Investigating Text-Based Similarity Measures for Automatic Image Description Evaluation

Jeroen H.M. Rooijmans
5887410

Bachelor thesis
Credits: 18 EC

Bachelor Opleiding Kunstmatige Intelligente
University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisors
Dr. D. Elliott
Centrum Wiskunde & Informatica (CWI)
Science Park 123
1098 XG Amsterdam

Dr. I. Titov
Institute for Logic, Language and Computation (ILLC)
Faculty of Science
University of Amsterdam
Science Park 107
1099 XG Amsterdam

June 26th, 2015
Acknowledgement

I would like to thank Desmond Elliott and Ivan Titov for supervising my research. Their guidance, support and motivational approach were crucial factors in developing this work. Thanks to their expertise I have had the chance to develop as an A.I. researcher.
I would also like to thank my girlfriend, family and friends for their support.
Abstract

Automatic image description systems aim to generate accurate descriptions of an image, capturing appropriate meaning of “what” is in the image and expressing that information in a human-like way. Despite increasing perceptive and expressive capabilities of these systems due to ingenious state-of-the-art approaches that are adopted from preceding fields of study such as computer vision and natural language processing, evaluation of computer-generated text has proven to be a persistent challenge. This re-evaluation of previous efforts explores what lexical and semantic properties are currently used in evaluation of machine generated text, and how performance of automatic image description systems is measured. I estimate the correlation of automatic evaluation measures: Bleu, TER, Rouge-SU4, Meteor and CIDEr with expert and crowdsourced human judgements. The main finding is that CIDEr shows the strongest correlation with human judgements, while TER shows the weakest correlation. I also explore automatic evaluation measure performance on crowdsourced human judgements and estimate inter-annotation agreement between expert and crowdsourced raters, results are difficult to interpreted hence future work is desirable.
Contents

1 Introduction 4

2 Methodology 6
  2.1 Data-sets ........................................ 7
  2.2 Automatic Evaluation Measures .................... 9
  2.3 Statistical Measures .............................. 13
    2.3.1 Correlation measures .......................... 14
    2.3.2 Annotation agreement measures ................ 15
  2.4 Protocol ........................................... 16

3 Results 18

4 Evaluation 19

5 Discussion 21

6 Conclusions 22
1 Introduction

Through advances in computer vision and natural language processing, research on automatic image description systems has seen a recent upsurge. The aim of such a system is to construct human-like descriptions that are based on the visual properties of an image (i.e. a sentences that accurately describes an image), and can be compared to task translating an image into text or summarizing an image. Human evaluation of these systems is extensive, but expensive and rarely reusable. Alternative evaluation measures are adopted from similar, preceding fields of study (e.g. machine translation (MT) and automatic summarisation).

The goal of automatic image description systems is to extract and reason about visual aspects of images and ultimately expressing this knowledge in a human-like description, advancing state-of-the-art in object recognition and natural language generation by placing it in the broader context of scene understanding. Such systems commonly consist of: (1) large data-sets containing images with corresponding gold standard image descriptions, (2) a computer vision system that is able to produce image descriptions by using object and scene recognition system (Elliott & Keller, 2013)(Vinyals et al., 2014)(Socher et al., 2014), (3) an automatic evaluation measure that evaluates automatic generated image descriptions, and (4) human evaluation of computer generated image descriptions, for the sake of evaluating the automatic evaluation measure. An visual representation of such a system can be seen in Figure 1.

![Figure 1: Visual representation of Image Description System](image)

Such measures have focused on computing a similarity score between the output of an MT system and reference sentences. **BLEU** (Papineni et al., 2002) is a measure of the effective overlap between human-created reference
sentences and a computer-generated candidate sentence, based on word error rate. While it is currently the most popular evaluation measure, the field of machine translation is argued to be overly reliant on BLEU (Callison-Burch et al., 2006). TER (Snover et al., 2006), a measure of translation error rate, calculates a score based on the number of modifications that transforms a candidate sentence to a gold-standard reference sentence. ROUGE (Lin, 2004), a popular evaluation measure from the summarization community, measures the longest common sub-sequence of tokens between a candidate and reference sentence. METEOR (Lavie, 2014) calculates a similarity score on sentence level based on exact, stemmed, synonym or paraphrase matching on word level. CIDEr (Vedantam et al., 2014) calculates similarity between a candidate and the consensus of multiple reference sentences, based on the average cosine similarity score, thereby inherently capturing notions of similarity, grammaticality, saliency, importance and accuracy.

These automatic evaluation measures are predominating research in the field of automatic image description, but without a well-grounded understanding of the validity of estimated correlations between these automatic metrics and human judgements, their practicality is debatable. A comparison done by Elliott & Keller (2014) lay bare the level of understanding of automatic evaluation tasks, considering state-of-the-art metrics show either weak or moderate correlated with human judgements. In addition, based on similar findings, it is argued that the field of machine translation is overly reliant on BLEU (Callison-Burch et al., 2006).

These findings could be seen as evidence showing that evaluation of computer generated text is still a challenging problem. Automatic evaluation measures that strongly correlate with human judgements are likely to be suitable performance indicators and could facilitate consistent, rapid and inexpensive evaluation of new ideas, algorithms, and data-sets to automatic image description research, concluding that the validity of automatic image description systems heavily rely on their evaluation measures. Therefor, for such a measure to be valid, an improved score should be necessary and sufficient for achieving actual improvements in the description of images using natural language. Such meaningful metrics should show a strong correlation with human judgements on the quality of computer-generated image description. Thus it is fruitful to create an understanding of how correlation between automatic evaluation measures and human judgements of sentences generated by automatic image description should be approached.

Techniques used in state-of-the-art evaluation measures are often adopted from related fields of work (e.g. speech recognition, automatic summarization and machine translation). The validity of these measures is not taken
without question, e.g. the validity of measures that exclusively focus on string level comparison is questioned by Reiter & Belz (2009), claiming they do not provide a useful measure of content quality. Further investigation into the practicality of automatic measures is done by estimating their correlation with human judgements (Elliott & Keller, 2014) (Lin & Och, 2004) (Chen et al., 2015). These results are limited by inconsistent human judgement data (Turian et al., 2006) and the use of different performance measures (Müller et al., 2001) generate results that are difficult to compare and interpret. While research on the agreement among different sources, e.g. expert and crowdsourced (Nowak & Rüger, 2010), and means, e.g. binary, relative or point-scale, to generate human judgements is sparse, issues concerning the measurements of such agreements are generally well-known, as described by Artstein & Poesio (2008).

By examining properties of frequently used measures, metrics, and judgements, a survey is presented. Firstly, by following the approach from Elliott & Keller (2014), correlation of evaluation of five automatic evaluation measures with human judgements is estimated, affirming reported performance. Secondly, the agreement between different annotators, in this case various approaches of human evaluation, is approximated to reveal influence of characteristic introduced by the method of collecting these judgements.

Results show evidence for a moderate correlation between top performing automatic evaluation measures and human judgements from various sources. Accompanying research into the agreement between expert and crowdsourced annotations support evidence that agreement lies above chance level, but is far below expectations and “ad hoc” standards. These results adhere to findings from related work (Elliott & Keller, 2014) (Passonneau et al., 2008), (Nowak & Rüger, 2010) (Reiter & Belz, 2009).

2 Methodology

Exploration of the conceptual, mathematical, and empirical implications of automatically evaluating image descriptions is critical for the validation of correlation between automatic evaluation measures and human judgements. An attempt to illustrate this consists of:

1. Re-evaluating correlation reported by Elliott & Keller (2014). This is done by calculating Spearman’s rank correlation coefficients regarding automatic measure scores and human judgements, include the Spearman’s correlation of CIDEr, a more recent proposed measure which has been reported to have a higher correlation with human judgements.
than existing measures. Computations are done using the Flickr8k data-set (Hodosh et al., 2013).

2. (1) Evaluating correlation between automatic evaluation measure scores and crowdsourced human judgements and (2) evaluating agreement between expert and crowdsources judgements, obtained using online labor markets such as Amazon’s Mechanical Turk (MTurk). Crowdsourcing has recently become a popular source of experimental data, and it is reasonable to have doubts about the quality of data provided by subjects recruited from such online labor markets. Although it has been argued that data obtained using MTurk could be at least as reliable as those obtained by traditional methods (Buhrmester et al., 2011) (Mason & Suri, 2012), it is valuable to validate the use crowdsourced judgements when appraising automatic evaluation measures within the field of automatic image description, for such sources of data could provide the means for rapid and inexpensive accumulation of high-quality data. On the other hand, fitting automatic evaluation measures to correlate with crowdsourced judgements would not achieve actual improvement in image description quality, when crowdsourced judgements would prove to be of poor descriptive quality.

2.1 Data-sets

Humans are effortlessly able to produce a description of an image that accurately identifies the portrayed objects and actions. Automatic generation of an accurate description of entities, events and scenes depicted in an image is possibly the most ambitious test of image understanding and natural language generation. Such an attempt would start with collecting images and corresponding descriptions, or using one of several available data-sets. First, the data-set should contain a large amount of images from a relative diverse domain. Second, the corresponding descriptions should be accurate and specific, and of human-like quality, for these will guide as gold-standards during evaluation.

My experiments are therefore based on the Flickr-8K data set (Hodosh et al., 2013). This data-set contains 8,092 images, each image is paired with five different captions that were purposely written to describe the image. Annotators were asked to write sentences that describe the depicted scenes, situations, events and entities. There is a considerable degree of variance in the way many images can be described, hence multiple captions are collected for each image. This data is evaluated using human studies consisting of
graded “expert” judgments, obtained from a small number of native American English speakers and relative “crowdsourced” human judgments, which are collected on a much larger scale using Mechanical Turk (MTurk), these relative judgements are eventually mapped to binary relevance judgments. All judgements are based purely on semantic properties, as the provided descriptions are written by people, lexical properties can thus be considered to be human-like.

Figure 2: Illustration by Hodosh et al. (2013), indicating the expert judgment rating scale, with actual examples from the data-set.

Figure 2 illustrates the rating scale, with actual image and description examples from the Flickr-8K data-set. The following indication of the rating scale was available for all expert judges: a score of 1 indicates no relation between the caption and the image, a score of 2 indicates the description is not useful, but describes some aspects of the image, a score of 3 almost describes the image (minor mistakes are allowed), and a score of 4 means the caption describes the image perfectly. For practical reasons, evaluation data is only collected for the highest correlated caption of each test image. An illustration of the data-structures of expert and human and judgements is presented in Table 1.

To collect human judgement on a larger scale, MTurk workers - sometimes referred to as *turkers* - were presented with images paired with ten different captions. Via checkboxes workers could indicate which of the captions describes the image, allowing for minor mistakes in the description, thus corresponding with a score of 3 or higher on the 4-point scale. Only judgements from workers who judged at least 70% of the control items correctly were collected, and each image-caption pair was annotated by three
to six judges.

### Expert Judgements

<table>
<thead>
<tr>
<th>image_id</th>
<th>caption_id</th>
<th>1-4 scale scores</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1056338697_4f7d7ce270.jpg</td>
<td>2718495608_d8533e3ac5.jpg#2</td>
<td>J_1:1, J_2:2, J_3:3</td>
<td></td>
</tr>
<tr>
<td>1119015538_e8c796281c.jpg</td>
<td>2534502836_7a75305655.jpg#2</td>
<td>J_1:2, J_2:3, J_3:4</td>
<td></td>
</tr>
</tbody>
</table>

### Crowdsourced Judgements

<table>
<thead>
<tr>
<th>image_id</th>
<th>caption_id</th>
<th>% yeses, # of yeses, # of noes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1056338697_4f7d7ce270.jpg</td>
<td>2248487950_c6230e31a9.jpg#2</td>
<td>%Y:0.33, #Y:1, #N:2</td>
<td></td>
</tr>
<tr>
<td>1056338697_4f7d7ce270.jpg</td>
<td>524105255_b346f288be.jpg#2</td>
<td>%Y:1, #Y:3, #N:0</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Illustration of structure of Expert and Crowdsourced Judgements found in the Flickr-8K data-set

### 2.2 Automatic Evaluation Measures

To appreciate and evaluate the characteristics of various automatic evaluation measures several popular measures are now described, illustrating their characteristic approach of computing a score expressing the quality of machine generated text. Such scores are based on strategies that attempt to measure one or several desirable properties of text. To provide a guideline for these measures, these properties are currently being measured using human studies by collecting human judgements on such properties including grammaticality, saliency, truthfulness or overall quality of image descriptions generated by computers.

**Bleu** (Papineni et al., 2002) is based on the central idea that good machine translation should be as close to a professional human translation as possible. This closeness is captured in a numerical metric, that is fashioned after the word-error-rate metric, modified for allowing legitimate differences in word choice and word order. Precision is computed by counting the number of word that occur in any of the reference sentences and dividing that number by the total number of words in the corpus C. Candidate word counts are clipped by the maximum count of that word in a reference sentence S to compensate improbable, high precision translations. The geometric mean of effective precision is calculated for a various lengths of n-grams, these are combined using the average logarithm with uniform weights. A sentence brevity penalty \( BP \) is introduced penalizing short lengthened, high-scoring candidate translations, through a comparison of the length of the candidate sentence \( c \) and the effective reference sentence length \( r \). A higher Bleu score is better. More formally,
\[ BP = \begin{cases} 
1 & \text{if } c > r \\
\frac{e^{(1-r/c)}}{c} & \text{if } c \leq r 
\end{cases} \]

then,

\[ \text{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \]

with,

\[ p_n = \frac{\sum_{S \in C} \sum_{\text{n-gram} \in S} \text{count}_{\text{clipped}}(\text{n-gram})}{\sum_{S \in C} \sum_{\text{n-gram} \in S} \text{count}(\text{n-gram})} \]

A similar approach is taken by the measure ROUGE (Lin, 2004), which counts the number of overlapping n-grams between a computer-generated candidate sentence and human-generated baseline sentences. Multiple ROUGE measures are available, where ROUGE-S, a skip-bigram version, contrary to BLEU, allows for parameterised intervening, meaning that a matching sequences can contain a skip-bigram, allowing for arbitrary gaps. Skip-bigrams are any pair of words in their sentence order. This way consecutive matches are not required, but sensitivity to word order is conserved. The length of the skip-bigrams are parameterised by \( d_{\text{skip}} \). A higher ROUGE score is better. The skip-bigram-based F-measure, given a reference sentence \( X \) of length \( m \) and a candidate description \( Y \) of length \( n \), is computed as follows:

\[ R_{\text{skip}2} = \frac{\text{SKIP2}(X,Y)}{C(m, 2)} \]

\[ P_{\text{skip}2} = \frac{\text{SKIP2}(X,Y)}{C(n, 2)} \]

\[ F_{\text{skip}2} = \frac{(1 + \beta^2)R_{\text{skip}2}P_{\text{skip}2}}{R_{\text{skip}2} + \beta^2 P_{\text{skip}2}} \]

Where \( \text{SKIP2}(X,Y) \) is the number of skip-bigram matches between \( X \) and \( Y \), \( \beta \) controls the relative importance of \( P_{\text{skip}2} \) and \( R_{\text{skip}2} \), and \( C \) is the combination function. To differentiate between candidate sentences that have no word-pair co-occurring with its reference but do have single-word co-occurrences, and sentences that do not have any word co-occurring, ROUGE-S is extended with the addition of unigram as counting units by adding a begin-of-sentence marker at the beginning of candidate and reference sentences. For the evaluation of the ROUGE metric, the extension called ROUGE-SU is used with \( d_{\text{skip}} = 4 \), for this setup shows good performance on headline-like summaries in an estimation of confidence intervals.
of correlations by Lin (2004), such headline-like summaries are similar to sentences found in image description tasks.

TER (Snover et al., 2006) takes a different approach to scoring “goodness” of machine generated sentence by computing the Translation Edit Rate (TER), which is in essence the minimum number of modifications that transforms a candidate sentence into a fluent sentences that has the correct meaning. A variant of this measure that finds the minimum of edits that transforms a machine generated sentence to a human-written reference sentence is defined as human-targeted TER (or HTER). This approach is applicable to the task of automatic image description, hence HTER will here be used to estimate correlation with human judgements. I note that a lower TER score is better. Specifically HTER is the number of edits to the closest reference, normalized by the average length of the references:

$$TER = \frac{\text{# of edits}}{\text{average # of reference words}}$$

Where # of edits is the minimum number of edits to transform a candidate sentence to the closest reference sentence. Possible edits include insertion, deletion and substitution of single words, as well as shifts of word sequences. All edits, have equal cost. The (de)-capitalization of words is treated as an edit and punctuation tokens are approached as normal words. Optimal calculation of edit-distance with move operations has been shown to be a NP-Complete problem (Shapira & Storer, 2002) and is calculated using dynamic programming. More recently TER-Plus (Snover et al., 2009) is presented as an extension of TER. TER-plus (TERp) addresses several of its weaknesses through the use of paraphrases, stemming, synonyms, as well as edit costs that can be automatically optimized to correlate better with various types of human judgments.

Similar to TERp, various measures have been proposed that focus on more intricate approaches trying to capture similarity in meaning alongside measuring the lexical similarity between sentences. For instance METEOR (Lavie, 2014), a measure of the harmonic mean of unigram precision and recall, calculates a similarity score on sentence-level. The space of possible alignments is constructed according to the following matchers:

1. Exact word matching, using the surface form of words.

2. Stemmed matching, using a language appropriate Snowball Stemmer (Porter, 2001), matches the stems are identical.

3. Synonym matching, occurs when words share membership in any synonym set according to the WordNet database (Miller, 1995).
4. Paraphrase matching, possible when both phrases are listen in a language appropriate paraphrase table.

By using machine learning techniques, a universal parameters set \((\alpha, \beta, \gamma, \delta \text{ and } w_1 \cdots w_n)\) and language specific function words and paraphrases are learned from the training bitext, allowing language specific evaluation to any target languages.

The METEOR score is calculated using a hypothesis \((h_c, h_f)\) and reference \((r_c, r_f)\) that consist of content and function words. For each matcher \(m_i\), the number of content and function words covered by this type of matching is counted in the hypothesis \((m_i(h_c), m_i(h_f))\) and reference \((m_i(r_c), m_i(r_f))\).

Weighted precision \((P)\) and recall \((R)\) is calculated, using matcher weights \((w_1 \cdots w_n)\) and content-function word weight \(\delta\), as follows:

\[
P = \frac{\sum_i w_i \cdot (\delta \cdot m_i(h_c) + (1 - \delta) \cdot m_i(h_f))}{\delta \cdot |h_c| + (1 - \delta) \cdot |h_f|}
\]

\[
R = \frac{\sum_i w_i \cdot (\delta \cdot m_i(r_c) + (1 - \delta) \cdot m_i(r_f))}{\delta \cdot |r_c| + (1 - \delta) \cdot |r_f|}
\]

The parameterized harmonic mean of precision and recall is then calculated:

\[
F_{\text{mean}} = \frac{R \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R}
\]

A fragmentation penalty is introduced to account for gaps and differences in word order, using the average number of matched words in the hypothesis and reference \((m)\) and the number of chunks \((ch)\):

\[
\text{Pen} = \gamma \cdot \left(\frac{ch}{m}\right)^\beta
\]

The METEOR score is then calculated: (note that a higher METEOR score is better)

\[
\text{METEOR} = (1 - \text{Pen}) \cdot F_{\text{mean}}
\]

CIDEr (Vedantam et al., 2014), a more recent measure that estimates the consensus of how most people describe an image, computing the average cosine similarity between a candidate description and the estimated consensus of a set of image descriptions \(S_i = \{s_{i1}, \cdots, s_{im}\}\), in an attempt of inherently capturing notions of similarity, grammaticality, saliency, importance and accuracy. The measure of consensus is based on the intuition that it should encode how often \(n\)-grams present in the candidate sentence
are simultaneously present in all reference sentences, with the added notion that n-grams not present in any of the reference sentences should not be present in the candidate sentence, assigning a lower-weight to n-grams that commonly occur across the whole data-set, hence are likely to be less informative. A higher CIDEr score is better. First the Term Frequency Inverse Document Frequency (TF-IDF) weighting \( g_k(s_{ij}) \) of each n-gram \( w_k \) is computed:

\[
TF = \frac{h_k(s_{ij})}{\sum_{w_l \in \Omega} h_l(s_{ij})}
\]

TF is the term frequency of each n-gram \( w_k \), placing higher weights on n-grams that frequently occur in the reference sentences describing an image.

\[
IDF = \frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))}
\]

IDF reduced the weight of n-grams that commonly occur in all reference sentences.

\[
g_k(s_{ij}) = TD \cdot \log(IDF)
\]

Where the number of times an n-gram \( w_k \) occurs in a reference sentence \( s_{ij} \) is denoted by \( h_k(s_{ij}) \) or \( h_k(s_{ij}) \) for the candidate sentence \( c_i \), \( \Omega \) is the vocabulary of all n-grams, and \( I \) is the set of images in the data-set. Then:

\[
\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{||g^n(c_i)|| \cdot ||g^n(s_{ij})||}
\]

Where \( m \) is the number of reference sentences, \( g^n(c_i) \) a vector of \( g_k(c_i) \) corresponding to all n-grams of length \( n \), \( ||g^n(c_i)|| \) is the magnitude of that vector. Similarly for \( g^n(s_{ij}) \). Scores of various n-gram lengths are combined to capture grammatical properties as well as richer semantics:

\[
\text{CIDEr} = \sum_{n=1}^N w_n \text{CIDEr}_n(c_i, S_i)
\]

Motivated by empirical results from Vedantam et al. (2014), uniform weights are chosen: \( w_n = \frac{1}{N} \) with \( N = 4 \).

### 2.3 Statistical Measures

Automatic evaluation measures that strongly correlate with human judgements on the quality of an image description sentence are likely to be suitable
performance indicators and could facilitate consistent, rapid and inexpensive evaluation of new ideas, algorithms, and data-sets used in automatic image description tasks. For such a measure to be valid, an improved metric score should be necessary and sufficient for achieving actual improvements in the description of images. Therefore it is essential to understand how correlation between automatic evaluation measures and human judgements of sentences generated by automatic image description should be approached. This would consequently facilitate easier and more accurate comparison of results from different analyses, furthermore it would aid in the evaluation of design-choices made when collecting data and conducting experiments.

2.3.1 Correlation measures

While no experiment can completely control for all sorts of complex variables, it is argued by Clark et al. (2011) that the impact of optimizer instability has been neglected by standard experimental methodology in machine translation research and suggest that replication should be adopted as standard practice, such statistical significance tests are provided in the form of bootstrap re-sampling methods (Koehn, 2004). However, if statistically reliable metric scores can substitute thorough human analysis is another debate. For this reason measuring the correlation of the automatic metrics with the human judgments of translation quality at the system-level is of practical use. Several statistical correlation coefficients are available to estimate correlation or agreement between automatic image description measures and human judgements, comparing these results proves to be difficult. Agreement scores are reported by Hodosh et al. (2013) using Cohen’s κ, unlike reported correlation scores using the Spearman’s rank correlation coefficient ρ (Elliott & Keller, 2014) (Callison-Burch et al., 2008), in Melamed et al. (2003) an argument is made to use Precision, Recall, and the F-measure, claiming for these standard measures for their intuitive interpretation and already broad practicing to evaluate natural language processing systems.

However (Powers, 2012, p. 1) argues: “It is becoming clear that traditional evaluation measures used in Computational Linguistics (including Error Rates, Accuracy, Recall, Precision and F-measure) are of limited value for unbiased evaluation of systems, and are not meaningful for comparison of algorithms unless both the dataset and algorithm parameters are strictly controlled for skew (Prevalence and Bias).”

As the collection of human judgments is often unsupervised, annota-
tions are likely to interact with the distributions of various categories in the labeled data, and with the strategies employed by individual annotators as reflected in their annotation decisions. (Passonneau et al., 2008). As stated in Elliott & Keller (2014), Cohen’s $\kappa$ requires transformation of real-valued scores into categorical values and is able to report agreement on just a thresholded subset of the data. Another, more suitable, coefficient is Spearman’s $\rho$, a non-parametric correlation coefficient, that is less sensitive to strong outlier data compared with Pearson’s $r$. For each data point (sample size = $n$), raw scores $X_i, Y_i$ are converted to rank $x_i, y_i$, now:

$$\rho = 1 - \frac{6 \sum (x_i - y_i)^2}{n(n^2 - 1)}$$

Spearman’s correlation coefficient ranges between $-1 \geq \rho \geq 1$, where coefficients -1 and 1 indicates perfect correlation and a lack of correlation is shows by a coefficient of 0. Further classification of the strength of correlation is, analogous with (Elliott & Keller, 2014), based on the guidance of Dancey & Reidy (2007). Spearman coefficients between 0.0-0.1 are uncorrelated, between 0.11-0.4 shows weak correlation, 0.41-0.7 is a moderate correlation, 0.71-0.9 is strong, and 0.91-1 is perfect correlation. Negative correlation can be classified identical using $|\rho|$, the absolute value of the coefficient. Such classification can be helpful for interpretation of results.

2.3.2 Annotation agreement measures

For the validation of the use of crowdsourced judgements, inter-annotation agreement studies are available and the methodology for large annotation efforts has improved, nonetheless they are often a race to report high scores and measure only percent agreement. This current state of affair is unsatisfactory, as weighted coefficients of agreement (e.g. Cohen’s $\kappa$ and Krippendorff’s $\alpha$) tend to provide more valuable indication of the quality of the resulting annotation agreement than percent agreement. Cohen’s $\kappa$ is defined as:

$$\kappa = \frac{\Pr(o) - \Pr(e)}{1 - \Pr(e)}$$

With $\Pr(o)$ as the relative observed agreement among raters, and $\Pr(e)$ being the probability of chance agreement (Cohen et al., 1960).

The general form of Krippendorff’s $\alpha$ (Krippendorff, 1970) is:

$$\alpha = 1 - \frac{D_o}{D_e}$$
Where $D_o$ is the observed disagreement among raters:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck} \text{metric} \delta_{ck}^2$$

And $D_e$ the disagreement expected to be attributable to chance:

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_k \text{metric} \delta_{ck}^2$$

Arguments (e.g. $o_{ck}, n_c, n_k, n$, and $\text{metric} \delta_{ck}^2$) refer to frequencies from coincidence matrices, defined in Krippendorff (2007). Both for Cohen’s $\kappa$ and Krippendorff’s $\alpha$ range is $1 \geq \text{metric} \geq 0$, when raters perfectly agree $\text{metric} = 1$ indicating perfect reliability, $\text{metric} = 0$ indicates raters are agreeing on chance level or, when $D_o = D_e$, an absence of reliability.

Unfortunately, weighted coefficients of agreement, while arguably more appropriate for annotation tasks that are common in evaluation of image descriptions, do not provide clear interpretation. Research should clearly report the methodology that was followed to collect the reliability data (number of judges, whether judges annotated independently, whether they relied exclusively on an annotation manual etc.). Atop of reporting the values of more than one coefficient of agreement, studies should also indicate whether agreement was statistically significant, and provide the confusion matrix or agreement table so that readers can find out whether overall figures of agreement hide disagreements on less common categories (Artstein & Poesio, 2008).

2.4 Protocol

My analyses of automatic evaluation measure performance consists of the following experiments:

1. Estimate correlation between automatic evaluation measures and expert human judgements.

For this re-evaluation of analysis done by (Elliott & Keller, 2014), sentence-level evaluation measure scores on tuples, consisting of an image description and a caption, are available, and are compared with corresponding expert human judgements using Spearman’s correlation estimate. Both data-files were first pre-processed. Automatic evaluation measure scores and judgement scores are put into vectors, these vectors are aligned and Spearman’s $\rho$ is estimated using `scipy.stats.spearmanr` (Zwillinger & Kokoska, 1999).
2. Estimate correlation between automatic evaluation measures and judgments on overlapped data.

To attain a comparison of automatic evaluation measure performance on expert and crowdsourced judgments, overlapped data is used. Image description and caption tuples that are annotated with both expert and crowdsourced judgments are selected by matching on image_id and reference_id columns by a python algorithm, a copy of the expert judgement vector is binarized by transforming low scores to 0: \((1, 2 \rightarrow 0)\) and high scores to 1: \((3, 4 \rightarrow 1)\). These score vectors are concatenated, together with corresponding automatic evaluation measure scores. A small number (1044) of crowdsourced judgement have more than three scores per image and caption combination. As the score vectors must be of the same length, scores that consist of more than three judgements are transposed to scores of three judgements, with respect to their original distribution. Using this data correlation is estimated using Spearman’s ρ.

3. Estimate correlation between automatic evaluation measures and crowdsourced judgements.

An attempt to validate the use of crowdsourced judgements to evaluate automatic evaluation measures is made by pre-processing the data to facilitate the use of the compareImageDescriptionMeasures framework (Elliott & Keller, 2014), that obtains BLEU, TER, and METEOR evaluation scores using MultEval (Clark et al., 2011) and ROUGE score using RELEASE-1.5.5.pl.\(^1\) To estimate the correlation, instead of using the included R script, Python’s implementation of Spearman’s ρ is used, making fair comparison with previous experiments possible.

4. Estimate agreement between expert and crowdsourced judgements.

The use of crowdsourced judgements is common in recent research, as is allows for cheap and fast annotation on a large-scale. However research that validates the sole use of crowdsourced judgements, as opposed to expert judgements or a combination of both, is sparse. Furthermore coefficients like Krippendorff’s α are hard to interpret without a given context. Here an attempt is made to validate the use of crowdsourced judgement by investigating expert and crowdsourced inter-annotation agreement. To facilitate accurate interpretation and comparison, several agreement coefficients are reported. Calculations are made using the online framework by Geertzen

\(^1\)http://www.isi.edu/licensed-sw/see/rouge/
Input is composed of a binarized expert judgement scores and crowd-sourced judgement scores concatenated in a vector:

\[
\begin{bmatrix}
\text{expert}_0 & \text{crowd}_0 \\
\text{expert}_1 & \text{crowd}_2 \\
\vdots & \vdots \\
\text{expert}_m & \text{crowd}_m
\end{bmatrix}
\]

Where expert\(_n\) and crowd\(_n\) both correspond to the same image and caption combination.

All code is released and available on GitHub.\(^2\) Likewise for the compareImageDescriptionMeasures framework (Elliott & Keller, 2014).\(^3\)

### 3 Results

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>E&amp;K(2014)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meteor</td>
<td>0.520</td>
<td>0.524</td>
</tr>
<tr>
<td>Rouge SU-4</td>
<td>0.437</td>
<td>0.435</td>
</tr>
<tr>
<td>Bleu-4</td>
<td>0.426</td>
<td>0.429</td>
</tr>
<tr>
<td>Ter</td>
<td>-0.278</td>
<td>-0.279</td>
</tr>
<tr>
<td>Cider</td>
<td>0.578</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Spearman’s \(\rho\) estimate of automatic evaluation measures against expert human judgements on the Flickr-8K data-set, compared with results by Elliott & Keller (2014).

Table 2 shows the estimate correlation co-efficients between automatic evaluation measures (Ter, Smoothed BLEU, Rouge-SU4, Meteor, and Cider) and expert human judgements, compared to results from Elliott & Keller (2014). Table 3 shows the correlation estimate computed against expert judgements, binarized expert judgements and crowdsourced judgements of 6798 corresponding image and caption combinations that occur in both the expert and crowdsourced data-set. In Table 4 the correlation estimates computed on the CrowdFlower data-set, annotated using crowdsourced human judgements, are reported. All reported correlations are significant at \(p < 0.001\).

\(^2\)https://github.com/jeroen-rooijmans/ThesisProject2015.git

\(^3\)https://github.com/elliottd/compareImageDescriptionMeasures.git
<table>
<thead>
<tr>
<th></th>
<th>Expert</th>
<th>Expert\textsubscript{bin}</th>
<th>Crowdsourced</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>METEOR</strong></td>
<td>0.517</td>
<td>0.438</td>
<td>0.403</td>
</tr>
<tr>
<td><strong>ROUGE SU-4</strong></td>
<td>0.481</td>
<td>0.408</td>
<td>0.387</td>
</tr>
<tr>
<td><strong>BLEU-4</strong></td>
<td>0.464</td>
<td>0.410</td>
<td>0.420</td>
</tr>
<tr>
<td><strong>TER</strong></td>
<td>-0.379</td>
<td>-0.360</td>
<td>-0.392</td>
</tr>
<tr>
<td><strong>CIDEr</strong></td>
<td>0.601</td>
<td>0.511</td>
<td>0.475</td>
</tr>
</tbody>
</table>

Table 3: Spearman’s $\rho$ estimate of automatic evaluation measures against expert, binarized expert and crowdsourced human judgements on a matched data-set.

<table>
<thead>
<tr>
<th></th>
<th>CrowdFlower</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>METEOR</strong></td>
<td>0.248</td>
</tr>
<tr>
<td><strong>ROUGE SU-4</strong></td>
<td>-0.025</td>
</tr>
<tr>
<td><strong>BLEU-4</strong></td>
<td>0.236</td>
</tr>
<tr>
<td><strong>TER</strong></td>
<td>-0.198</td>
</tr>
</tbody>
</table>

Table 4: Spearman’s $\rho$ estimate of automatic evaluation measures against crowdsourced human judgements.

Figure 3 shows inter-annotation agreement between 6798 corresponding expert judgement (binarized) and crowdsourced judgements. Several agreement coefficients, including Fleiss’ $\kappa$, Krippendorff’s $\alpha$, are reported. Calculations are made using the online framework by Geertzen (2012).\footnote{https://mlnl.net/jg/software/ira/}

4 Evaluation

Correlation estimate results on the Flickr-8K data-set with expert human judgements (Table 2) are similar as observed in the Elliott & Keller (2014) paper. Results show a weak correlation with expert human judgements for TER, and moderate correlation for BLEU, ROUGE-SU4, METEOR, and CIDEr, with the recent proposed measure CIDEr showing the strongest correlation. Small variation in the coefficient estimates can be described to candidate and reference sentences combinations with an equal rank, these tied-ranks are handled in different ways by Spearman’s ranked correlation coefficient implementations in R and Python, explaining the small differences in the results.

Results on the matched data-set (Table 3) show a similar pattern, how-
ever a variation in the correlation with expert judgements compared to the results in Table 2 exist. This difference can be caused by the difference of the data-set size (17466 judgements in Flickr-8K and only 6798 judgements in the matched data-set). Correlation estimates drops when expert judgements are binarized, with significant drops for METEOR, ROUGE-SU4, and CIDEr. Such a decrease of correlation could be expected when masking more fine-grained scores to only two categories, thus losing information, but could also show that reported automatic evaluation measures have a preference for fine-grained judgements. Standard variation measures show: mean\text{expert}(\rho) = 0.488 with \sigma = 0.08, mean\text{bin(expert)}(\rho) = 0.425 with \sigma = 0.06, and mean\text{crowd}(\rho) = 0.415 with \sigma = 0.04.

Experiments on correlation between automatic evaluation measure scores and crowdsourced judgements show unsatisfying results (Table 4). Correlation estimates are far below expectation, also considering the correlation pattern found between metric scores and crowdsourced judgement that overlapped with expert judgements. Further evaluation is needed to verify these results.

An inter-annotator agreement (Table 3) of 0.481 is estimated using Fleiss’ \( \kappa \) and Krippendorff’s \( \alpha \). Showing agreement almost halfway between chance and perfect agreement. This is comparable with the Cohen’s \( \kappa \) score of 0.495. As these measures are closely related, similar scores are to be expected.
5 Discussion

Apart from small differences in the estimated coefficients, allocated to tied-rank handling by different Spearman’s ρ implementation in R and Python, results validate the analysis of Elliott & Keller (2014). By using Spearman’s ρ to report correlation coefficients, there is no need to categorizes results, therefore all real-valued data is applicable without any transformation, as apposed when using the correlation coefficient Cohen’s κ. Comparison to other related work (Hodosh et al., 2013) (Turian et al., 2006) (Reiter & Belz, 2009) is difficult due to differences in analyses, mainly results of difference in reported correlation coefficients, used data and judgements, and reported automatic evaluation measures. However conclusion drawn across these surveys are similar as those presented here, suggesting CIDEr and METEOR as the most appropriate measures for image description performance. Such conclusions agree with suggestions by Vedantam et al. (2014), claiming that adding word-level semantic matching (e.g. synonym and paraphrase matching) could improved performance on data-sets with a “small” number of reference descriptions per image (1-3 sentences per image). Current data-sets commonly allow evaluation at up to 5 sentences, thus it is reasonable that such extensions result into improvements for current measures that focus on lexical properties of description sentences.

Discrepancy between fine-grained judgements and their binarized rendition can be accredit to the loss of information. Considering that stronger correlated measure experience the greatest variance due to the binarization, it could suggest these measure are correlated with more fine-grained judgements. To further validate these findings I suggest similar research is done using a larger data-set. Correlation estimated between automatic evaluation measures TER, BLEU, ROUGE, and METEOR and the crowdsourced judgement data-set CrowdFlower is far below expectations. Future work is needed to evaluate the methodology that is carried out here. I suspect that given the results, alignment between presented images, candidate descriptions, and reference descriptions is lost in the process of obtaining scores of the automatic evaluation measures using the MultiEval (Clark et al., 2011) and RELEASE-1.5.5 packages.

Interpretation of results regarding inter-annotation agreement is difficult, given the nature of weighted agreement coefficients. There does not seem to be a general standard on how to report inter-annotation agreement results, but interpretation and comparison can be facilitated by reporting several different agreement coefficients are reported, including Krippendorff’s α and Fleiss’ κ as these measures calculate the degree of agreement over that
which would be expected by chance and can assess multiple raters simultaneously. Krippendorff’s $\alpha$ is applicable to various numbers of judgements (binary, nominal, ordinal, interval, ratio, polar, and circular metrics) and also generalizes with several other measures of inter-coder agreement (Scott’s $\pi$, Fleiss’ $\kappa$, Spearman’s $\rho$, and Pearson’s $r$) (Krippendorff, 1970). Supplementary work, using a larger data-set containing expert and crowdsourced judgement collected using the same means, could provide additional insight in the relation between crowdsourced and expert human judgements.

As there is no clear consensus on what constitutes a good image description, further research into the performance and validity of current automatic evaluation measure is necessary to establish automatic evaluation measures for which an improved scores is necessary and sufficient for achieving actual improvements in the description of images. Such future work could frame image description as a ranking task (Hodosh et al., 2013), as this feels inherently more similar to how humans evaluate image descriptions. Furthermore researchers should be cautious optimizing performance of image description systems with weak correlated measures, as over-fitting is a potential pitfall. Results\(^5\) from the recent MS COCO Captioning Challenge 2015 (Chen et al., 2015) show how problematic measuring performance of automatic evaluation measures is, as automatic image description system CIDEr-D, among others, currently outperform humans in their image description task.

### 6 Conclusions

During this thesis, sentence-level correlation analysis of automatic evaluation measures against expert, binarized expert, and crowdsourced human judgements is performed, to validate measure performance for the automatic image description tasks. It is found that TER shows a weak correlation with human judgements, smoothed BLEU, skip-bigram ROUGE, METEOR, and CIDEr show stronger, but moderate correlations. With CIDEr showing the strongest correlation overall, followed by METEOR. I recommend adopting evaluation measures that show strong correlation with human judgements, either CIDEr or METEOR. Results of analysis of the agreement between binarized expert and crowdsourced judgements show above chance agreement, but do not validate the use of crowdsourced judgement on its own, suggesting careful and appropriate collection of human judgement is necessary to further improve automatic image description systems.

\(^5\)http://mscoco.org/dataset/#leaderboard-cap
References


annual meeting on association for computational linguistics (pp. 311–318). Stroudsburg, PA, USA: Association for Computational Linguistics.


