Learning-based Composite Metrics for Improved Caption Evaluation

Naeha Sharif, Lyndon White, Mohammed Bennamoun and Syed Afaq Ali Shah,

{naeha.sharif, lyndon.white}@research.uwa.edu.au,
and {mohammed.bennamoun, afaq.shah}@uwa.edu.au

The University of Western Australia.
35 Stirling Highway, Crawley, Western Australia

Abstract

The evaluation of image caption quality is a challenging task, which requires the assessment of two main aspects in a caption: adequacy and fluency. These quality aspects can be judged using a combination of several linguistic features. However, most of the current image captioning metrics focus only on specific linguistic facets, such as the lexical or semantic, and fail to meet a satisfactory level of correlation with human judgements at the sentence-level. We propose a learning-based framework to incorporate the scores of a set of lexical and semantic metrics as features, to capture the adequacy and fluency of captions at different linguistic levels. Our experimental results demonstrate that composite metrics draw upon the strengths of stand-alone measures to yield improved correlation and accuracy.

1 Introduction

Automatic image captioning requires the understanding of the visual aspects of images to generate human-like descriptions (Bernardi et al., 2016). The evaluation of the generated captions is crucial for the development and fine-grained analysis of image captioning systems (Vedantam et al., 2015). Automatic evaluation metrics aim at providing efficient, cost-effective and objective assessments of the caption quality. Since these automatic measures serve as an alternative to the manual evaluation, the major concern is that such measures should correlate well with human assessments. In other words, automatic metrics are expected to mimic the human judgement process by taking into account various aspects that humans consider when they assess the captions.

The evaluation of image captions can be characterized as having two major aspects: adequacy and fluency. Adequacy is how well the caption reflects the source image, and fluency is how well the caption conforms to the norms and conventions of human language (Toury, 2012). In the case of manual evaluation, both adequacy and fluency tend to shape the human perception of the overall caption quality. Most of the automatic evaluation metrics tend to capture these aspects of quality based on the idea that “the closer the candidate description to the professional human caption, the better it is in quality” (Papineni et al., 2002). The output in such case is a score (the higher the better) reflecting the similarity.

The majority of the commonly used metrics for image captioning such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) are based on the lexical similarity. Lexical measures (n-gram based) work by rewarding the n-gram overlaps between the candidate and the reference captions. Thus, measuring the adequacy by counting the n-gram matches and assessing the fluency by implicitly using the reference n-grams as a language model (Mutton et al., 2007). However, a high number of n-gram matches cannot always be indicative of a high caption quality, nor a low number of n-gram matches can always be reflective of a low caption quality (Giménez and Màrquez, 2010). A recently proposed semantic metric SPICE (Anderson et al., 2016), overcomes this deficiency of lexical measures by measuring the semantic similarity of candidate and reference captions using *Scene Graphs*. However, the major drawback of SPICE is that it ignores the fluency of the output caption.

Integrating assessment scores of different measures is an intuitive and reasonable way to improve the current image captioning evaluation methods. Through this methodology, each metric plays the role of a judge, assessing the quality of captions in terms of lexical, grammatical or semantic accuracy. For this research, we use the scores conferred

Proceedings of ACL 2018, Student Research Workshop, pages 14–20
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

by a set of measures that are commonly used for captioning and combine them through a learning-based framework. In this work:

1. We evaluate various combinations of a chosen set of metrics and show that the proposed composite metrics correlate better with human judgements.

2. We analyse the accuracy of composite metrics in terms of differentiating between pairs of captions in reference to the ground truth captions.

2 Literature Review

The success of any captioning system depends on how well it transforms the visual information to natural language. Therefore, the significance of reliable automatic evaluation metrics is undeniable for the fine-grained analysis and advancement of image captioning systems. While image captioning has drawn inspiration from the Machine Translation (MT) domain for encoder-decoder based captioning networks (Vinyals et al., 2015), (Xu et al., 2015), (Yao et al., 2016), (You et al., 2016), (Lu et al., 2017), it has also benefited from automatic metrics which were initially proposed to evaluate machine translations/text summaries, such as BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and ROUGE (Lin, 2004).

In the past few years, two metrics CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016) were developed specifically for image captioning. Compared to the previously used metrics, these two show a better correlation with human judgements. The authors in (Liu et al., 2016) proposed a linear combination of SPICE and CIDEr called *SPIDEr* and showed that optimizing image captioning models for SPIDEr's score can lead to better quality captions. However, SPIDEr was not evaluated for its correlation with human judgements. Recently, (Kusner et al., 2015) proposed the use of a document similarity metric *Word Mover's Distance* (WMD), which uses the word2vec (Mikolov et al., 2013) embedding space to determine the distance between two texts.

The metrics used for caption evaluation can be broadly categorized as lexical and semantic measures. Lexical metrics reward the n-gram matches between candidate captions and human generated reference texts (Giménez and Màrquez, 2010), and can be further categorized as unigram and n-gram based measures. Unigram based methods such as BLEU-1 (Papineni et al., 2002), assess only the lexical correctness of the candidate. However, in the case of METEOR or WMD, where some sort of synonym-matching/stemming is also involved, unigram-overlaps help to evaluate both the lexical and to some degree the semantic aptness of the output caption. N-gram based metrics such as ROUGE and CIDEr primarily assess the lexical correctness of the caption, but also measure some amount of syntactic accuracy by capturing the word order.

The lexical measures have received criticism based on the argument that the n-gram overlap is neither an adequate nor a necessary indicative measure of the caption quality (Anderson et al., 2016). To overcome this limitation, semantic metrics such as SPICE, capture the sentence meaning to evaluate the candidate captions. Their performance however is highly dependent on a successful semantic parsing. Purely syntactic measures, which capture the grammatical correctness, exist, and have been used in MT (Mutton et al., 2007), but not in the captioning domain.

While fluency (well-formedness) of a candidate caption can be attributed to the syntactic and lexical correctness (Fomicheva et al., 2016), adequacy (informativeness) depends on the lexical and semantic correctness (Rios et al., 2011). We hypothesize that by combining scores from different metrics, which have different strengths in measuring adequacy and fluency, a composite metric that is of overall higher quality is created (Sec. 5).

Machine learning offers a systematic approach to integrate the scores of stand-alone metrics. In the MT evaluation, various successful learning paradigms have been proposed (Bojar et al., 2016), (Bojar et al., 2017) and the existing learning-based metrics can be categorized as *binary functions*–"which classify the candidate translation as good or bad" (Kulesza and Shieber, 2004), (Guzmán et al., 2015) or *continuous functions*–"which score the quality of translation on an absolute scale" (Song and Cohn, 2011), (Albrecht and Hwa, 2008). Our research is conceptually similar to the work in (Kulesza and Shieber, 2004), which induces a *"human-likeness"* criteria. However, our approach differs in terms of the learning algorithm as well as the features used. Moreover, the focus of this work is to assess various combinations of metrics (that capture the caption quality at

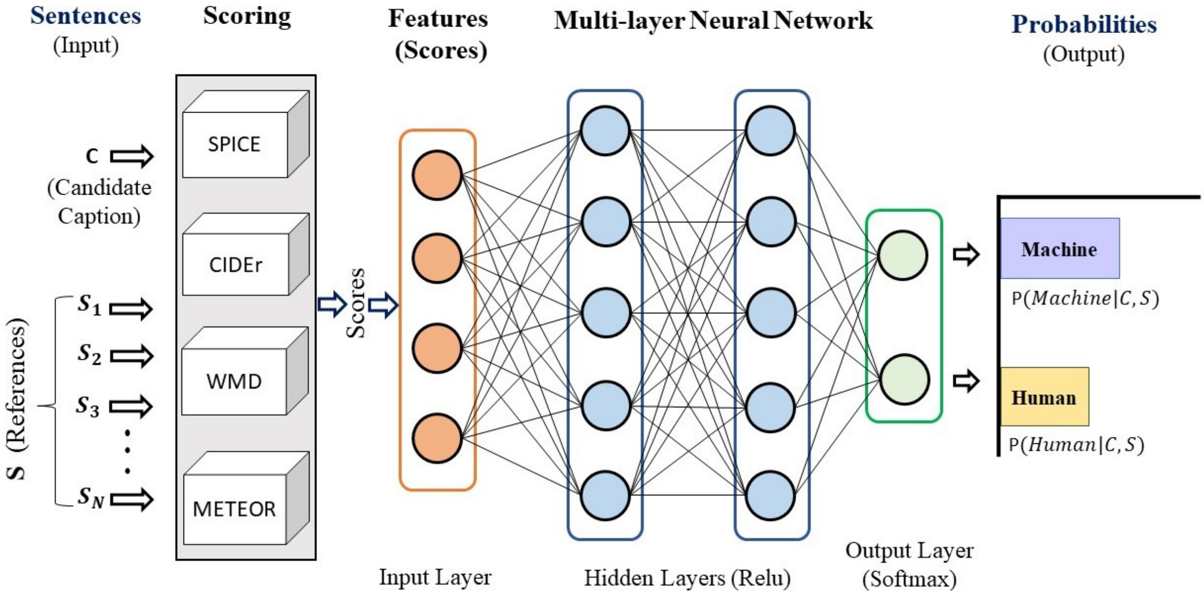

Figure 1: Overall framework of the proposed Composite Metrics

different linguistic levels) in terms of their correlation with human judgements at the sentence level.

3 Proposed Approach

In our approach, we use scores conferred by a set of existing metrics as an input to a multi-layer feed-forward neural network. We adopt a training criteria based on a simple question: *is the caption machine or human generated?* Our trained classifier sets a boundary between good and bad quality captions, thus classifying them as human or machine produced. Furthermore, we obtain a continuous output score by using the class-probability, which can be considered as some "measure of believability" that the candidate caption is human generated. Framing our learning problem as a classification task allows us to create binary training data using the human generated captions and machine generated captions as positive and negative training examples respectively.

Our proposed framework shown in Figure 1 first extracts a set of numeric features using the candidate "C" and the reference sentences "S". The extracted feature vector is then fed as an input to our multi-layer neural network. Each entity of the feature vector corresponds to the score generated by one of the four measures: METEOR, CIDEr, WMD[1] and SPICE respectively. We chose these measures because they show a relatively better correlation with human judgements compared to the other commonly used ones for captioning (Kilickaya et al., 2016). Our composite metrics are named $Eval_{MS}$, $Eval_{CS}$, $Eval_{MCS}$, $Eval_{WCS}$, $Eval_{MWS}$ and $Eval_{MWCS}$. The subscript letters in each name corresponds to the first letter of each individual metric. For example, $Eval_{MS}$ corresponds to the combination of METEOR and SPICE. Figure 2 shows the linguistic aspects captured by the stand-alone[2] and the composite metrics. SPICE is based on sentence meanings, thus it evaluates the semantics. CIDEr covers the syntactic and lexical aspects, whereas Meteor and WMD assess the lexical and semantic components. The learning-based metrics mostly fall in the region formed by the overlap of all three major linguistics facets, leading to better a evaluation.

We train our metrics to maximise the classification accuracy on the training dataset. Since we are primarily interested in maximizing the correlation with human judgements, we perform early stopping based on Kendalls τ (rank correlation) with the validation set.

4 Experimental Setup

To train our composite metrics, we source data from Flicker30k dataset (Plummer et al., 2015) and three image captioning models namely: (1) show and tell (Vinyals et al., 2015), (2) show, attend and tell (soft-attention) (Xu et al., 2015), and (3) adaptive attention (Lu et al., 2017). Flicker30k

[1] We convert the WMD distance score to similarity by using a negative exponential, to use it as a feature.

[2] The stand-alone metrics marked with an * in the Figure 2 are used as features for this work.

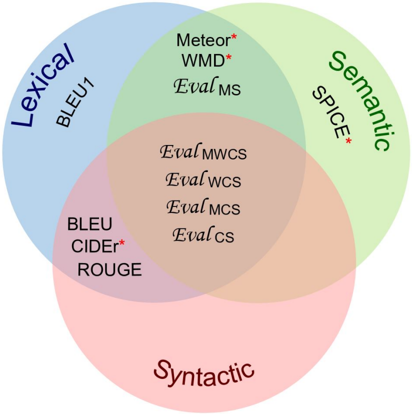

Figure 2: Various automatic measures (standalone and combined) and their respective linguistic levels. See Sec. 3 for more details.

Table 1: Kendall's correlation co-efficient of automatic evaluation metrics and proposed composite metrics against human quality judgements. All correlations are significant at p<0.001

Individual Metrics	**Kendall** τ	**Composite Metrics**	**Kendall** τ
BLEU	0.202	$Eval_{MS}$	**0.386**
ROUGE-L	0.216	$Eval_{CS}$	0.384
METEOR	0.352	$Eval_{MCS}$	**0.386**
CIDEr	0.356	$Eval_{WCS}$	0.379
SPICE	0.366	$Eval_{MWS}$	0.367
WMD	0.336	$Eval_{MWCS}$	0.378

dataset consists of 31,783 photos acquired from Flicker[3], each paired with 5 captions obtained through the Amazon Mechanical Turk (AMT) (Turk, 2012). For each image in Flicker30k, we randomly select three of the human generated captions as positive training examples, and three machine generated (one from each image captioning model) captions as negative training examples. We combined the Microsoft COCO (Chen et al., 2015) training and validation set (containing 123,287 images in total, each paired with 5 or more captions), to train the image captioning models using their official codes. These image captioning models achieved state-of-the-art performance when they were published.

In order to obtain reference captions for each training example, we again use the human written descriptions of Flicker30k. For each negative training example (machine-generated caption), we randomly choose 4 out of 5 human written captions originally associated with each image. Whereas, for each positive training example (human-generated caption), we use the 5 human written captions associated with each image, selecting one of these as a human candidate caption (positive example) and the remaining 4 as references. In Figure 3, a possible pairing scenario is shown for further clarification.

For our validation set, we source data from Flicker8k (Young et al., 2014). This dataset contains 5,822 captions assessed by three expert judges on a scale of 1 (the caption is unrelated to the image) to 4 (the caption describes the image without any errors). From our training set, we remove the captions of images which overlap with the captions in the validation and test sets (discussed in Sec. 5), leaving us with a total of 132,984 non-overlapping captions for the training of the composite metrics.

5 Results and Discussion

5.1 Correlation

The most desirable characteristic of an automatic evaluation metric is its strong correlation with human scores (Zhang and Vogel, 2010). A stronger correlation with human judgements indicates that a metric captures the information that humans use to assess a candidate caption. To evaluate the sentence-level correlation of our composite metrics with human judgements, we source data from a dataset collected by the authors in (Aditya et al., 2017). We use 6993 manually evaluated human and machine generated captions from this set, which were scored by AMT workers for correctness on the scale of 1 (low relevance to image) to 5 (high relevance to image). Each caption in the dataset is accompanied by a single judgement. In Table 1, we report the Kendalls τ correlation coefficient for the proposed composite metrics and other commonly used caption evaluation metrics.

It can be observed from Table 1 that composite metrics outperform stand-alone metrics in terms of sentence-level correlation. The combination of Meteor and SPICE ($Eval_{MS}$) and METEOR, CIDEr and SPICE ($Eval_{MCS}$) showed the most promising results. The success of these composite metrics can be attributed to the individual strengths of Meteor, CIDEr and SPICE. METEOR is a strong lexical measure based on unigram matching, which uses additional linguistic knowledge for word matching, such as the morphological variation in words via stemming and

[3] https://www.flickr.com/

(a)

Human Generated Captions

S_1: "young toddler raking leaves in the outdoors "
S_2: "a little boy raking the leaves with a rake "
S_3: "a young boy in a red t-shirt rakes leaves "
S_4: "a little boy is raking leaves "
S_5: "a young boy raking leaves "

Machine Generated Captions

M_1: "a little girl holding a baseball bat on a field"
M_2: "a little boy playing with a ball in a field"
M_3: "a little boy holding a bat in his hands"

(b)

	Candidate Captions	References
1	M_1 (machine)	$\{S_1, S_2, S_3, S_4\}$
2	M_2 (machine)	$\{S_2, S_1, S_3, S_5\}$
3	M_3 (machine)	$\{S_1, S_4, S_5, S_2\}$
4	S_1 (human)	$\{S_2, S_3, S_4, S_5\}$
5	S_2 (human)	$\{S_1, S_3, S_5, S_4\}$
6	S_3 (human)	$\{S_1, S_2, S_4, S_5\}$

(c)

Figure 3: Shows an example of a candidate and reference pairing that is used in the training set. (a) Image, (b) human and machine generated captions for the image, and (c) candidate and reference pairings for the image.

dictionary based look-up for synonyms and paraphrases (Banerjee and Lavie, 2005). CIDEr uses higher order n-grams to account for fluency and down-weighs the commonly occurring (less informative) n-grams by performing Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram in the dataset (Vedantam et al., 2015). SPICE on the other hand is a strong indicator of the semantic correctness of a caption. Together these metrics assess the lexical, semantic and syntactic information. The composite metrics which included WMD in the combination achieved a lower performance, compared to the ones in which WMD was not included. One possible reason is that WMD heavily penalizes shorter candidate captions when the number of words between the output and the reference captions are not equal (Kusner et al., 2015). This penalty might not be consistently useful as it is possible for a shorter candidate caption to be both fluent and adequate. Therefore, WMD is a better suited metric for measuring document distance.

5.2 Accuracy

We follow the framework introduced in (Vedantam et al., 2015) to analyse the ability of a metric to discriminate between pairs of captions with reference to the ground truth caption. A metric is considered accurate if it assigns a higher score to the caption preferred by humans. For this experiment, we use PASCAL-50s (Vedantam et al., 2015), which contains human judgements for 4000 triplets of descriptions (one reference caption with two candidate captions). Based on the pairing, the triplets are grouped into four categories (comprising of 1000 triplets each) i.e., Human-Human Correct (HC), Human-Human Incorrect (HI), Human-Machine (HM), Machine-Machine (MM). We follow the original approach of (Vedantam et al., 2015) and use 5 reference captions per candidate to assess the accuracy of the metrics and report them in Table 2. Table 2 shows that on average composite measures produce better accuracy compared to the individual metrics. Amongst the four categories, HC is the hardest, in which all metrics show the worst performance. Differentiating between two good quality (human generated) correct captions is challenging as it involves a fine-grained analysis of the two candidates. $Eval_{MS}$ achieves the highest accuracy in HC category which shows that as captioning systems continue to improve, this combination of lexical and semantic metrics will continue to perform well. Moreover, human generated captions are usually fluent. Therefore, a combination of strong indicators of adequacy such as SPICE and METEOR is the most suitable for this task. $Eval_{MCS}$ shows the highest accuracy in differ-

Table 2: Comparative accuracy results (in percentage) on four kinds of pairs tested on PASCAL-50s

Metrics	HC	HI	HM	MM	AVG
BLEU	53.7	93.2	85.6	61.0	73.4
ROUGE-L	56.5	95.3	93.4	58.5	75.9
METEOR	61.1	97.6	**94.6**	62.0	78.8
CIDEr	57.8	98.0	88.8	68.2	78.2
SPICE	58.0	96.7	88.4	71.6	78.7
WMD	56.2	98.4	91.7	71.5	79.5
$Eval_{MS}$	**62.8**	97.9	93.5	69.6	**80.9**
$Eval_{CS}$	59.5	98.3	90.7	71.3	79.9
$Eval_{MCS}$	60.2	98.3	91.8	**71.8**	80.5
$Eval_{WCS}$	58.2	**98.7**	91.7	70.6	79.8
$Eval_{MWS}$	56.9	98.4	91.3	71.2	79.4
$Eval_{MWCS}$	59.0	98.5	90.7	70.2	79.6

entiating between machine captions, which is another important category as one of the main goals of automatic evaluation is to distinguish between two machine algorithms. Amongst the composite metrics, $Eval_{MS}$ is again the best in distinguishing human captions (good quality) from machine captions (bad quality) which was our basic training criteria.

6 Conclusion and Future Works

In this paper we propose a learning-based approach to combine various metrics to improve caption evaluation. Our experimental results show that metrics operating along different linguistic dimensions can be successfully combined through a learning-based framework, and they outperform the existing metrics for caption evaluation in term of correlation and accuracy, with $Eval_{MS}$ and $Eval_{MCS}$ giving the best overall performance.

Our study reveals that the proposed approach is promising and has a lot of potential to be used for evaluation in the captioning domain. In the future, we plan to integrate features (components) of metrics instead of their scores for a better performance. We also intend to use syntactic measures, which to the best of our knowledge have not yet been used for caption evaluation (except in an indirect way by the n-gram measures which capture the word order) and study how they can improve the correlation at the sentence level. Majority of the metrics for captioning focus more on adequacy as compared to fluency. This aspect also needs further attention and a combination of metrics/features that can specifically assess the fluency of captions needs to be devised.

Acknowledgements

The authors are grateful to Nvidia for providing Titan-Xp GPU, which was used for the experiments. This research is supported by Australian Research Council, ARC DP150100294.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermüller. 2017. Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding*.

Joshua S Albrecht and Rebecca Hwa. 2008. Regression for machine translation evaluation at the sentence level. *Machine Translation*, 22(1-2):1.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank, et al. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.(JAIR)*, 55:409–442.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the wmt16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 199–231.

Ondřej Bojar, Jindřich Helcl, Tom Kocmi, Jindřich Libovickỳ, and Tomáš Musil. 2017. Results of the wmt17 neural mt training task. In *Proceedings of the Second Conference on Machine Translation*, pages 525–533.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Marina Fomicheva, Núria Bel, Lucia Specia, Iria da Cunha, and Anton Malinovskiy. 2016. Cobaltf: a fluent metric for mt evaluation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 483–490.

Jesús Giménez and Lluís Màrquez. 2010. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24(3-4):209–240.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 805–814.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2016. Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600*.

Alex Kulesza and Stuart M Shieber. 2004. A learning approach to improving sentence-level mt evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 75–84.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2016. Improved image captioning via policy gradient optimization of spider. *arXiv preprint arXiv:1612.00370*.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. Gleu: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2641–2649. IEEE.

Miguel Rios, Wilker Aziz, and Lucia Specia. 2011. Tine: A metric to assess mt adequacy. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 116–122. Association for Computational Linguistics.

Xingyi Song and Trevor Cohn. 2011. Regression and ranking based optimisation for sentence level machine translation evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 123–129. Association for Computational Linguistics.

Gideon Toury. 2012. *Descriptive Translation Studies and beyond: revised edition*, volume 100. John Benjamins Publishing.

Amazon Mechanical Turk. 2012. Amazon mechanical turk. *Retrieved August*, 17:2012.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2016. Boosting image captioning with attributes. *OpenReview*, 2(5):8.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Ying Zhang and Stephan Vogel. 2010. Significance tests of automatic machine translation evaluation metrics. *Machine Translation*, 24(1):51–65.