, когда конкретно определимся)

Литература

- [1] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," *CoRR*, vol. abs/1611.04230, 2016.
- [2] X. Wan, H. Li, and J. 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, (Stroudsburg, PA, USA), pp. 917–926, Association for Computational Linguistics, 2010.
- [3] X. Wan, F. Luo, X. Sun, S. Huang, and J. ge Yao, "Cross-language document summarization via extraction and ranking of multiple summaries," *Knowledge and Information Systems*, jan 2018.
- [4] E. L. Pontes, S. Huet, J.-M. Torres-Moreno, and A. C. Linhares, "Cross-language text summarization using sentence and multi-sentence compression," in *Natural Language Processing and Information Systems*, pp. 467–479, Springer International Publishing, 2018.
- [5] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016.
- [6] C. Zhang, S. Sah, T. Nguyen, D. Peri, A. Loui, C. Salvaggio, and R. Ptucha, "Semantic sentence embeddings for paraphrasing and text summarization," in *GlobalSIP*, pp. 705–709, IEEE, 2017.
- [7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.
- [8] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," *CoRR*, vol. abs/1701.02810, 2017.