Comparing Automatic Evaluation Measures for Image Description

Desmond Elliott and **Frank Keller**
Institute for Language, Cognition, and Computation
School of Informatics, University of Edinburgh
d.elliott@ed.ac.uk, keller@inf.ed.ac.uk

Abstract

Image description is a new natural language generation task, where the aim is to generate a human-like description of an image. The evaluation of computer-generated text is a notoriously difficult problem, however, the quality of image descriptions has typically been measured using unigram BLEU and human judgements. The focus of this paper is to determine the correlation of automatic measures with human judgements for this task. We estimate the correlation of unigram and Smoothed BLEU, TER, ROUGE-SU4, and Meteor against human judgements on two data sets. The main finding is that unigram BLEU has a weak correlation, and Meteor has the strongest correlation with human judgements.

1 Introduction

Recent advances in computer vision and natural language processing have led to an upsurge of research on tasks involving both vision and language. State of the art visual detectors have made it possible to hypothesise *what* is in an image (Guillaumin et al., 2009; Felzenszwalb et al., 2010), paving the way for automatic image description systems. The aim of such systems is to extract and reason about visual aspects of images to generate a human-like description. An example of the type of image and gold-standard descriptions available can be seen in Figure 1. Recent approaches to this task have been based on slot-filling (Yang et al., 2011; Elliott and Keller, 2013), combining web-scale n-grams (Li et al., 2011), syntactic tree substitution (Mitchell et al., 2012), and description-by-retrieval (Farhadi et al., 2010; Ordonez et al., 2011; Hodosh et al., 2013). Image description has been compared to translating an image into text (Li et al., 2011; Kulkarni et al., 2011) or summarising an image

1. An older woman with a small dog in the snow.
2. A woman and a cat are outside in the snow.
3. A woman in a brown vest is walking on the snow with an animal.
4. A woman with a red scarf covering her head walks with her cat on snow-covered ground.
5. Heavy set woman in snow with a cat.

Figure 1: An image from the Flickr8K data set and five human-written descriptions. These descriptions vary in the adjectives or prepositional phrases that describe the woman (1, 3, 4, 5), incorrect or uncertain identification of the cat (1, 3), and include a sentence without a verb (5).

(Yang et al., 2011), resulting in the adoption of the evaluation measures from those communities.

In this paper we estimate the correlation of human judgements with five automatic evaluation measures on two image description data sets. Our work extends previous studies of evaluation measures for image description (Hodosh et al., 2013), which focused on unigram-based measures and reported agreement scores such as Cohen's κ rather than correlations. The main finding of our analysis is that TER and unigram BLEU are weakly corre-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 452–457,
Baltimore, Maryland, USA, June 23-25 2014.

lated against human judgements, ROUGE-SU4 and Smoothed BLEU are moderately correlated, and the strongest correlation is found with Meteor.

2 Methodology

We estimate Spearman's ρ for five different automatic evaluation measures against human judgements for the automatic image description task. Spearman's ρ is a non-parametric correlation coefficient that restricts the ability of outlier data points to skew the co-efficient value. The automatic measures are calculated on the sentence level and correlated against human judgements of semantic correctness.

2.1 Data

We perform the correlation analysis on the Flickr8K data set of Hodosh et al. (2013), and the data set of Elliott and Keller (2013).

The test data of the Flickr8K data set contains 1,000 images paired with five reference descriptions. The images were retrieved from Flickr, the reference descriptions were collected from Mechanical Turk, and the human judgements were collected from expert annotators as follows: each image in the test data was paired with the highest scoring sentence(s) retrieved from all possible test sentences by the TRI5SEM model in Hodosh et al. (2013). Each image–description pairing in the test data was judged for semantic correctness by three expert human judges on a scale of 1–4. We calculate automatic measures for each image–retrieved sentence pair against the five reference descriptions for the original image.

The test data of Elliott and Keller (2013) contains 101 images paired with three reference descriptions. The images were taken from the PASCAL VOC Action Recognition Task, the reference descriptions were collected from Mechanical Turk, and the judgements were also collected from Mechanical Turk. Elliott and Keller (2013) generated two-sentence descriptions for each of the test images using four variants of a slot-filling model, and collected five human judgements of the semantic correctness and grammatical correctness of the description on a scale of 1–5 for each image–description pair, resulting in a total of 2,042 human judgement–description pairings. In this analysis, we use only the first sentence of the description, which describes the event depicted in the image.

2.2 Automatic Evaluation Measures

BLEU measures the effective overlap between a reference sentence X and a candidate sentence Y. It is defined as the geometric mean of the effective n-gram precision scores, multiplied by the brevity penalty factor BP to penalise short translations. p_n measures the effective overlap by calculating the proportion of the maximum number of n-grams co-occurring between a candidate and a reference and the total number of n-grams in the candidate text. More formally,

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$p_n = \frac{\sum_{c \in cand} \sum_{ngram \in c} count_{clip}(ngram)}{\sum_{c \in cand} \sum_{ngram \in c} count(ngram)}$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

Unigram BLEU without a brevity penalty has been reported by Kulkarni et al. (2011), Li et al. (2011), Ordonez et al. (2011), and Kuznetsova et al. (2012); to the best of our knowledge, the only image description work to use higher-order n-grams with BLEU is Elliott and Keller (2013). In this paper we use the smoothed BLEU implementation of Clark et al. (2011) to perform a sentence-level analysis, setting $n = 1$ and no brevity penalty to get the unigram BLEU measure, or $n = 4$ with the brevity penalty to get the Smoothed BLEU measure. We note that a higher BLEU score is better.

ROUGE measures the longest common subsequence of tokens between a candidate Y and reference X. There is also a variant that measures the co-occurrence of pairs of tokens in both the candidate and reference (a skip-bigram): ROUGE-SU*. The skip-bigram calculation is parameterised with d_{skip}, the maximum number of tokens between the words in the skip-bigram. Setting d_{skip} to 0 is equivalent to bigram overlap and setting d_{skip} to ∞ means tokens can be any distance apart. If $\alpha = |SKIP2(X,Y)|$ is the number of matching skip-bigrams between the reference and the candidate, then skip-bigram ROUGE is formally defined as:

$$R_{SKIP2} = \alpha \,/ \binom{\alpha}{2}$$

ROUGE has been used by only Yang et al. (2011) to measure the quality of generated descriptions, using a variant they describe as ROUGE-1. We set $d_{skip} = 4$ and award partial credit for unigram only matches, otherwise known as ROUGE-SU4. We use ROUGE v.1.5.5 for the analysis, and configure the evaluation script to return the result for the average score for matching between the candidate and the references. A higher ROUGE score is better.

TER measures the number of modifications a human would need to make to transform a candidate Y into a reference X. The modifications available are insertion, deletion, substitute a single word, and shift a word an arbitrary distance. TER is expressed as the percentage of the sentence that needs to be changed, and can be greater than 100 if the candidate is longer than the reference. More formally,

$$TER = \frac{|\text{edits}|}{|\text{reference tokens}|}$$

TER has not yet been used to evaluate image description models. We use v.0.8.0 of the TER evaluation tool, and a lower TER is better.

Meteor is the harmonic mean of unigram precision and recall that allows for exact, synonym, and paraphrase matchings between candidates and references. It is calculated by generating an alignment between the tokens in the candidate and reference sentences, with the aim of a 1:1 alignment between tokens and minimising the number of chunks *ch* of contiguous and identically ordered tokens in the sentence pair. The alignment is based on exact token matching, followed by Wordnet synonyms, and then stemmed tokens. We can calculate precision, recall, and F-measure, where m is the number of aligned unigrams between candidate and reference. Meteor is defined as:

$$M = (1 - Pen) \cdot F_{mean}$$

$$Pen = \gamma \left(\frac{ch}{m} \right)^{\theta}$$

$$F_{mean} = \frac{PR}{\alpha P + (1-\alpha) R}$$

$$P = \frac{|m|}{|\text{unigrams in candidate}|}$$

$$R = \frac{|m|}{|\text{unigrams in reference}|}$$

We calculated the Meteor scores using release 1.4.0 with the package-provided free parameter settings of 0.85, 0.2, 0.6, and 0.75 for the matching components. Meteor has not yet been reported to evaluate the performance of different models on the image description task; a higher Meteor score is better.

	Flickr 8K co-efficient $n = 17,466$	E&K (2013) co-efficient $n = 2,040$
METEOR	0.524	0.233
ROUGE SU-4	0.435	0.188
Smoothed BLEU	0.429	0.177
Unigram BLEU	0.345	0.097
TER	-0.279	-0.044

Table 1: Spearman's correlation co-efficient of automatic evaluation measures against human judgements. All correlations are significant at $p < 0.001$.

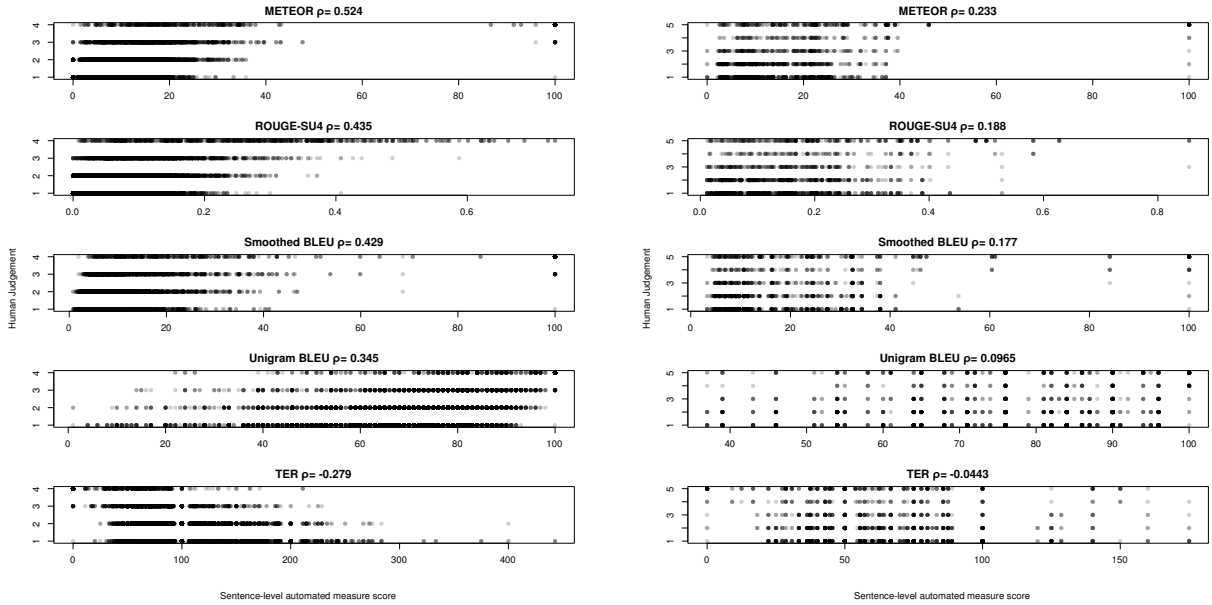
2.3 Protocol

We performed the correlation analysis as follows. The sentence-level evaluation measures were calculated for each image–description–reference tuple. We collected the BLEU, TER, and Meteor scores using MultEval (Clark et al., 2011), and the ROUGE-SU4 scores using the RELEASE-1.5.5.pl script. The evaluation measure scores were then compared with the human judgements using Spearman's correlation estimated at the sentence-level.

3 Results

Table 1 shows the correlation co-efficients between automatic measures and human judgements and Figures 2(a) and (b) show the distribution of scores for each measure against human judgements. To classify the strength of the correlations, we followed the guidance of Dancey and Reidy (2011), who posit that a co-efficient of 0.0–0.1 is uncorrelated, 0.11–0.4 is *weak*, 0.41–0.7 is *moderate*, 0.71–0.90 is *strong*, and 0.91–1.0 is *perfect*.

On the Flickr8k data set, all evaluation measures can be classified as either *weakly* correlated or *moderately* correlated with human judgements and all results are significant. TER is only weakly correlated with human judgements but could prove useful in comparing the types of differences between models. An analysis of the distribution of TER scores in Figure 2(a) shows that differences in candidate and reference length are prevalent in the image description task. Unigram BLEU is also only weakly correlated against human judgements, even though it has been reported extensively for this task.



(a) Flick8K data set, n=17,466.

(b) E&K (2013) data set, n=2,042.

Figure 2: Distribution of automatic evaluation measures against human judgements. ρ is the correlation between human judgements and the automatic measure. The intensity of each point indicates the number of occurrences that fall into that range.

Figure 2(a) shows an almost uniform distribution of unigram BLEU scores, regardless of the human judgement. Smoothed BLEU and ROUGE-SU4 are moderately correlated with human judgements, and the correlation is stronger than with unigram BLEU. Finally, Meteor is most strongly correlated measure against human judgements. A similar pattern is observed in the Elliott and Keller (2013) data set, though the correlations are lower across all measures. This could be caused by the smaller sample size or because the descriptions were generated by a computer, and not retrieved from a collection of human-written descriptions containing the gold-standard text, as in the Flickr8K data set.

Qualitative Analysis

Figure 3 shows two images from the test collection of the Flickr8K data set with a low Meteor score and a maximum human judgement of semantic correctness. The main difference between the candidates and references are in deciding *what* to describe (content selection), and *how* to describe it (realisation). We can hypothesise that in both translation and summarisation, the source text acts as a lexical and semantic framework within which the translation or summarisation process takes place. In Figure 3(a), the authors of the descriptions made different decisions on *what* to describe. A decision has been made to describe the role of the officials in the candidate text, and not in the reference text. The underlying cause of this is an active area of research in the human vision literature and can be attributed to bottom-up effects, such as saliency (Itti et al., 1998), top-down contextual effects (Torralba et al., 2006), or rapidly-obtained scene properties (Oliva and Torralba, 2001). In (b), we can see the problem of deciding how to describe the selected content. The reference uses a more specific noun to describe the person on the bicycle than the candidate.

4 Discussion

There are several differences between our analysis and that of Hodosh et al. (2013). First, we report Spearman's ρ correlation coefficient of automatic measures against human judgements, whereas they report agreement between judgements and automatic measures in terms of Cohen's κ. The use of κ requires the transformation of real-valued scores into categorical values, and thus loses information; we use the judgement and evaluation measure scores in their original forms. Second, our use of Spearman's ρ means we can readily use all of the available data for the correlation analysis, whereas Hodosh et al. (2013) report agreement on thresholded subsets of the data. Third, we report the correlation coefficients against five evaluation measures,

Candidate: Football players gathering to contest something to collaborating officials.
Reference: A football player in red and white is holding both hands up.

(a)

Candidate: A man is attempting a stunt with a bicycle.
Reference: Bmx biker Jumps off of ramp.

(b)

Figure 3: Examples in the test data with low Meteor scores and the maximum expert human judgement. (a) the candidate and reference are from the same image, and show differences in *what* to describe, in (b) the descriptions are retrieved from different images and show differences in *how* to describe an image.

some of which go beyond unigram matchings between references and candidates, whereas they only report unigram BLEU and unigram ROUGE. It is therefore difficult to directly compare the results of our correlation analysis against Hodosh et al.'s agreement analysis, but they also reach the conclusion that unigram BLEU is not an appropriate measure of image description performance. However, we do find stronger correlations with Smoothed BLEU, skip-bigram ROUGE, and Meteor.

In contrast to the results presented here, Reiter and Belz (2009) found no significant correlations of automatic evaluation measures against human judgements of the *accuracy* of machine-generated weather forecasts. They did, however, find significant correlations of automatic measures against *fluency* judgements. There are no fluency judgements available for Flickr8K, but Elliott and Keller (2013) report grammaticality judgements for their data, which are comparable to fluency ratings. We failed to find significant correlations between grammatlicality judgements and any of the automatic measures on the Elliott and Keller (2013) data. This discrepancy could be explained in terms of the differences between the weather forecast generation and image description tasks, or because the image description data sets contain thousands of texts and a few human judgements per text, whereas the data sets of Reiter and Belz (2009) included hundreds of texts with 30 human judges.

5 Conclusions

In this paper we performed a sentence-level correlation analysis of automatic evaluation measures against expert human judgements for the automatic image description task. We found that sentence-level unigram BLEU is only weakly correlated with human judgements, even though it has extensively reported in the literature for this task. Meteor was found to have the highest correlation with human judgements, but it requires Wordnet and paraphrase resources that are not available for all languages. Our findings held when judgements were made on human-written or computer-generated descriptions.

The variability in what and how people describe images will cause problems for all of the measures compared in this paper. Nevertheless, we propose that unigram BLEU should no longer be used as an objective function for automatic image description because it has a weak correlation with human accuracy judgements. We recommend adopting either Meteor, Smoothed BLEU, or ROUGE-SU4 because they show stronger correlations with human judgements. We believe these suggestions are also applicable to the ranking tasks proposed in Hodosh et al. (2013), where automatic evaluation scores could act as features to a ranking function.

Acknowledgments

Alexandra Birch and R. Calen Walshe, and the anonymous reviewers provided valuable feedback on this paper. The research is funded by ERC Starting Grant SYNPROC No. 203427.

References

Jonathon H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA.

Christine Dancey and John Reidy, 2011. *Statistics Without Maths for Psychology*, page 175. Prentice Hall, 5th edition.

Desmond Elliott and Frank Keller. 2013. Image Description using Visual Dependency Representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, U.S.A.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision*, pages 15–29, Heraklion, Crete, Greece.

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. 2010. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Matthieu Guillaumin, Thomas Mensink, Jakob J. Verbeek, and Cornelia Schmid. 2009. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *IEEE 12th International Conference on Computer Vision*, pages 309–316, Kyoto, Japan.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task : Data , Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1601–1608, Colorado Springs, Colorado, U.S.A.

Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. 2012. Collective Generation of Natural Image Descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 359–368, Jeju Island, South Korea.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, Portland, Oregon, U.S.A.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daumé III. 2012. Midge : Generating Image Descriptions From Computer Vision Detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France.

Aude Oliva and Antonio Torralba. 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems 24*, Granada, Spain.

Ehud Reiter and A Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychologial Review*, 113(4):766–786.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-Guided Sentence Generation of Natural Images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 444–454, Edinburgh, Scotland, UK.