Reproducing and extending ”Pragmatic Issue-Sensitive Image Captioning”

Jan-Felix Klumpp, Daniela Verratti Souto

April 11, 2023

Abstract

We reproduce and extend a paper by Nie et al. [2020] on issue-sensitive image captioning (ISIC). While an entire reproduction is not possible due to missing detailed information regarding some aspects, and while the results do not exactly match the original results due to unknown reasons, they confirm the original authors' conclusion that RSA-based issue-sensitive image captioners produce pragmatically more adequate captions than comparable issue-insensitive captioners. We further show that these conclusions can be generalized onto two issues as well.

1 Introduction

In their paper, Nie et al. [2020] define and evaluate issue-sensitive speaker agents for image captioning. Issue sensitivity means that the generated caption differs according to an issue. An issue is understood to be a partition of a set of images according to a relevant attribute (e.g., a set of images subdivided into two cells; one containing birds that have a curved beak and another one with birds that do not).

If successful, the issue-sensitive speaker agent would produce a caption that addresses exactly this issue and no other ones. Nie et al. [2020] base their speaker agents on the Rational Speech Act model (RSA). In RSA, the choice of an utterance by the speaker is not only dependent on the correctness of the description, but also on the informativity and discriminative power of the utterance given a set of possible situations. For caption generation, this means that the caption should describe a specific feature of the shown object, while leaving out others that are not considered relevant.

As an example, given a set of images of birds (the CUB birds data, which is used both in Nie et al. [2020]'s paper as well as in our reproduction), an issue-sensitive speaker should ideally be able to produce a caption that describes a specific property of the bird (an issue), e.g. the color of the wings, while leaving out other properties considered irrelevant. As it turns out, this goal is very difficult to achieve. However, while no speaker agent considered in Nie et al. [2020] addresses all and only the relevant issues, their issue-sensitive speaker agents considerably outperform the issue-insensitive ones.

1.1 Speaker agents

Nie et al. [2020] compare five speaker agents:

1. $S0$: This is a literal speaker based on the GVE (Generating Visual Explanations) model [Hendricks et al., 2016]. Based on the LRCN model [Donahue et al., 2015], GVE uses neural networks trained to produce captions for a given image. This is done by taking the LSTM-transformed caption embeddings and appending them to the image features, which have previously passed through a linear and a ReLu layer and been concatenated with class labels for the training captions.

   This model takes into account neither the discriminative power of the caption in general (as RSA-based speaker agents do), nor any specific issue.

2. $S1$: This is a issue-insensitive pragmatic RSA speaker. It takes the output probability distribution over possible captions from the $S0$ speaker but modifies the probability distribution according to their discriminative usefulness to a pragmatic listener.

3. $S1^C$: This is an issue-sensitive RSA speaker agent, which maximizes the discriminative power according to the relevant issue. This way, $S1^C$ will produce captions that help the listener select the cell in the partition that contains the target image.

4. $S1^{CH}$: In producing captions that unequivocally pick out the correct cell, $S1^C$ might closely describe a specific image that shares a cell with the target, but which contains characteristics that are in fact not present in the desired image. Additionally, it might include information that is true about the target image but that is not relevant to the issue at hand. To avoid both of these problems, $S1^{CH}$ penalizes uneven probability distributions over the images in the correct cell by using a utility function based on information-theoretical entropy.

5. $S0$ Avg: To obtain roughly issue-sensitive captions without the need to apply RSA principles, they propose a literal speaker $S0$ Avg that takes the average of the images in the current cell. This should help evaluate whether there is actual value in extending a literal speaker with the RSA model or whether similar results can be obtained in a simpler manner.

1.2 S0 Average speaker agent

In the code provided by Nie et al. [2020], no $S0$ Avg is implemented. In our repository we include two possibilities for implementation.

The first one consisted of obtaining all images in the target image cell and calculating their average for each issue; this average tensor was then used as an input for the semantic speaker. However, when generating captions with this agent, we noticed that all captions for all images were exactly the same despite the inputted average being different for each issue. Upon further inspection, we determined that this is usually the case when averaging more than three image tensors. Since we do not consider this to be a sensible cell size for capturing issues, we discarded this code. It can still be found in a comment, as we believe this attempt is quite true to the approach described in the paper.

On the other hand, a simpler implementation, consisting of passing all the image features of the similar cell (without averaging) as inputs to the literal speaker, yields similar results to those reported by Nie et al. [2020]. These results are what we report in this document and what is run in the current version of our caption generation code.

2 Technical details

The code was obtained from the repository that accompanies the original paper and later modified and extended by us. The data can be downloaded by following the instructions provided in our repository (`https://github.com/codes-witch/NLG_Klumpp_Verratti.git`).

We reproduced the whole training process for CUB data, using the same parameters as described in the original paper. However, to avoid an error in evaluation arising within the pycocoevalcap.eval module, we did not obtain evaluation metrics. By generating and comparing captions for the $S0$ condition, we could make sure that the results of our training are equivalent to the training checkpoint provided in the original repository. The latter was then used to generate all captions for issue alignment evaluation as well as for the two-issue experiment.

We did not reproduce the training process for COCO data, which is not relevant to the experiments done here.

3 CUB Experiments

Nie et al. evaluate the quality of their issue-sensitive speaker agents using two measures, Attribute Coverage and Issue Alignment, concluding that indeed issue-sensitive speaker agents perform better than speaker-insensitive ones. As it turned out, with the information given in their paper and repository, only Issue Alignment can be reproduced, though even there, some aspects are not totally unambiguous. Our results approximate those reported by Nie et al. [2020] but do not exactly match them. The reason for this is not totally clear. Nevertheless, we can confirm their conclusion that issue-sensitive speaker agents perform better on the experiment task than issue-insensitive ones.

3.1 Attribute Coverage

Attribute coverage is a measure describing which of the attributes discernible from an image are featured in the generated captions. Nie et al. evaluate attribute coverage for each image, concatenating the captions for all resolvable issues. This means that the measure is not evaluated individually for each issue, but instead looking at all issue conditions simultaneously. Nevertheless, Nie et al. expect (and observe) a better performance of their issue-sensitive captioners, because by generating different captions for each issue, issue-sensitive captions cover a wider range of attributes overall.

The evaluation of attribute coverage needs a definition of the mappings between keywords (i.e., the expressions that actually occur in the caption) and attributes (i.e., more abstract descriptions). However, Nie et al. [2020] only report which keywords they regard as resolving each issue, but not how they resolve it. For example, they provide a list of words describing beak shapes, but not the mapping between those natural-language keywords to the controlled language of the abstract CUB descriptions. While their KeywordExtractor class provides a method to obtain keywords associated with each attribute that could be used to solve this problem, it is very unreliable and probably not what was used in the original paper. Hence, given the lack of information on the mapping between natural-language expressions and abstract attributes, a reproduction of Nie et al. [2020]'s results for attribute coverage is not possible.

3.2 Issue Alignment

The second measure evaluated in Nie et al. [2020]'s paper is Issue Alignment. This measure looks at whether each resolvable issue is actually resolved in the associated caption. Nie et al. [2020] expect and observe the issue-sensitive speaker agents to perform better than the issue-insensitive $S0$ speaker here, from which they conclude that issue-sensitive speakers produce captions which address the target issue better than issue-insensitive speakers.

The definition of Issue Alignment given in Nie et al. [2020] is slightly ambiguous in two ways: First, they state that they measure whether ”the produced caption precisely resolve[s]” a given issue (p. 1930). ”Precisely” here could either refer to whether the issue is resolved in a precise way, i.e. correctly, or whether it is exactly the target issue that is resolved. Even though it seems the less intuitive measure for judging the quality of speaker agents, we assume the second option for the following reasons:

1. The description of the measure of ”precision” in this context contrasts resolving the target issue with resolving other issues. It does not, however, contrast resolving the target issue correctly with resolving the target issue incorrectly.

2. The code provided by Nie et al. [2020] allows easily to identify whether an issue is resolved. It is, however, not possible to identify whether an issue is resolved correctly with the given information (for the reasons given in the section on Attribute Coverage above).

3. The scores for the issue-insensitive $S0$ speaker agent for issue alignment and attribute coverage are different in Nie et al. [2020], which should not be the case if only correctly resolved issues are counted (under the assumption that averaging is done over images, and not over image-issue pairs, see below).

Second, Nie et al. [2020] do not report how their scores are calculated. Precision and recall could either be calculated for each image and then averaged, or they could be derived directly from the number of resolved image-issue pairs. This becomes relevant where (as it is the case for the CUB dataset) the number of resolvable issues differs between images. Since it is not possible to resolve this ambiguity, both methods of calculation are included for the following results.

3.3 Results

Table 1 gives the issue alignment scores for all five speaker agents in our reproduction (for both averaging over images and direct calculation from resolved image-issue pairs), juxtaposed to Nie et al. [2020]'s original results.

As can be seen, neither of the results corresponds exactly to Nie et al. [2020]'s scores, and in the case of $S0$ recall the difference is rather large, but we cannot explain these discrepancies. One possibility

Condition	for images	for image-issue pairs	original paper
Precision			
$S0$	11.4	10.6	10.5
$S0$ Avg	11.2	10.8	12.1
$S1$	11.8	11.2	11.2
$S1^C$	20.3	18.8	18.7
$S1^{CH}$	15.4	14.6	16.6
Recall			
$S0$	29.6	28.5	21.1
$S0$ Avg	29.3	28.6	29.0
$S1$	22.2	21.8	21.7
$S1^C$	43.4	42.7	42.5
$S1^{CH}$	43.3	42.4	46.6
F_1			
$S0$	15.7	15.5	15.5
$S0$ Avg	16.1	15.7	17.0
$S1$	14.4	14.8	14.8
$S1^C$	26.8	26.1	25.9
$S1^{CH}$	22.4	21.7	24.5

Table 1: Issue alignment scores for our reproduction and Nie et al. [2020]'s original paper.

could be the size of the context window. This size is not explicitly stated in the paper, but as a default is set to 3 in their code. We obtained the results above using the same setting, but using different window sizes slightly alters the scores. However, no window size between 3 and 6 produces exactly the same values as they report, so the question of how these discrepancies arise remains unanswered.

Interestingly, in our reproduction, $S0$ Avg did not perform better than $S0$, and $S1$ even performed worse (except for precision). Similarly, we did not find any improvement in recall from $S1^C$ to $S1^{CH}$, though in this case the difference was not very large in the original paper either.

Nevertheless, our results unambiguously confirm Nie et al. [2020]'s conclusion that in general, issue-sensitive speaker agents as described in their paper perform better on this image-captioning task than the issue-insensitive literal speaker.

4 Two-Issue Image Captioning

4.1 Idea

Nie et al. [2020] describe situations where it could be important to resolve a specific issue in a caption while ignoring others. For example, they note that in a description of a sports event, a caption should usually identify the player in the picture, but rather not comment on their ethnicity. The same reasoning, however, can be easily extended to situations where there is a larger number of issues to resolve: for example, we might want not only to identify the player by name, but also to mention the team in which they play, or there might be more than one player to be identified in the picture. In such cases, an image captioner should be able to address any number of relevant issues. To be able to estimate whether an RSA-based approach can be used to solve this task, we extend Nie et al. [2020]'s speaker agents to address two issues. If such a two-issue issue-sensitive speaker agent performs better than an issue-insensitive speaker agent, we take it as an indication that Nie et al. [2020]'s approach can indeed be extended to more than one issue.

Nie et al. [2020] define two utility functions, U_1 and U_2. U_1^C is defined as $U_1^C(i, w, C) = log(\sum_{i\prime \in \mathcal{I}} \delta_{[C(i)=C(i\prime)]} L_1(i\prime|w))$, where $\delta_{[C(i)=C(i\prime)]}$ returns 1 if i and $i\prime$ are in the same cell of the partition C, and 0 otherwise. If we have two issues C and D, we can simply replace $\delta_{[C(i)=C(i\prime)]}$ by $\delta_{[C(i)=C(i\prime)]}\delta_{[D(i)=D(i\prime)]}$. This means that i and $i\prime$ have to be in the same cell for both partitions to make the term 1, else it is 0. The same is also done in the second utility function U_2.

4.2 Implementation

The file two_issues.py contains the additional code that is needed to generate and evaluate captions for two issues.

The easiest approach to implementation is to re-use most of the code for one-issue caption generation, only changing the partition construction. There are, however, several imaginable ways to do so. Intuitively, starting from the formula above (with C and D being our issue partitions), we could try to split all images into the intersection of the cells with the target image in both partitions vs. the union of distractors (such that in the new partition, there is one cell with all images that are in the same cell as the target in both C and D, and one cell with all others). However, in practice, this does not work as expected: since in the actual implementation, images are **sampled** from the partitions, deriving the new distractor cell from a union of the distractors in C and D means that some images that actually are in the same cell as the target in C and D come to land in the distractors because they were sampled only once. Hence, instead of this naive approach, we implemented two alternatives for defining the distractors:

1. In the "narrow" definition of distractors, it is not the union, but the intersection that is used to sort images into the distractor cell. This avoids misassignment of images that are sampled only once. It also leads to an approximately equal size of both cells of the resulting partition. On the other hand, however, it means that many images are completely disregarded and not assigned to any cell.

2. In the "wide" definition of distractors, the new distractor set is the union of the following: 1. the intersection of the distractors of C and D, 2. the intersection of the target cell of C and the distractors of D, 3. the intersection of the distractors of C and the target cell of D. This corresponds closest to the intended formula: By considering all and only images that are sampled for both C and D, problems arising from the sampling process are avoided without making further changes to the partition construction.

Both these approaches ("narrow" and "wide") were evaluated for issue alignment.

4.3 Results

Table 2 gives the issue alignment results for the conditions $S0$, $S1$, $S1^C$ (narrow and wide), and $S1^{CH}$ (narrow and wide), for both averaging over images and equal weighting of image-issue-issue triples. Recall is divided into "strict" recall, where both issues have to be resolved, and "lax" recall, where at least one has to be resolved. As can be expected, precision and lax recall are higher for two issues than for one issue, while strict recall is lower. However, what is more important are the differences between the issue-sensitive and issue-insensitive conditions. Here, just as for one issue, the issue-sensitive captioners perform better than the issue-insensitive ones, from which we can conclude that indeed Nie et al. [2020]'s approach can be extended to more than one issue.

The results also indicate that the narrow definition of distractors (where the distractor cell only contains images that are different from the target image on both issues) performs somewhat better than the wide definition (where the distractor cell also includes images that correspond to the target on exactly one issue).

It has to be noted that the setting of parameters was the same as for the one-issue experiments (for the sake of better comparability), i.e. rationality was 3 for $S1$ and 10 for the other captioners, and the entropy parameter was 0.4. This led to sometimes ungrammatical captions for $S1$ and $S1^C$ (but not for $S1^{CH}$), which means that by adjusting the rationality parameter, some further improvement might be possible. None of this invalidates the general conclusions, since no grammaticality problems were encountered with $S0$, $S0$ Avg or $S1^{CH}$, and here the RSA-based issue-sensitive captioner $S1^{CH}$ clearly performs better than both $S0$ and $S0$ Avg.

5 Conclusion

After having carried out these experiments, we come to the conclusion that Nie et al. [2020]'s results are partially reproducible. Despite the fact that the code they provide to accompany the paper does not

	by image			by image-issue-issue triple		
Condition	Precision	Recall, strict	Recall, lax	Precision	Recall, strict	Recall, lax
$S0$	22.8	8.5	50.6	20.2	7.3	47.6
$S0Avg$	21.6	7.8	49.0	20.7	7.2	47.5
$S1$	22.9	6.2	41.2	20.8	5.5	39.8
$S1^C$, narrow	29.6	10.6	54.2	27.2	9.8	52.7
$S1^C$, wide	28.2	7.6	42.8	26.4	7.2	42.2
$S1^{CH}$, narrow	27.6	13.3	63.3	25.3	12.2	61.0
$S1^{CH}$, wide	27.9	12.6	61.0	25.5	11.5	58.8

Table 2: Issue Alignment scores for two issues.

contain all the information necessary for a full reproduction and although some of the explanations as to how their scores were calculated were missing or ambiguous, the results we obtained broadly resemble the ones they report and align with their conclusion that an issue-sensitive image captioner based on RSA principles performs better than other issue-insensitive neural image captioners.

Furthermore, we have determined that the same can also be said about two-issue issue-sensitive captioners, as the captions generated using Nie et al. [2020]'s method also performed better in terms of issue alignment when two issues are considered.

References

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 3–19. Springer, 2016.

Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. Pragmatic issue-sensitive image captioning. *arXiv preprint arXiv:2004.14451*, 2020.