Pre-gen Metrics: Predicting Caption Quality Metrics Without Generating Captions

Marc Tanti(✉), Albert Gatt, and Adrian Muscat

University of Malta, Msida 2080, MSD, Malta
{marc.tanti.06,albert.gatt,adrian.muscat}@um.edu.mt

Abstract. Image caption generation systems are typically evaluated against reference outputs. We show that it is possible to predict output quality without generating the captions, based on the probability assigned by the neural model to the reference captions. Such pre-gen metrics are strongly correlated to standard evaluation metrics.

Keywords: Image captioning · Neural architectures
Evaluation metrics

1 Introduction

Automatic metrics for image description generation (IDG) compare c, a generated caption, to a set of reference sentences, $R_1 ... R_n$. We therefore refer to these as **post-gen**(eration) metrics. In most neural IDG architectures generation is performed by an algorithm such as beam search that samples the vocabulary at every timestep, selecting a likely next word after a given sentence prefix (according to the neural network) and attaching it to the end of the prefix, and repeating this procedure until the entire caption is produced. Given that the output thus generated is evaluated against a gold standard, post-gen metrics actually evaluate the neural network's ability to predict the words in the reference captions given an image. Unfortunately, generating sentences is a time consuming process due to the fact that every word in a sentence requires its own forward pass through the neural network. This means that generating a 20-word sentence requires calling the neural network 20 times. As an indicative example, it takes 20.8 min to generate captions for every image in the MSCOCO test set on a standard hardware setup (GeForce GTX 760) using a beam width of just 1.

Our question is whether a system's performance can be assessed *prior* to the generation step, by exploiting the fact that the output is ultimately based on this core sampling mechanism. We envisage a scenario in which a neural caption generator is evaluated based on the extent to which its estimated softmax probabilities over the vocabulary maximise the probability of the words in the reference sentences $R_1 ... R_n$. We refer to this as a **pre-gen**(eration) evaluation metric, as it can be computed prior to generating any captions. A well-known

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11132, pp. 114–123, 2019.
https://doi.org/10.1007/978-3-030-11018-5_10

example of a pre-gen metric is language model perplexity although, as we show below, this metric is not the best pre-gen candidate in terms of its correlation to standard evaluation measures for IDG systems.

From a development perspective, the advantage of a pre-gen metric lies in that all the word probabilities in a sentence are immediately available to the network in one forward pass, whereas a post-gen metric can only be computed following a relatively expensive process of word-by-word generation requiring repeated calls to a neural network. To return to the earlier example, on the same hardware setup it only takes 28 sec to compute model perplexity.

Thus, if pre-gen metrics can be shown to correlate strongly with established post-gen metrics, they could serve as a proxy for such metrics. This would speed up processes requiring repeated caption quality measurement such as during hyperparameter tuning.

Finally, from a theoretical and empirical perspective, if caption quality, as measured by one or more post-gen metric(s), can be predicted prior to generation, this would shed further light on the underlying reasons for the observed correlations of such metrics [14].

All code used in these experiments is publicly available.[1]

The rest of this paper is organised as follows; background on metrics is covered in Sect. 2, the methodology and experimental setup in Sect. 3, and the results are given in Sect. 4; the paper is concluded in Sect. 5.

2 Background: *Post-gen* Metrics for Image Captioning

In IDG, automatic metrics originally developed for Machine Translation or Summarisation, such as BLEU [21], ROUGE [18], and METEOR [2], were initially adopted, followed by metrics specifically designed for image description, notably CIDEr [24] and SPICE [1]. Lately, Word Mover's Distance (WMD) [17], originally from the document similarity community, has also been suggested for IDG [14]. Like BLEU, ROUGE and METEOR, CIDEr makes use of n-gram similarities, while WMD measures the semantic distance between texts on the basis of word2vec [20] embeddings. All of these metrics are purely linguistically informed. By contrast, SPICE computes similarity between sentences from scene graphs [13], obtained by parsing reference sentences. This method is also linguistically informed; however the intuition behind it is that the human authored sentences should be an accurate reflection of image content.

A typical IDG experiment reports several post-gen metrics. One reason is that the metrics correlate differently with human judgments, depending on task and dataset [3], echoing similar findings in other areas of NLP [4–7,10,11,22,26]. Thus, BLEU, METEOR, and ROUGE correlate weakly [12,15,16] and yield different system rankings compared to human judgments [25]. METEOR has a reportedly higher correlation than BLEU/ROUGE [8,9], with stronger relationships reported for CIDEr [24] and SPICE [1]. Meta-evaluation of the ability of metrics to discriminate between captions have also been somewhat inconsistent [14,24].

[1] See: https://github.com/mtanti/pregen-metrics.

The extent to which post-gen metrics correlate with each other also varies, with stronger relationships among those based on n-grams on the one hand, and more semantically-oriented ones on the other [14], suggesting that these groups assess complementary aspects of quality, and partially explaining their variable relationship to human judgments in addition to variations due to dataset.

For neural IDG architectures, post-gen metrics have one fundamental property in common: they compare reference outputs to generated sentences which are based on sampling at each time-step from a probability distribution. Our hypothesis is that it is possible to exploit this, using the probability distribution itself to directly estimate the quality of captions, prior to generation.

3 Pre-gen Metrics

Given a prefix, a neural caption generator predicts the next word by sampling from the softmax's probabilities estimated over the vocabulary. Let R be a reference caption of length m. Given a prefix $R^{0\cdots k}$ (where R^0 is the start token), $k \leq m$, a neural caption generator can be used to estimate the probability of the next word (or the end token) in the reference caption, R^{k+1}. The intuition underlying pre-gen metrics is that the higher the estimated probability of R^{k+1}, for all $k \leq m$, the more likely it is that the generator will approximate the reference caption. Note that the idea is to estimate the probability of *reference* captions based on a trained IDG model.

Pre-gen metrics produce a score by aggregating the word probabilities predicted by the generator for all reference captions (combined with their respective image) over prefixes of different lengths. To find the best way to do this, we define a search space by setting options at four different algorithmic steps which we call 'tiers'. Each tier represents a function and the composition of all four tiers together constitutes a pre-gen function. We try several different options for each tier in order to find the best pre-gen function. Figure 1 shows an example of how tiers form a pre-gen function.

Given a set of images with their corresponding reference captions, the process starts by computing each reference caption's individual words probabilities (given the image) according to the model. Note that the model may not predict every word in a reference caption as the most likely in the vocabulary.

The first tier is a filter that selects which predicted word probabilities should be considered in the next tier. We consider three possible filters: (a) the filter *none* passes all probabilities; (b) *filter0* filters out the word probabilities that are not ranked as most probable in the vocabulary by the model, i.e are not predicted to be maximally probable continuations of the current prefix; and (c) *prefix0* selects the longest prefix of the caption such that the model predicts all words in the prefix as being the most likely in the vocabulary.

At the second tier, we aggregate the selected word probabilities in each reference sentence into a single score for each sentence. We define four possible functions: (a) *prob* multiplies all probabilities; (b) *pplx* computes the perplexity; (c) *count* counts the number of word probabilities that were selected in the first

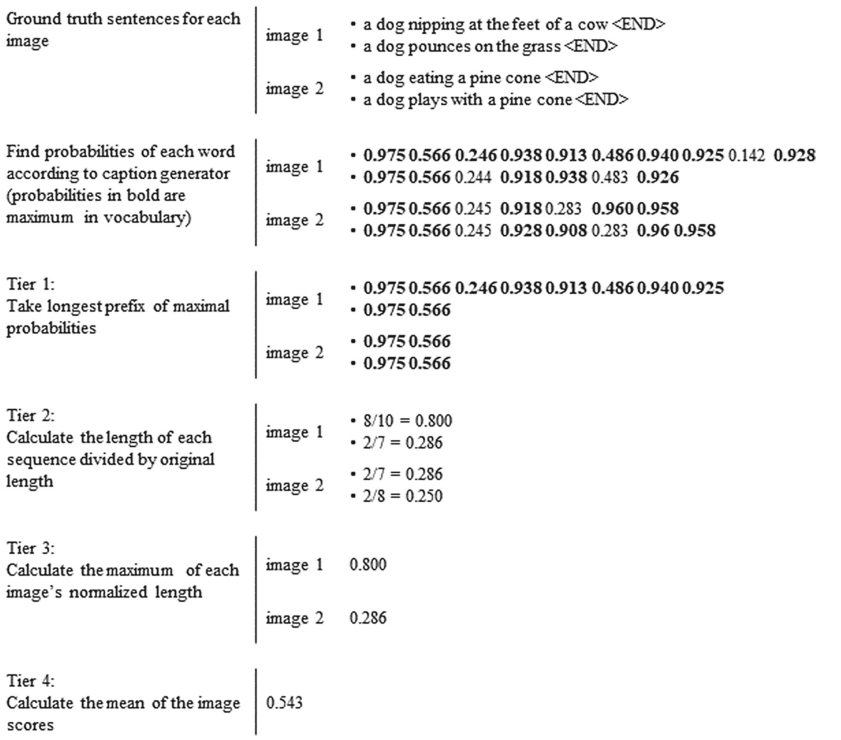


Fig. 1. An example illustrating how tiers work. This illustration shows the best pre-gen metric found: *mean_max_normcount_prefix0*.

tier; and (d) *normcount* normalises *count* by the total number of words in the reference sentence.

The third tier aggregates the scores obtained for all reference sentences into a single score for each image. We explore six possibilities: (a) *sum*; (b) *mean*; (c) *median*; (d) *geomean*, the geometric mean; (e) *max*; and (f) *min*. We also consider (g) *join*, whereby all the image-sentence scores are joined into a single list without aggregation so that they are all aggregated together in the next tier.

The fourth tier aggregates the image scores into a single dataset score, which is the final pre-gen score of the caption generator. For this aggregation, we use the same six functions as in the previous tier (excluding *join*).

The above possibilities result in 504 unique combinations. In what follows, we adopt the convention of denoting a pre-gen metric by the sequence of function names that compose it, starting from tier four e.g. *mean_max_normcount_prefix0*. In our experiments, we compute all of these different combinations and compare their predictions to standard post-gen metrics, namely METEOR, CIDEr, SPICE, and WMD. All metrics except WMD were computed using the MSCOCO

Evaluation toolkit[2]. Since the toolkit does not include WMD, we created a fork of the repository that includes it.[3]

3.1 Experimental Setup

For our experiments, we used a variety of pre-trained neural caption generators (36 in all) from [23].[4] These models are based on four different caption generator architectures. Each was trained and tested over three runs on Flickr8k [12], Flickr30k [27], and MSCOCO [19]. The four architectures differ in terms of how the CNN image encoder is combined with the RNN: `init` architectures use the image vector as the initial hidden state of the RNN decoder; `pre` architectures treat the image vector as the first word of a caption; `par` architectures are trained on captions where each word is coupled with an image vector at each time-step; and `merge` architectures keep the image out of the the RNN entirely, merging the image vector with the RNN hidden state in a final feedforward layer, prior to prediction.

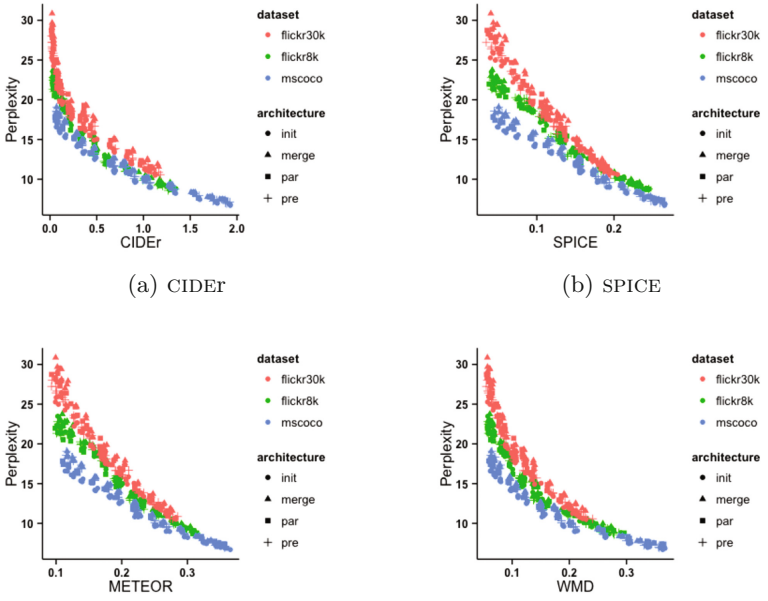
Since only the final trained versions of the models are available, there is a bias towards good quality post-gen metric results. This renders the values of the post-gen metrics rather similar and concentrated in a small range. A pre-gen metric is useful if it makes good predictions on models of any quality not just good ones. Rather than re-training all the models and saving the parameters at different intervals during training, we opted to stratify the dataset on the basis of how well each individual image is rated by the CIDEr metric.

We grouped images into the best and worst halves on the basis of the CIDEr score (since CIDEr is the post-gen metric that best correlates with the other post-gen metrics [14]) of their sentences as generated by a model. This creates two datasets, one where the model performs well and one where the model performs badly. We stratified the dataset into different numbers of equal parts and not just two, namely: 1 (whole), 2, 3, 4 and 5, resulting in a 15-fold increase over the original 36 averaged results and more importantly, over a wide dynamic range in CIDEr scores, which we required to study the correlation in between pre- and post-gen metrics.

4 Results

We evaluate the correlation between pre- and post-gen metrics using the Coefficient of Determination, or R^2, defined as the square of the Pearson correlation coefficient. The reason for this is twofold. First, R^2 reflects the magnitude of a correlation, irrespective of whether it is positive or negative (the pre-gen metrics based on perplexity would be expected to be negatively correlated with post-gen

[2] See: https://github.com/tylin/coco-caption.
[3] See: https://github.com/mtanti/coco-caption.
[4] See: https://github.com/mtanti/where-image2.



(a) CIDEr (b) SPICE

(c) METEOR (d) WMD

Fig. 2. Relationship between perplexity and post-gen metrics by dataset and architecture. The overall correlation has an R^2 of 0.76. (Best viewed in colour.) (Color figure online)

metrics). Second, given a linear model in which a pre-gen metric is used to predict the value on a post-gen metric, R^2 indicates the proportion of the variance in the latter that the pre-gen metric predicts.

As a baseline, we show the scatter plot for the relationship between language model perplexity and the post-gen metrics in Fig. 2. In terms of the description in the previous section, perplexity is defined as *geomean_join_pplx_none*. As can be seen, perplexity performs somewhat poorly on low scoring captions. Our question is whether a better pre-gen metric can be found.

For each of the 4 post-gen metrics, we identified the top 5 best correlated pre-gen metrics, based on the R^2 value computed over all the data (i.e. aggregating scores across architectures and datasets). The top 4 pre-gen metrics were the same for all post-gen metrics, namely:

1. *mean_max_normcount_prefix0*;
2. *mean_mean_normcount_prefix0*;
3. *mean_join_normcount_prefix0*;
4. *mean_sum_normcount_prefix0*

Note that all the best performing metrics are based on the variable *prefix0*. This is not surprising since when generating a sentence, it is probably the word with the maximum probability in the vocabulary which gets selected as a next

word in a prefix. On the other hand, the fifth most highly correlated pre-gen metric differed for each post-gen metric, as follows:

- CIDER: *mean_min_count_filter0*;
- METEOR: *mean_mean_count_prefix0*;
- SPICE: *geomean_min_pplx_filter0*;
- WMD: *mean_join_count_prefix0*

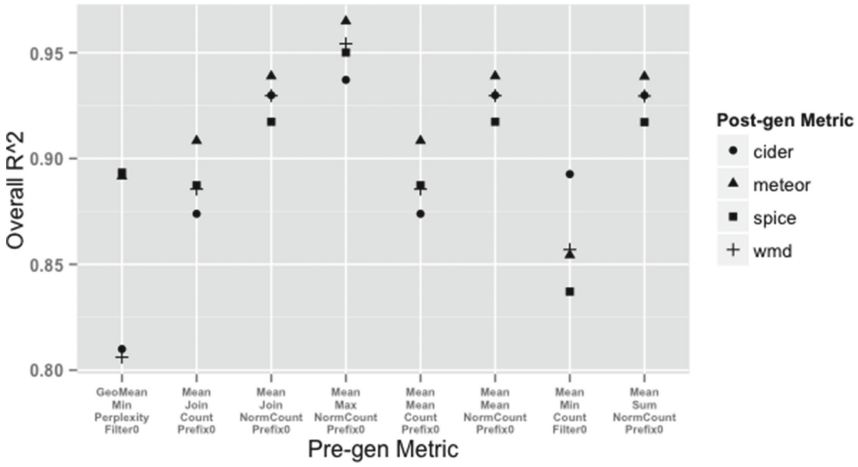
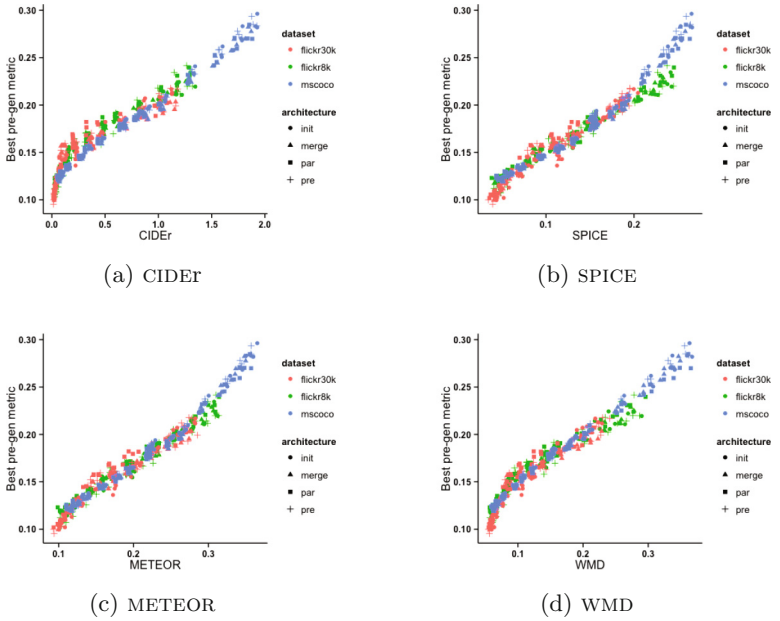


Fig. 3. Overall R^2 between the 4 post-gen metrics and their 5 most highly correlated pre-gen metrics. Scores average over architectures and datasets.

Figure 3 displays the relationship between these pre-gen metrics and the post-gen scores. Note that all R^2 scores are above 0.8, indicating a very strong correlation.[5] The top 4 scores have $R^2 \geq 0.9$.

To investigate the relationship between pre- and post-gen metrics more closely, we focus on the best pre-gen metric (that is, *mean_max_normcount_prefix0*) and consider its relationship to each post-gen metric individually. This is shown in Fig. 4. Irrespective of architecture and/or dataset, we observe a broadly linear relationship, despite some evidence of non-linearity at the lower ends of the scale, especially for CIDEr and WMD. This supports the hypothesis made at the outset, namely, that it is possible to predict the quality of captions, as measured by a standard metric, by considering the probability of the reference captions in the test set, without the need to generate the captions themselves.

[5] All correlations are significant at $p < 0.001$.



(a) CIDEr (b) SPICE

(c) METEOR (d) WMD

Fig. 4. Best pre-gen metric (*mean_max_normcount_prefix0*) vs post-gen metrics. The overall correlation has an R^2 of 0.94. (Best viewed in colour.) (Color figure online)

5 Discusison and Conclusion

We have introduced and defined the concept of pre-gen metrics and described a methodology to search for useful variants of these metrics. Our results show that pre-gen metrics closely approximate a variety of standard evaluation measures.

These results can be attributed to the fact that neural captioning models share core assumptions about the sampling mechanisms that underlie generation, and that standard evaluation metrics ultimately assess the output of this sampling process. Thus, it is possible to predict the quality of the output, as measured by a post-gen metric, using the probability distribution that a trained model predicts over prefixes of varying length in the reference captions. The practical implication is that pre-gen metrics can act as quick and efficient evaluation proxies during development. The theoretical implication is that the correlations among standard evaluation metrics reported in the literature are due, at least in part, to core sampling mechanisms shared by most neural generation architectures.

In future work, we plan to experiment with tuning captioning models using pre-gen metrics. We also wish to compare pre-gen metrics directly to human judgments.

Acknowledgments. The research in this paper is partially funded by the Endeavour Scholarship Scheme (Malta). Scholarships are part-financed by the European Union - European Social Fund (ESF) - Operational Programme II Cohesion Policy 2014–2020 Investing in human capital to create more opportunities and promote the well-being of society.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: semantic propositional image caption evaluation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part V. LNCS, vol. 9909, pp. 382–398. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_24
2. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings on the Workshop on Intrinsic and extrinsic evaluation measures for machine translation and/or summarization, vol. 29, pp. 65–72 (2005)
3. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B.: Automatic description generation from images: a survey of models, datasets, and evaluation measures. JAIR **55**, 409–442 (2016)
4. Cahill, A.: Correlating human and automatic evaluation of a German surface realiser. In: Proceedings of the ACL-IJCNLP 2009, pp. 97–100 (2009). https://doi.org/10.3115/1667583.1667615, http://dl.acm.org/citation.cfm?id=1667583.1667615, http://www.aclweb.org/anthology-new/P/P09/P09-2025.pdf
5. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the role of BLEU in machine translation research. In: Proceedings of the EACL 2006, pp. 249–256 (2006)
6. Caporaso, J.G., Deshpande, N., Fink, J.L., Bourne, P.E., Bretonnel Cohen, K., Hunter, L.: Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. Pac. Symp. Biocomput. **13**, 640–651 (2008). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2517250/
7. Dorr, B., Monz, C., Oard, D., President, S., Zajic, D., Schwartz, R.: Extrinsic evaluation of automatic metrics. Technical report, Institute for Advanced Computer Studies, University of Maryland, College Park, College Park, MD (2004)
8. Elliott, D., Keller, F.: Image description using visual dependency representations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1292–1302. Association for Computational Linguistics, Seattle, Washington, October 2013. http://www.aclweb.org/anthology/D13-1128
9. Elliott, D., Keller, F.: Comparing automatic evaluation measures for image description. In: Proceedings of the ACL 2014, pp. 452–457 (2014)
10. Espinosa, D., Rajkumar, R., White, M., Berleant, S.: Further Meta-evaluation of broad-coverage surface realization. In: Proceedings of the EMNLP 2010, pp. 564–574 (2010). http://www.aclweb.org/anthology/D10-1055
11. Gatt, A., Belz, A.: Introducing shared tasks to NLG: the TUNA shared task evaluation challenges. In: Krahmer, E., Theune, M. (eds.) EACL/ENLG -2009. LNCS (LNAI), vol. 5790, pp. 264–293. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15573-4_14
12. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. JAIR **47**(1), 853–899 (2013). https://doi.org/10.1109/cvprw.2013.51

13. Johnson, J., et al.: Image retrieval using scene graphs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2015. https://doi.org/10.1109/cvpr.2015.7298990
14. Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., Erdem, E.: Re-evaluating automatic metrics for image captioning. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/e17-1019
15. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. CoRR 1411.2539 (2014)
16. Kulkarni, G., et al.: Baby talk: understanding and generating simple image descriptions. In: CVPR 2011. IEEE, June 2011. https://doi.org/10.1109/cvpr.2011.5995466
17. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 957–966. PMLR, Lille (2015). http://proceedings.mlr.press/v37/kusnerb15.html
18. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the ACL 2004 (2004)
19. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: Proceedings of the ECCV 2014, pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. CoRR 1301.3781 (2013)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the ACL 2002, pp. 311–318 (2002)
22. Reiter, E., Belz, A.: An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Comput. Linguist. **35**(4), 529–558 (2009)
23. Tanti, M., Gatt, A., Camilleri, K.P.: Where to put the image in an image caption generator. Nat. Lang. Eng. **24**(3), 467–489 (2018). https://doi.org/10.1017/S1351324918000098. https://www.cambridge.org/core/journals/natural-language-engineering/article/where-to-put-the-image-in-an-image-caption-generator/A5B0ACFFE8E4AEAA5840DC61F93153F3#fndtn-information
24. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: consensus-based image description evaluation. In: Proceedings of the CVPR 2015 (2015)
25. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 652–663 (2017). https://doi.org/10.1109/tpami.2016.2587640
26. Wubben, S., van den Bosch, A., Krahmer, E.: Sentence simplification by monolingual machine translation. In: Proceedings of the ACL 2012, pp. 1015–1024 (2012). http://www.aclweb.org/anthology/P12-1107
27. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. TACL **2**, 67–78 (2014)