ГЕНЕРАЦИЯ ЕСТЕСТВЕННОГО ЯЗЫКА, ПАРАФРАЗ И АВТОМАТИЧЕСКОЕ ОБОБЩЕНИЕ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ С ПОМОЩЬЮ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ

Тарасов Д. С. (dtarasov3@gmail.com)

Интернет-портал reviewdot.ru, Россия, Казань

Ключевые слова: генерация естественного языка, генерация парафраз, автоматическое обобщение отзывов, рекуррентные нейронные сети

NATURAL LANGUAGE GENERATION, PARAPHRASING AND SUMMARIZATION OF USER REVIEWS WITH RECURRENT NEURAL NETWORKS

Tarasov D. S. (dtarasov3@gmail.com)

ReviewDot Research, Kazan, Russia

Multi-Document summarization and sentence generation are important challenges in natural language processing. This paper presents recurrent neural network (RNN) architecture capable of producing abstractive document summaries, as well as generating novel paraphrases of input sentences in the same language. We demonstrate practical application of our system on the task of multiple consumer reviews summarization.

Keywords: natural language generation, paraphrase generation, automatic summarization of user reviews, recurrent neural networks

1. Introduction

The main role of automatic document summarization is to help readers to understand most important points of long documents without much effort. One particular area of document summarization that attracted a lot of research attention is automatic

summarization of consumer reviews, also called opinion summarization. It is traditionally based on feature selection, feature rating and identifying important sentences, leading to so called extractive summaries (summaries that consists of original sentences extracted from user reviews) [Mei et al, 2007; Liu J. et al, 2012, Liu C. et al, 2012, Raut and Londhe, 2014]. Another kind of summaries is abstractive summaries (texts that summarize essential facts mentioned in reviews without using original sentences). Such texts tend to have better coverage for a particular level of conciseness, and to be less redundant and more coherent [Carenini et al, 2006]. They also can be constructed to target particular goals, such as summarization, comparison or recommendation.

Abstractive summarizers rely on natural language generation systems, that are currently designed using a lot of expert linguistic knowledge, heuristics and complex pipelines (that typically include text planner, sentence planner and surface realizer) [Fabbrizio et al, 2014]. Therefore adapting such systems to new languages and domains can be difficult. Up until now, only a few works considered machine learning based (trainable) language generation systems, and their success was limited [Ratnaparkhi, 2000;Hammervold, 2000]. However, recent research on neural networks demonstrated their capabilities to generate novel descriptions of pictures using purely machine-learning methods [Mao et al, 2014].

In this work we explore application of similar methodology to the domain of consumer reviews. We describe and evaluate recurrent neural network (RNN) model capable of generating novel sentences and document summaries.

To achieve this, we train recurrent neural network language model on a large number of sentences describing positive and negative aspects of various consumer products. In our setup, RNN task is to predict next word given current word and additional sentence-level semantic information that include sentence polarity, sentence length, product category and bag of aspects vector. In the test phase we give RNN sentence-level features vector and generate corresponding sentence.

We demonstrate that such relatively simple model can generate novel paraphrases that capture original meaning and show that this ability can be used to “compress” multiple important points about the product in one statement, thus producing concise multi-document summary. To do this, we first compute semantic vectors for all sentences in all available user reviews of a given product, combine them into two semantic vectors—positive (containing bag of positive aspects) and negative (containing bag of negative aspects). We then feed these vectors to language-generating RNN, obtaining sentences that sum up negative and positive product sides.

2. Related work

Convolutional neural networks were used for generation of extractive summaries of movie reviews [Denil et al, 2014]. In [Iyyer, 2014] paraphrase generation using tree-based autoencoders was demonstrated, however, no evolution of paraphrase quality was presented aside from few paraphrase examples. The approach of [Iyyer, 2014] also relies on dependency parse trees. Our method in contrast, does not use sentences parsers. It can be viewed as similar to encoder-decoder machine

translation models [Cho et al, 2014], while our RNN architecture is different and inspired by method of [Mao et al, 2014] where RNN was used to generate descriptions of pictures. We are not aware of any prior application of such models to abstractive text summarization or paraphrase generation.

3. Methods and algorithms

3.1. Datasets

We use database of 820,000 consumer reviews in Russian language from reviewdot.ru that was obtained by automatic crawling of more than 200 different web-resources. From that database we selected 120,000 reviews in 15 different product categories that had three sections (positive points, negative points and comments). These three sections are commonly used in Russian consumer reviews websites and reviewdot.ru crawler automatically detects them using heuristics-based algorithm. We then exclude sentences with unknown polarity and those with length more than 25 words, resulting in 56,000 training sentences. All sentences were padded with <START> and <END> special symbols.

3.2. Summarization Recurrent neural network model

The structure of our summarization recurrent neural network (s-RNN) is shown in Figure 1. The s-RNN model is deeper than the simple RNN model and similar to multimodal RNN introduced in [Mao et al, 2014]. It has five layers in each time frame: the input word layer, one projection layer, the recurrent layer, the summarization layer, and the softmax layer.

Projection layer implements table-lookup operation, converting word to real-valued embedding vector. Embedding vectors are obtained by training recurrent neural network language model [Mikolov et al, 2010] on 30M words dataset of consumer reviews.

Recurrent layer implements standard Elman-type [Elman, 1990] recurrent function:

$$h(t) = f\left(Wx(t) + Vh(t-1) + b\right)$$

Here f is a nonlinear function, (in our case hyperbolic tangent function), W and V are weight matrices between the projection and recurrent layer, and between the hidden units. U is the output weight matrix, b is bias vector connected to hidden and output units.

After the recurrent layer, we set up a summarization layer that connects the language model part and sentence-level semantics in s-RNN model. The language model part includes the projection layer and the recurrent layer. The sentence-level semantics contains the sentence features vector. We use sentence polarity, product category, bag-of-aspect-terms vector and sentence length as sentence-level features. While it is possible to incorporate more complex features, including these learned

by unsupervised neural network models, for this proof-of-principle experiment we avoid these additional complexities.

The softmax layer on top of the network generates the probability distribution of the next word.

Our s-RNN model was trained using backpropagation through time (BPTT) [Werbos, 1990] method with mini-batch gradient descent using one sentence per mini-batch as described in [Mesnil et al, 2013].

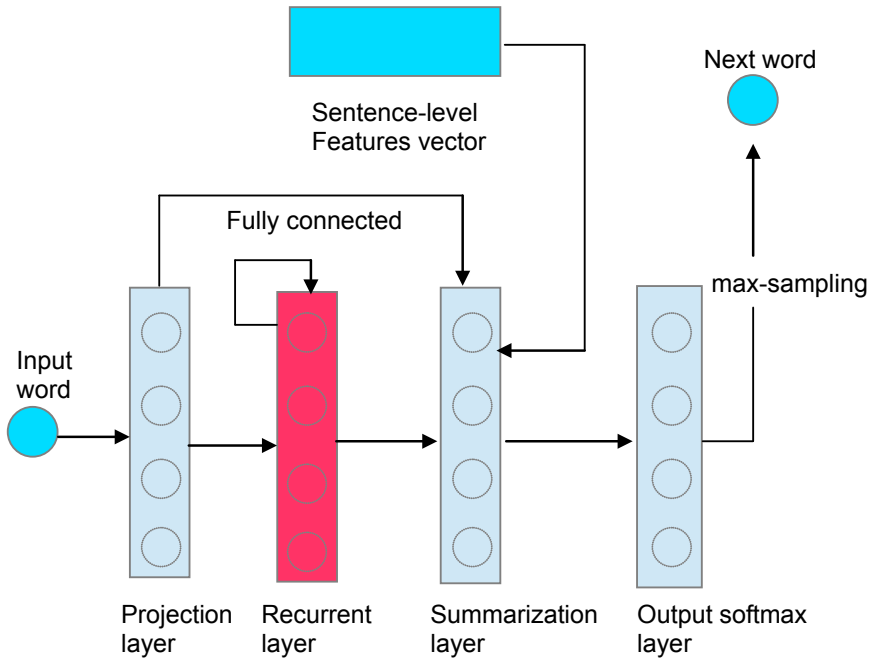

Figure 1. Architecture of summarization recurrent neural network

4. Results and discussion

4.1. Paraphrasing

To produce paraphrases, we give network sentence-level features vector of original sentence and then generate new sentence word-by-word, beginning from "<START>" symbol and stopping after network generates "<END>" symbol. Sentences quality was assessed by two human judges, who were asked to label sentences as "grammatically correct/not correct" and "conveying original meaning, not conveying original meaning". Results are summarized in table 1

Table 1. Human judgment of generated paraphrases (total of 75 sentences)

Human judgment	**Percentage of sentences (average value from results of two human judges)**
Grammatically correct and conveying original meaning	65%
Conveying original meaning but not necessary correct	78%
Correct, but not conveying original meaning	18%

Table 2. Example of original and paraphrased text

Original	**Paraphrased by neural network (two samples)**
Качественный и красивый корпус. (good and nice looking case)*	Красивый, качественный корпус, живучий. (Nice looking, good case, reliable) металический корпус довольно красиво выглядит (Metallic case looks reasonable good)
Большие кнопки с цифрами,хорошее качество изготовления,приятный корпус, лёгкий, громкий сигнал вызова (Big buttons with numbers, good production quality, nice case, lightweight, loud signal)	Вполне нравится пластик, цифры кнопок / качество изготовления металлический корпус. (Good quality of plastics, number of buttons / quality of production, metallic case) но кнопки , корпус , хорошее качество фотографий (But buttons, case, good quality of photo)
Довольно сносное качество исполнения, легкий и вполне удобный Quality is reasonable good, lightweight and convenient enough.	Маленький, легкий, удобный, хорошее качество звучания (Small, lightweight, convenient, good quality of sound) легкий , удобный , хорошее качество фотографий (lightweight, convenient, good quality of photo)
Очень простой, понятный и удобный в использовании. (Very simple, understandable and convenient in use)	Простой в использовании, удобный, недорогой. (Simple in use, convenient, not high-priced) мощный, простой, понятный аппарат удобный (powerful, simple, understandable device is convenient)

* English translations are human made, with an effort to preserve important sentence features.

As shown in table 2, most common mistakes are omissions of some original points and additions of new information that was not present in original sentence.

4.2. Language generation

Our design allows certain degree of control over the meaning of generated sentences. By choosing sentence-level features vector we can instruct the network, for example, to "say something good about screen and sound quality in about ten words".

We found, that better sentences are produced when number of words is set to roughly triple of the number of aspect terms. With smaller sentences, RNN just lists all aspects, and with larger values it tend to produce long phrases without well-defined meaning ("bright display from outside") and undesired additions such as "smart helps" (Table 3).

Table 3. Examples of sentences produced by s-RNN (polarity set to "positive" and aspects set to "battery, screen, convenience")

Desired sentence length	output
3	батарея, экран, удобный (battery, screen, convenient)
5	аккумулятор, размер дисплея солидный, эргономика (accumulator, impressive display size, ergonomics)
10	быстрый аккумулятор, яркий внешне дисплей, удобный функционал, умный помогает. (fast accumulator, bright display from outside, convenient functions, smart helps)

4.3. Summarization of multiple user reviews

Language-generating capacity of our RNN can be used for producing abstractive summaries of multiple user reviews. To achieve that we generate synthetic sentence-level feature vectors by running aspect-based sentiment analysis over all sentences of reviews subjected to summarization, using extracted aspect terms and polarities to generate feature vectors.

The major obstacle here is that our feature vectors capture only coarse-grained information (i.e. they can tell that display is good, but information why it is good is lost). Thus direct application of s-RNN usually leads to production of rather generic or plainly incorrect summaries.

To circumvent this problem, we use additional dynamic training step that consists of running one iteration of gradient descent over all sentences with aspect terms. We found that this method considerably improves quality of summaries, and allows incorporating fine-grained device-specific information.

Quality of review summaries were evaluated by two human judges who were given original reviews and asked to rate summary quality as good, acceptable or unacceptable. Table 4 presents averaged results.

Overall we found, that our method often produces summaries of reasonable quality, while still making a number of mistakes. Most commonly observed problem is inclusion

of seemingly irrelevant statements, such as "lot of different days". Also, we observed significant number of ungrammatical sentences, that can be result of relatively small training sample size, failure of RNN to capture long-term grammatical dependencies, and/or grammatical errors in the training samples (since user reviews typically contain certain number of ungrammatical phrases). The extent to which these factors contribute to generation of grammar errors is presently unknown and needs further investigation.

Still, we find it impressive that such relatively simple method can be used to solve multi-document summarization task—a problem that is generally considered difficult in natural language processing. Future work should include evaluation of proposed methods on different datasets and also investigation of possible use of trainable sentence-level feature vectors instead of pre-defined ones.

Table 4. Human evaluation of review summaries (100 summaries total).

Quality rating	Percentage of review summaries
Good	35%
Acceptable	44%
Unacceptable	21%

Table 5. Examples of generated summaries for two different mobile phones

Positives	Negatives
Качество звука, удобный интерфейс, очень долго держит заряд. Отзывчивый экран, громкий звонок, крупный шрифт, рабочий день. Приятно лежит в руках, 2 сим—карты выручают. Качество сборки, батарея, удобное меню, устойчив к воздействию воды. Явно лидируют, сочный дисплей, качество связи, плеер, фонарь. Хорошая фотокамера, динамик (Quality of sound, convenient user interface, very long battery life. Responsive screen, loud calling signal, large font, working day. Lies in hands nicely, 2 sim cards help. Quality of production, convenient menu, waterproof. Obviously leading, nice display, player, bright light. Good photo-camera, speaker).	Не обнаружено (not found)
Аккумулятор, скорость красивая. Дизайн, звук, функционал, масса разных дней хватает. Красив, несколько назад, процессор отзывчивый сенсор. Красивый экран, цветопередача. Дизайн, батарея, не тормозят, практичный. (Accumulator, speed is beautiful. Design, sound, functions, lot of different days. Beatiful, few days ago, processor, responsive sensor. Nice screen, color reproduction. Design and battery is not slow, practical).	Скользкий панель громкости тиховат. Стирается, заметно ос виснет, появляется белый экран. (Slippery panel of volume is too quiet. Noticable shabby, OS hangs and white screen appears)

Acknowledgements

Author thanks anonymous reviewers for helpful comments on earlier drafts of the manuscript.

References

1. *Carenini, G., Ng, R. T., & Pauls, A.* (2006, April). Multi-Document Summarization of Evaluative Text. In EACL.
2. *Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y.* (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.
3. *Denil, M., Demiraj, A., Kalchbrenner, N., Blunsom, P., & de Freitas, N.* (2014). Modelling, Visualising and Summarising Documents with a Single Convolutional Neural Network. arXiv preprint arXiv:1406.3830.
4. *Elman J.* (1990). Finding structure in time. Cognitive science, 14(2):179–211.
5. *Fabbrizio D., Stent A. J., & Gaizauskas, R.* (2014). A Hybrid Approach to Multi-document Summarization of Opinions in Reviews. INLG, 54.
6. *Hammervold K.* (2000, June). Sentence generation and neural networks. In Proceedings of the first international conference on Natural language generation-Volume 14 (pp. 239–246). Association for Computational Linguistics.
7. *Iyyer M., Boyd-Graber J., Daumé H.* (2014). Generating Sentences from Semantic Vector Space Representations. NIPS Workshop on Learning Semantics.
8. *Liu C., Hsaio W.-H., Lee C.-H., Lu G., Jou E.* (2012). Movie Rating and Review Summarization in Mobile Environment. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 42, No. 3, May, pp. 397–406.
9. *Liu J., Seneff S., Zue V.* (2012). Harvesting and Summarizing User-Generated Content for Advanced Speech-Based HCI. IEEE Journal of Selected Topics in Signal Processing, Vol. 6, No. 8, pp.982–992
10. *Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L.* (2014). Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090.
11. *Mei Q., Ling X., Wondra M., Su H., ZHAI C.* (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In WWW '07: Proceedings of the 16th international conference on World Wide Web. ACM, New York, NY, USA, 171–180.
12. *Mesnil, G., He, X., Deng, L. & Bengio, Y.* (2013). Investigation of recurrent neural network architectures and learning methods for spoken language understanding. In INTERSPEECH pp. 3771–3775 : ISCA.
13. *Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S.* (2010, January). Recurrent neural network based language model. In INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26–30, 2010 (pp. 1045–1048).
14. *Ratnaparkhi A.* (2000, April). Trainable methods for surface natural language generation. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (pp. 194–201). Association for Computational Linguistics.
15. *Raut B., Londhe D. Survey on Opinion Mining and Summarization of User Reviews on Web* (2014). International Journal of Computer Science and Information Technologies, Vol. 5 (2), 1026–1030
16. *Werbos, P. J.* (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10), pp. 1550–1560