A survey on entity-aware news image captioning

Abstract

Nowadays, as information is exploding, automatic news image captioning has become meaningful to replace human labor for the extremely time-consuming captioning work. The challenge of news image captioning is generating captions with Named Entities, many of which are out-of-vocabulary. Template-based and End-to-end methods have been proposed to overcome the problem by infusing the Named Entities extracted from the given article into the generated captions. However, they still failed to extract Named Entities accurately enough, or combining text and visual features effectively. The potential solutions may be supplementing current datasets with image location information within articles, capturing relationships from the images, and proposing new metrics focusing on both recall and correctness. Besides, this survey also indicates three possible research directions in future: adapting the task formulation by extending the source of Named Entities, interpreting how the models work, and proposing new pipelines to leverage the advance of Large Language Models.

1. Introduction

News image captioning aims to generate brief descriptions of images in news reports automatically, given an image and the corresponding complete news article as input. Compared with standard image captioning tasks, news image captioning has two extra challenges: complex images and a high ratio of named entities. Specifically, news images always contain various objects while depicting an event, which imposes more requirements on datasets and models' image feature extraction sections. Besides, many Named Entities are out-of-vocabulary, which makes it difficult for models to embed and generate. Template-based[7] and end-to-end[8][9][10] methods have been proposed to handle that, but they still can be improved in many aspects, like context selection (3.1), multi-modal features combination (3.3), etc.

This survey aims to compare existing research and identify potential future directions. Specifically, the rest of the survey will start from existing datasets and then compare some state-of-the-art methods from the aspect of relevant context selection, image feature extraction, and captions generation, which are three main components of news image captioning. After clarifying what really matters for this task, the survey will turn to evaluation metrics in the next section. Finally, we envision some future directions with the hope of inspiring research in future.

2. Dataset

Currently, three benchmarks (GoodNews[7], NYTimes800k[8], Visual News[9]) are used for the news image captioning task due to their large scale and diverse information resources. Although there have already been many benchmarks for standard image captioning (like PASCAL[3], NLVR[4]), they are not sufficient for news image captioning owing to the low proportion of named entities and the limited complexity of objects depicted in the images. Besides, it is noticeable that the current three available datasets can lead a model to perform variously. We try investigating the reasons behind the discrepancies and exploring ways to improve existing datasets here.

Articles' average length can impact the effectiveness of certain modules. In practice, a normal challenge for news image captioning is that the articles are often too long to be fully processed by Large Language Models. Consequently, datasets with an average article length of less than 512 (like GoodNews) ignore the models' capability to handle lengthy articles. For example, Tell [8] suffered worse performance on GoodNews[7] after integrating the context selection strategy, which had been proven effective on other datasets with longer articles. Therefore, a suitable dataset is expected to have an average article length properly longer than 512 to ensure that a model can work effectively on long news articles in real-world scenarios.

Besides, image location information (i.e., the given image refers to which part of the article) can enhance the models' performance by filtering out irrelevant noise. Currently, Times800k is the only dataset providing such location information at the paragraph level, which enables almost all models to attain the highest recall and precision in predicting which Named Entities should appear in the generated captions. Thus, supplementing existing datasets with more precise location information at the sentence level may yield more benefits for the task (3.1).

Almost all works generate captions with Named Entities by extracting from the given articles, but few have noticed a problem: only 50% - 70% of Named Entities in ground truths are covered by the given articles. It is essential to align the uncovered ratio of the datasets with the data distribution in practice, as the value can significantly impact the performance of models (like Tell [8]). Additionally, exploring solutions from a methodological standpoint (3.3) is necessary.

3. Methodology

3.1 Relevant Context Extraction

Different from standard image captioning tasks, in addition to the given image, news image captioning also needs to extract

Named Entities from the accompanying news articles. A successful context selection strategy is vital in improving the accuracy of Named Entities extraction and caption generation. It serves two primary purposes: 1) filtering out the sentences unrelated to the image (as a news report may contain multiple images), and 2) reducing the input texts to a length shorter than the Large Language Models' maximum input length (512).

One current bottleneck is the low relevant context extraction accuracy. As Table 1 shows, with fully correct context extraction (Oracle), context selection module can enhance baseline models by two times more than now. Suffering from the lack of annotated labels, current works all select contexts via unsupervised methods (like selecting 512 tokens surrounding the image, and sentences related to the image). However, human-defined standards are subjective and not necessarily consistent with the implicit pattern of judging if a sentence is relevant to the image. Supplementing the image-relevance labels at the sentence level and formulating context selection as a sequence labelling task may be a solution.

Besides, visual information can assist the relevant context extraction. Zhou et al. (2022) simultaneously extracted entities from the image and the entities that cooccur in the same sentence from the article. However, the improvement is limited since merely selecting cooccurred entities introduces too much noise. Leveraging knowledge graphs for additional filtering may help to obtain more precise candidate named entities.

		BLEU-4	ROUGE	CIDER	Named Entity	
					Precision	Recall
GoodNews	Tell (Tran et al., 2020)	6.0	21.4	53.8	22.2	18.7
	Tell (Oracle) (Zhou et al., 2022)	7.2(↑1.2)	24.2(↑3.2)	67.4(↑14.4)	30.0(↑7.8)	24.5(↑5.8)
	Tell (Auto) (Zhou et al., 2022)	6.3(↑0.3)	22.4(↑1.0)	60.3(↑6.5)	24.2(↑2.0)	20.9(↑2.2)
NYTimes800K	Tell(Tran et al., 2020)	6.3	21.7	54.4	24.6	22.2
	Tell (Oracle) (Zhou et al., 2022)	10.3(↑4.0)	27.6(↑5.9)	84.5(↑30.1)	39.8(↑15.2)	35.5(↑13.3)
	Tell (Auto) (Zhou et al., 2022)	7.0(↑0.7)	22.9(↑1.2)	63.6(↑9.2)	29.8(↑5.2)	25.9(↑3.7)

Table 1: Performance of Tell (Tran et al., 2020), Tell with Oracle optimal context selection (Zhou et al., 2022), and Tell with Automatic context selection (Zhou et al., 2022) on GoodNews and NYTimes800k. Oracle optimal context selection means conducting the experiments with manually extracted context which contains the named entities appearing in ground truth. Auto means extracting relevant context via proposed strategy automatically.

3.2 Image Feature Extraction

Image Feature Extraction is one significant challenge for news captioning tasks. To demonstrate an event, news images always contain multiple objects, which are related to some named entities in articles. So, it is essential to detect the important regions with objects and extract the features from the image.

Capturing all crucial visual information is challenging. Tran et al. (2020) improved previous methods (directly inputting the image to visual feature extractors) by extracting faces and objects exclusively as two extra input features to emphasize the Named Entities in the image. However, it still fails to consider the relationships between objects. As news images aim to describe an event, relationships between objects are often essential. Based on the outcomes of Scene Graph Generation (e.g. Motifs[14], VCTree[15]), news image captioning may be benefited by capturing the complex relationships in the image.[16]

Furthermore, applying APIs (like Google Cloud, Amazon Azure) to recognize the entities in the image directly can also help to supplement information not covered by the given article.

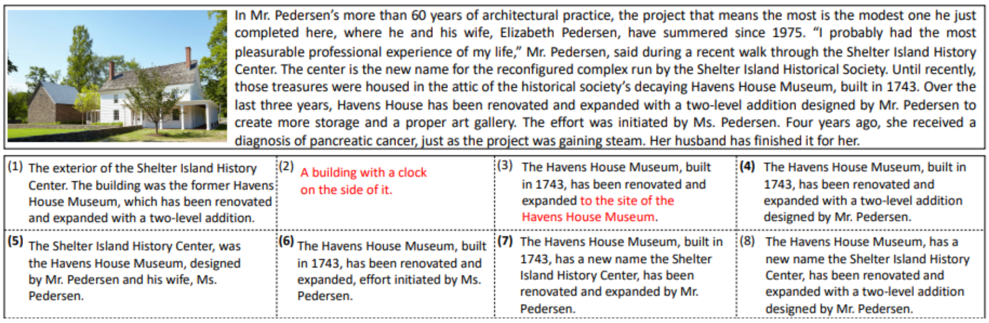
3.3 Caption Generation

Different from standard captioning tasks, news captioning has two extra challenges while generating the captions: 1) combining the captions and the extracted Named Entities, and 2) including some necessary elements (like who, when, where, etc.) to make the news image captions specific.

For now, templated-based methods and end-to-end methods are two categories of methods. Template-based methods mean generating a template first and then inserting named entities into the template, and end-to-end methods mean directly generating a caption according to the text and image input. Templated-based methods can help ensure the existence of necessary elements, but Named Entities insertion task is quite challenging. Captions generated by end-to-end methods are linguistically rich but do not always contain all necessary elements.

To benefit from both approaches, Yang et al. (2022) proposed a template-guided end-to-end model (JoGANIC[10]). It begins by identifying which components should be contained and then utilizes end-to-end methods to generate coherent captions under the guidance of determined composition. Despite currently boasting the highest performance, it shares a common issue encountered by all existing models: an excessive emphasis on the input text at the expense of the image. As a result, the generated captions are more like summarizing the entire news story rather than depicting the image. (Figure 1)

A potential cause of the problem is that simple concatenation cannot effectively combine the text and visual features, and further research on explaining how this multi-modal model works is required to validate the hypothesis. Besides, we can propose a more interpretable and feasible pipeline here: generating a caption of the image first as the basis of the caption and then supplementing details mentioned in the given article to the generated caption gradually. This pipeline is worth exploring because 1) it mimics humans' behaviours of generating a news image caption and may lead to better performance. 2) The intermediate results help to explain the model.



In Mr. Pedersen's more than 60 years of architectural practice, the project that means the most is the modest one he just completed here, where he and his wife, Elizabeth Pedersen, have summered since 1975. "I probably had the most pleasurable professional experience of my life," Mr. Pedersen, said during a recent walk through the Shelter Island History Center. The center is the new name for the reconfigured complex run by the Shelter Island Historical Society. Until recently, those treasures were housed in the attic of the historical society's decaying Havens House Museum, built in 1743. Over the last three years, Havens House has been renovated and expanded with a two-level addition designed by Mr. Pedersen to create more storage and a proper art gallery. The effort was initiated by Ms. Pedersen. Four years ago, she received a diagnosis of pancreatic cancer, just as the project was gaining steam. Her husband has finished it for her.

(1) The exterior of the Shelter Island History Center. The building was the former Havens House Museum, which has been renovated and expanded with a two-level addition.	(2) A building with a clock on the side of it.	(3) The Havens House Museum, built in 1743, has been renovated and expanded to the site of the Havens House Museum.	(4) The Havens House Museum, built in 1743, has been renovated and expanded with a two-level addition designed by Mr. Pedersen.
(5) The Shelter Island History Center, was the Havens House Museum, designed by Mr. Pedersen and his wife, Ms. Pedersen.	(6) The Havens House Museum, built in 1743, has been renovated and expanded, effort initiated by Ms. Pedersen.	(7) The Havens House Museum, built in 1743, has a new name the Shelter Island History Center, has been renovated and expanded by Mr. Pedersen.	(8) The Havens House Museum, has a new name the Shelter Island History Center, has been renovated and expanded with a two-level addition designed by Mr. Pedersen.

Figure 1 An example of news caption generation. The captions are generated by: (1) human (ground truth caption). (2) conventional image captioning model SAT. (3) Tell. (4) JoGANIC. (5) JoGANIC+NEE. (6) JoGANIC+MSTR. JoGANIC+MSTR+NEE (7) auto, (8) oracle, with template. The wrong information is highlighted as red.

4. Evaluation Metric

Existing research mainly uses automatic metrics and human evaluation to assess models' performances. However, human evaluation carries more credibility than automatic metrics in this open-ended task, despite being more time-consuming and labor-intensive. Besides, results from Biten et al. (2019)[7] show a significant drop from the template generation task to the news captioning task, which indicates that named entity insertion is challenging for news image captioning. Hence, the precision and recall for named entity extraction are also suitable evaluation metrics for this task.

Furthermore, human-annotated ground truths are not the only standard in this task because: 1) what components a caption should contain largely depends on the authors' writing styles[10], which cannot be judged as right or wrong. 2) Any caption that effectively summarizes the context of the image can be considered satisfactory. Thus, as long as the length of a generated caption is within a reasonable scope, it should not be penalized for containing additional components and enhancing its clarity. In future, we need to consider two standards while proposing a new metric: 1) whether the generated caption covers most information points in the ground truths (recall), and 2) the consistency of the caption with the facts indicated by the given news articles (correctness).

5. Future Direction

After analyzing previous research on news image captioning, the preceding sections have suggested directions for further study regarding datasets, methodology and metrics. However, there remain unexplored directions that differ from all existing works and are worth investigating.

To begin with, the formulation of the task needs to be extended. All existing works are limited to extracting Named Entities solely from the given articles. However, no matter in the datasets or the real-world scenarios, a large proportion of the Named Entities present in the ground truths are not mentioned in the accompanying articles. Hence, to make this research more practical and meaningful, it is advisable to expand the scope of the Named Entities source beyond the provided articles, like detecting directly from the image or applying knowledge graphs to find related entities.

In addition, a bottleneck for all existing news image captioning methods is that they tend to prioritize the input text over the input images (3.3). To understand the underlying reasons, it is necessary to conduct research that explains how the models work (especially at the layer where multi-modal features are combined), which can also inspire the development of a more effective approach to aligning text features with visual features.

Furthermore, the task could leverage the advancements of Large Language Models, which excel in generating linguistically rich descriptions and performing basic reasoning. While there have been some studies on generating image captions using GPT, it is crucial to discuss the following aspects for this task specifically: 1) exploring effective methods to integrate Named Entities with image captions, 2) devising a pipeline that emulates human behaviour to generate captions for news images (mentioned in 3.3) and makes the process more interpretable.

Reference

[1] Shu K, Sliva A, Wang S, et al. Fake news detection on social media: A data mining perspective[J]. ACM SIGKDD explorations newsletter, 2017, 19(1): 22-36.

[2] Nallapati R, Zhou B, Gulcehre C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond[J]. arXiv preprint arXiv:1602.06023, 2016.

[3] Farhadi A, Hejrati M, Sadeghi M A, et al. Every picture tells a story: Generating sentences from images[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2010: 15-29.

[4] Suhr A, Zhou S, Zhang A, et al. A corpus for reasoning about natural language grounded in photographs[J]. arXiv preprint arXiv:1811.00491, 2018.

[5] Feng Y, Lapata M. Automatic caption generation for news images[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(4): 797-812.

[6] Ramisa A, Yan F, Moreno-Noguer F, et al. Breakingnews: Article annotation by image and text processing[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(5): 1072-1085.

[7] Biten A F, Gomez L, Rusinol M, et al. Good news, everyone! context driven entity-aware captioning for news images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12466-12475.

[8] Tran A, Mathews A, Xie L. Transform and tell: Entity-aware news image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13035-13045.

[9] Liu F, Wang Y, Wang T, et al. Visual news: Benchmark and challenges in news image captioning[J]. arXiv preprint arXiv:2010.03743, 2020.

[10] Yang X, Karaman S, Tetreault J, et al. Journalistic Guidelines Aware News Image Captioning[J]. arXiv preprint arXiv:2109.02865, 2021.

[11] Zhou M, Luo G, Rohrbach A, et al. Focus! Relevant and Sufficient Context Selection for News Image Captioning[J]. arXiv preprint arXiv:2212.00843, 2022.

[12] Hu A, Chen S, Jin Q. Icecap: Information concentrated entity-aware image captioning[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 4217-4225.

[13] Da J, Forbes M, Zellers R, et al. Edited Media Understanding: Reasoning About Implications of Manipulated Images[J]. arXiv preprint arXiv:2012.04726, 2020.

[14] Zellers R, Yatskar M, Thomson S, et al. Neural motifs: Scene graph parsing with global context[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5831-5840.

[15] Tang K, Zhang H, Wu B, et al. Learning to compose dynamic tree structures for visual contexts[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6619-6628.

[16] Milewski V, Moens M F, Calixto I. Are scene graphs good enough to improve image captioning?[J]. arXiv preprint arXiv:2009.12313, 2020.

[17] Zhao S, Sharma P, Levinboim T, et al. Informative image captioning with external sources of information[J]. arXiv preprint arXiv:1906.08876, 2019.