# Semantic Tuples for Evaluation of Image to Sentence Generation

Lily D. Ellebracht[1], Arnau Ramisa[1], Pranava Swaroop Madhyastha[2],
Jose Cordero-Rama[1], Francesc Moreno-Noguer[1], and Ariadna Quattoni[3]

[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC
[2]TALP Research Center, UPC
[3]Xerox Research Centre Europe

## Abstract

The automatic generation of image captions has received considerable attention. The problem of evaluating caption generation systems, though, has not been that much explored. We propose a novel evaluation approach based on comparing the underlying visual semantics of the candidate and ground-truth captions. With this goal in mind we have defined a semantic representation for visually descriptive language and have augmented a subset of the Flickr-8K dataset with semantic annotations. Our evaluation metric (BAST) can be used not only to compare systems but also to do error analysis and get a better understanding of the type of mistakes a system does. To compute BAST we need to predict the semantic representation for the automatically generated captions. We use the Flickr-ST dataset to train classifiers that predict STs so that evaluation can be fully automated [1].

## 1 Introduction

In recent years, the task of automatically generating image captions has received considerable attention. The task of evaluating such sentences, though, has not been that much explored, and mainly holds on metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2003), originally proposed for evaluating machine translation systems. These metrics have been shown to poorly correlate with human evaluations (Vedantam et al., 2014). Their main problem comes from the fact that they uniquely consider n-grams agreement between the reference and candidate sentences, focusing thus only on the lexical information and obviating the agreement at the visual semantic level. These limitations are illustrated in Figure 1.

Vedantam et al. (2014) have proposed to address these limitations by making use of a Term Frequency Inverse Document Frequency (TF-IDF) that places higher weight on n-grams that frequently occur in the reference sentence describing an image, while reducing the influence of popular words that are likely to be less visually informative.

In this paper, we consider a different alternative to overcome the limitations of BLEU and ROUGE metrics, by introducing a novel approach specifically tailored to evaluate systems for image caption generation. To do this, we first define a semantic representation for visually descriptive language, that allows measuring to which extent an automatically generated caption of an image matches the underlying visual semantics of human authored captions.

To implement this idea we have augmented a subset of the Flickr-8K dataset (Nowak and Huiskes, 2010) with a visual semantic representation, which we call Semantic Tuples (ST). This representation shares some similarity with the more standard PropBank (Kingsbury and Palmer, 2002) style Semantic Roles (SRL). However, SRL was designed to have high coverage of all the linguistic phenomena present in natural language sentences. In contrast, our ST representation is simpler and focuses on the aspects of the predicate structure that are most relevant for capturing the semantics of visually descriptive language.

This ST representation is then used to measure the agreement between the underlying semantics of an automatically generated caption and the semantics of the gold reference captions at different levels of granularity. We do this by aggregating the STs from the gold captions and forming a Bag of Aggregated Semantic Tuples represen-

[1]System and data are made available here: `https://github.com/f00barin/semtuples`

Ref: A man sliding down a huge sand dune on a sunny day

SA: A man slides during the day on a dune.

SB: A dinosaur eats huge sand and remembers a sunny day.

| System | 1-gram | 2-gram | 3-gram | 4-gram |
|---|---|---|---|---|
| A | 0.47 | 0.29 | 0.16 | 0.11 |
| B | **0.49** | **0.36** | **0.23** | **0.17** |

Figure 1: The limitations of the BLEU evaluation metric: **SA** and **SB** are two automatically generated sentences that we wish to compare against the manually authored **Ref**. However, while **SB** does not relate to the image, it obtains higher n-gram similarity than **SA**, which is the basis of BLEU and ROUGE.

tation (BAST) that describes the image. We do the same for the automatically generated sentences and compute standard agreement metrics between the gold and predicted BAST. One of the appeals of the proposed metric is that it can be used not only to compare systems but also to do error analysis and get a better understanding of the type of mistakes a system does.

In the experimental section we use the ST augmented portion of the Flickr-8K dataset (Flickr-ST) as a benchmark to evaluate two publicly available pre-trained models of the Multimodal Recurrent Neural Network proposed by (Vinyals et al., 2014) and (Karpathy and Fei-Fei, 2014) that generate image captions directly from images. To compute BAST we need to predict STs for the automatically generated captions. This is sub-optimal because, ideally, we would like a metric that can be computed without human intervention. We therefore use the Flickr-ST dataset to train classifiers that predict STs from sentences. While this might add some noise to the evaluation, we show that the STs can be predicted from sentences with a reasonable accuracy and that they can be used as a good proxy for the human annotated STs.

In summary our main contributions are:

- A definition of a linguistic representation (the ST representation) that models the relevant semantics of visually descriptive language.
- Using ST we propose a new approach to evaluate sentence generation systems that measures caption-gold agreement with respect to the underlying visual semantics expressed in the reference captions.
- A new dataset (Flickr-ST) of captions augmented with corresponding semantic tuples.
- A new metric BAST (Bag of Aggregated Semantic Tuples) to compare systems. In addition, this metric is useful to understand the types of errors made by the systems.
- A new fully automated metric that uses trained classifiers to predict STs for candidate sentences.

The rest of the paper is organized as follows: Section 2 presents the evaluation approach, including the proposed ST representation, the human annotation process to produce a dataset of captions and STs and the proposed BAST metric computed over the ST representation. Section 3 describes in detail the proposed BAST metric. Section 4 describes the annotation process and the creation of the Flickr-ST dataset. Section 5 gives some details about the automatic sentence to ST predictors used to compute the (fully automatic) BAST metric. Section 6 discusses related work. Finally, Section 7 presents experiments using the proposed metric to evaluate state-of-the-art Multimodal Recurrent Neural Networks for caption generation.

## 2 Semantic Representation of Visually Descriptive Language

We next describe our approach for evaluating sentence generation systems. Figure 3 illustrates the steps involved in the evaluation of a generated caption. Given a caption we first generate a set of semantic tuples (STs) which capture the underlying semantics. While these STs could be generated by human annotators this will not be feasible for an arbitrarily large number of generated captions. Thus, in Section 5 we describe an approach to automatically generate STs from captions.

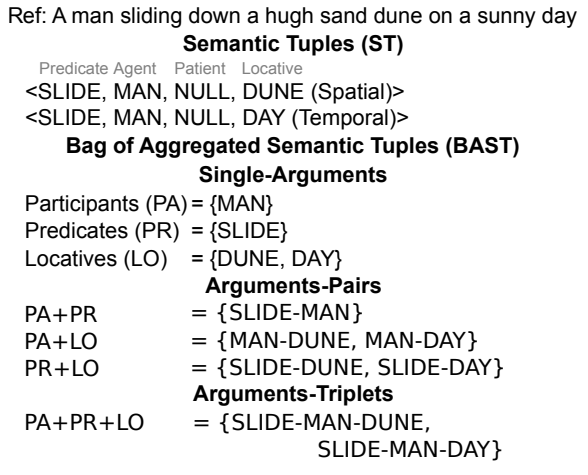


Figure 2: Bag of Aggregated Semantic Tuples.

In the second step of the evaluation we map the set of STs for the caption to a bag of arguments representation which we call BAST. Finally, we compare the BAST of the caption to that of the gold captions. The proposed metric allows us to measure the precision and recall of a system in predicting different components of the underlying visual semantics.

In order to define a useful semantic representation of Visually Descriptive Language (VDL) (Gaizauskas et al., 2015) we follow a basic design principle: we strive for the simplest representation that can cover most of the salient information encoded in VDL and that will result in annotations that are not too sparse. The last requirement means that in many cases we will prefer to map two slightly different visual concepts to the same semantic argument and produce a coarser semantic representation.

In contrast, the PropBank representation (SRL) (Kingsbury and Palmer, 2002) is what we would call a fine-grained representation which was designed with the goal of covering a wide range of semantic phenomena, i.e. cover small variations in semantic content. Furthermore, the SRL representation is designed so that it can represent the semantics of any natural language sentence whereas our representation focuses on covering the semantics present in VDL. Our definitions of semantic tuples are more similar to the proto-roles described by Dowty (1991).

Given an image caption we wish to generate a representation that captures the main underlying visual semantics in terms of the events or actions (we call them predicates), who and what are the participants (we call them agents and patients) and where or when is the action taking place (we call them locatives). For example, the caption "A brown dog is playing and holding a ball in a crowded park" would have the associated semantic tuple: [predicate = *play*; agent = *dog*; patient = *null*; locative = *park*] and [predicate = *hold*; agent = *dog*; patient = *ball*; locative = *park*]. We call each field of a tuple an argument; an argument consists of a semantic type and a set of values. For example the first argument of the first semantic tuple is a predicate with value *play*. Notice that arguments of type agent, patient and locative can take more than one value. For example: "A young girl and an old woman eat fruits and bread in a park on a sunny day" will have the associated semantic tuple: [predicate = *eat*; agent = *girl, woman*; patient = *fruits, bread*; locative = *park, day*].

Note also that we use italics to represent argument values and distinguish them from variables (over some well defined discrete domain) and words or phrases in the caption that we might regard as lexical evidence for that value. For example, the caption "A brown dog is playing and holding a ball in a crowded park" will have the associated semantic tuple: [predicate = *play*; agent = *dog*; patient = *null*; locative = *park*]. The word associated with the predicate *play* is playing, but *play* is a variable. In this case we are assuming that the domain for the predicate variable is the set of all lemmatized verbs.

Argument values will in most cases have some word or phrase in the caption that can be regarded as the lexical realization of the value. We refer to such a realization as the 'span' of the value on the caption. From the previous example, the span of the predicate is 'playing', and its value is *play*. Not all values will have an associated span, since as we describe below, argument values might have tacit spans which can be inferred from the information contained in the caption but they are not explicitly mentioned. In practice to generate the semantic representation we will ask human annotators to mark the spans in the caption corresponding to the argument values (for non-tacit values). We will define the argument variable to be a 'canonical' representation of the span. How this 'canonical' representation is defined will be described in more detail in the next section, where we discuss the annotation process.


Figure 3: Computation of the BAST metric.

## 3 The Bag of Semantic Tuples Metric

As mentioned earlier, our semantic representation is 'coarser' than PropBank style semantic role annotations. Furthermore, there are two other important differences: 1) We do not represent the semantics of atomic sentences but that of captions that might actually consist of multiple sentences, and 2) Our representation is truly semantic meaning that resolving the argument value of a predicate might involve making logical inferences. For example we would annotate the caption: "A man is standing on the street. He is holding a camera" with [predicate = *standing*; agent = *man*; patient = *null*; locative = *street*] and [predicate = *hold*; agent = *man*; patient = *null*; locative = *street*]. This means that in contrast to the SRL representation, our semantic representation will not, in general, be 'aligned' with the syntax of the caption.

We now give a more detailed description of each argument type:

- The **Predicate** is the main event described by the sentence. We consider two types of predicates, those that describe an action and those that describe a state. Action predicates are in most cases expressed in the caption using verb-phrases. However, some action predicates might not be explicitly mentioned in the caption but can be naturally inferred. For example, the caption "A woman in a dark blue coat, cigarette in hand" would be annotated with the tuple: [predicate = *hold*; agent = *woman*; patient = *cigarette*; locative = *null*]. In the case that the predicate is indicating a state of being, there is typically a conjugation of the verb "to be", i.e. is, are, was. For example: "A person is in the air on a bike near a body of water."

- The **Agent** is defined as the entity that is performing the action. Roughly speaking, it is the answer to the question: Who is doing the action? For example: in the sentence "The man is sleeping under a blanket in the street

as the crowds pass by" we have the predicate = *sleeping* with agent = *man*, and predicate = *pass* with agent = *crowd*. In the case of predicates that describe a state of being such as "A person is in the air on a bike near a body of water", we define the agent to be the answer to the question: Whose state is the predicate describing? Thus for the given example we would have agent = *person*.

- The **Patient** is the entity that undergoes a state of change or is affected by the agent performing some action. For example, the caption "A woman in a dark blue coat, cigarette in hand." would have: [patient = *cigarette*]. Unlike the predicate and agent, the patient is not always present, for example in "Two people run in the sand at the beach." The patient is never present with state-of-being predicates as: "A person is in the air on a bike near a body of water". When there is no patient we say that the argument value is *null*.

- The **Locative** is defined as the answer to the question: Where or When is the action taking place? So there are two main types of locatives, spatial locatives such as *on the water* and temporal locatives such as *at night*. Spatial locatives in turn can be of different types, they can be scenes such as *on-beach* or they can express the relative location of the action with respect to a reference object such as *under-blanket* in the caption "A man sleeping under the blanket". The locatives are actually composed of two parts: a preposition (if present), which expresses the temporal or spatial situation, and the main object or scene. Locatives, like the patient, are not always present. Thus the locative might also take the value *null*.

We could also consider a richer semantic representation that includes modifiers of the arguments, for example for the caption: "A brown dog is playing and holding a ball in a crowded park" we would have the associated semantic tuples: [predicate = *play*; agent = *dog*; agent-mod = *brown* patient = *null*; locative = *park*] and [predicate = *hold*; agent = *dog*; patient = *ball*; locative = *park*, locative-mod = *crowded*]. For the first version of the ST dataset, however, we opted for keeping the representation as simple as possible and decided not to annotate argument modifiers. One of the reasons is that we observed that in most cases if we can properly identify the main arguments extracting their modifiers can be done automatically by looking at the syntactic structure of the sentence. For example if we can obtain a dependency parse tree for the reference caption, extracting the syntactic modifiers of *dog* is relatively easy.

## 4 The Flickr-ST Dataset: Human Annotation of Semantic Tuples

We believe that one of the main reasons why most of the evaluations used to measure caption generation performance involve computing surface metrics is that until now there was no dataset annotated with underlying semantics.

To address this limitation we decided to create a new dataset of images annotated with semantic tuples as described in the previous section. Our dataset has the advantage that every image is annotated with both the underlying semantics in the form of semantic tuples and natural language captions that constitute different lexical realizations of the underlying visual semantics. To create our dataset we used a subset of the Flickr-8K dataset with captions, proposed in (Hodosh et al., 2013). This dataset consists of 8,000 images of people and animals performing some action taken from Flickr, with five crowd-sourced descriptive captions for each one. These captions are sought to be concrete descriptions of what can be seen in the image rather than abstract or conceptual descriptions of non-visible elements (e.g. people or street names, or the mood of the image).

We asked human annotators to annotate 250 image captions, corresponding to 50 images taken from the development set of Flickr-8K. In order to ensure the alignment between the information contained in the captions and their corresponding semantic tuples, annotators were not allowed to look at the referent image while annotating every caption.

Annotators were asked to list all the unique tuples present in the caption. Then, for each argument of the tuple, they had to decide if its value is *null*, *tacit* or *explicit* (i.e. an argument value that can be associated with a text span in the caption). For explicit argument values we asked the annotator to mark the corresponding span in the text. That is, instead of giving a value for the argument, we ask them to mark in the caption the evidence for that argument.

To create the STs that we use for evaluation we first need to compute the argument values. We assume that we can compute a function that maps spans of text to argument variables, and we call this the grounding function. Currently, we use a very simple mapping from spans to argument values: they map to lowercase lemmatized forms. Given the annotated data and a grounding function, we refer to the process of computing argument values for argument spans as projecting the annotations.

With our approach for decoupling surface (i.e. argument spans) from semantics (argument values) we can address some common problems in caption generation evaluation. The idea is simple, we can use the same annotation with different grounding functions to get useful projections of the original annotation. One clear problem when evaluating caption generation systems is how to handle synonymity, i.e. the fact that two surface forms might refer to the same semantic concept. For example, if the reference caption is: "A boy is playing in a park", the candidate caption: "A kid playing on the park" should not be penalized for using the surface form boy instead of kid. We can address this problem by building a grounding function that maps the argument span boy and the argument span kid to the same argument variable. We could automatically build such function using a thesaurus.

Another common problem when evaluating caption generation is the fact that the same visual entity can be described with different levels of specificity. For example, for the previous reference caption it is clear that "A person is playing in a park" should have a higher evaluation score than "A dog playing in a park". This is because any human reading the caption would agree that person is just a 'coarser' way of referring to the same entity. With our approach we could handle this problem by having a coarser grounding function that maps the argument span kid and the argument span person to the same argument value *human*. The important thing is that for any grounding function we can project the annotations and compute the evaluation, thus we can analyze the performance of a system in different dimensions.

Our goal is to define an evaluation metric that measures the similarity between the STs of the ground-truth captions for an image and the STs of a generated image caption. We wish to define a metric that is useful not only to compare systems, but also that allows for error analysis and some insight on the types of mistakes performed by any given system.

To do this we will first use the STs corresponding to the ground-truth captions to compute what we call a Bag of Aggregated Semantic Tuples representation (BAST). Figure 2 shows a reference caption and its corresponding STs and BAST. Notice that for simplicity we show a single reference caption, in reality if there are $k$ captions for an image, we will first compute the STs corresponding to all of them. The BAST representation is computed in the following manner:

1. For the locatives and predicate arguments compute the union of all the corresponding argument values appearing in any ST. For the patient and agent we will compute a single set which we refer to as the *participants* set. We call this portion of the BAST the bag of single arguments representation.

2. We compute the same representation but now we look at pairs of argument values, meaning: predicate+participant, participant+locative and predicate+locative. We call these the bag of argument pairs.

3. Similarly we can also compute a bag of argument triplets for predicate+participant+locative

We can also compute the BAST representation of an automatically generated caption. This can be done via human annotation of the caption's STs or using a model that predicts STs from captions (such a model is described in the next section). Now if we have the ground-truth BAST and the BAST of the candidate caption we can compute standard precision, recall and F1 metrics over the different components of the BAST. More specifically, for the single argument component of the BAST we compute:

- Predicate-Precision: This is the number of predicted predicates present in the BAST of the candidate caption that where also present in the BAST of the ground-truth reference captions for the corresponding image. That is this is the number of correctly predicted predicates.

- Predicate-Recall: This is the number of predicted predicates present in the BAST of the ground-truth captions that were also present in the BAST of the candidate caption.

- Predicate-F1: This is the standard metric, i.e. the harmonic mean of precision and recall.

We can compute the same metrics for other arguments and for argument pairs and triplets of arguments. Figure 3 shows an example of computing the BAST evaluation metric for two captions.

## 5 Automatic Prediction of Semantic Tuples from Captions

To compute the BAST metric we need to have STs for the candidate captions, one option is to perform a human annotation. The problem is that collecting human annotations is an expensive and time consuming task. Instead we would prefer to have a fully automated metric. In our case that means that we need an automated way of generating STs for candidate captions. We show in this section that we can use the Flickr-ST dataset to train a model that maps captions to their underlying ST representation.

We would like to point out that while this task has some similarities to semantic-role labeling, it is different enough so that the STs can not be directly derived from the output of an SRL system, in fact our model uses the output of an SRL system in conjunction with other lexical and syntactic features.

Our model exploits several linguistic features of the caption extracted with state-of-the-art tools. These features range from shallow part of speech tags to dependency parsing and semantic role labeling(SRL). More specifically, we use the FreeLing lemmatizer (Carreras et al., 2004), Stanford part of speech(POS) tagger (Toutanova et al., 2003), TurboParser (Martins et al., 2013) for dependency parsing and Senna (Collobert et al., 2011) for semantic role labeling. We also tried using state-of-the art SRL system from Roth and Woodsend (2014), but we observed that Senna performed better on our dataset.

We extract the predicates by looking at the words tagged as verbs by the POS tagger. Then, the extraction of arguments for each predicate is resolved as a classification problem. More specifically, for each detected predicate in a sentence we regard each noun as a positive or negative training example of a given relation depending on whether the candidate noun is or is not an argument of the predicate. We use these examples to train an SVM that decides if a candidate noun is or is not an argument of a given predicate in a given sentence. This classifier exploits several linguistic features computed over the syntactic path of the dependency tree connecting the candidate noun and the predicate and features of the predicted semantic roles of the predicate.

| | Model 1 | Model 2 |
|---|---|---|
| Participants (PA) | 0.967 | 0.865 |
| Predicates (PR) | 0.703 | 0.808 |
| Locatives (LO) | 0.793 | 0.819 |
| PA-PR | 0.884 | 0.812 |
| PR-LO | 0.779 | 0.723 |
| PA-LO | 0.849 | 0.757 |
| PA-PR-LO | 0.815 | 0.704 |

Table 1: F1 score of the automatic BAST extractor taking as reference the manually annotated tuples for the sentences generated by the two models.

Table 1 shows the F1 of our predicted STs compared against manually annotated STs for the two caption generation systems that we evaluate in the experiments section.

## 6 Related Work

Our definition of semantic tuple is reminiscent in spirit to Farhadi et al. (2010) scene-object-action triplets. In that work, the authors proposed to use a triplet meaning representation as a bridge between images and natural language descriptions. However, the similarity ends there because their goal was neither to develop a formal semantic representation of VDL nor to provide a semantically annotated dataset that could be used for automatic evaluation of captioning systems. At the end, their dataset was created in a very simplistic manner by extracting subject-verb, object-verb and locative-verb pairs from a labeled dependency tree by checking for dependencies where the head and modifier matched a small fix set of possible objects, actions and scenes. As we have illustrated with multiple caption examples, the semantics of VDL can be quite complex and it can be very 'loosely aligned' with the syntactic (e.g. dependency structure) of the sentence. There has also been some recent work on semantic image

retrieval based on scene graphs (Johnson et al., 2015), where they model semantic representation of image content to retrieve semantically related images.

BLEU has been the most popular metric used for evaluation, its limitations when used in the context of evaluation of caption quality have been investigated in several works (Kulkarni et al., 2013; Elliott and Keller, 2013; Callison-Burch et al., 2006; Hodosh et al., 2013). Another common metric is ROUGE which has been shown to have some weak correlation with human evaluations (Elliott and Keller, 2013). An alternative metric for caption evaluation is METEOR which seems to be better correlated with human evaluations than BLEU and ROUGE (Elliott and Keller, 2014). Recently a new consensus based metric was proposed by Vedantam et al. (2014), here, the main idea is to measure similarity of a caption to the majority of ground-truth reference captions. One of the limitations of metrics based on consensus is that they are better suited for cases when many ground-truth annotations exist for each image. We take a different approach, instead of augmenting a dataset with more captions, we directly augment it with annotations which reflect what are the most relevant pieces of information in the available captions.

Hodosh et al. (2013) propose a different metric for evaluating image-caption ranking systems and it can not be directly applied to evaluate sentence generation systems (i.e. systems that output novel sentences).

## 7 Experiments

### 7.1 The evaluated models

The evaluated models are two instances of the Multimodal Recurrent Neural Network described in (Simonyan and Zisserman, 2014a) and (Karpathy and Fei-Fei, 2014), that takes an image and generates a caption. content of the image in natural language).

This model addresses the caption generation task combining recent advances in Machine Translation and Image Recognition: it combines a Convolutional Neural Network (CNN) initially trained to extract image features, and a Long Short Term Memory Recurrent Neural Network (RNN-LSTM), which is used as a Language Model conditioned by the image features to generate the captions one word at a time.

Both networks can then be re-trained (or fine-tuned) together by back-propagation for the task of generating sentences. However, in this work we use the pre-trained models provided by Karpathy [2] for both the CNN and the RNN, which have been trained sequentially. is fed by the features extracted by the CNN during the training process).

The CNN used in our experiments is the 16-layer model described in (Simonyan and Zisserman, 2014b), which achieves state-of-the-art result in many image recognition tasks, provided by the authors of the paper, and we used the standard feature extraction procedure.

For the RNN-LSTM part, we have evaluated two models to generate two distinct sets of captions that then could be evaluated using the BAST metric. The architecture is the same in both networks but one is trained using the Flickr-8K (LSTM-RNN-Flickr-8K) train set, dubbed *Model 1* in the rest of the paper, and the other is trained using MicrosoftCOCO (LSTM-RNN-MsCOCO) training set, dubbed *Model 2*. Both networks can be downloaded from the NeuralTalk project web-page. Results for the two models using the existing metrics[3] can be seen in Table 2; notice that our installation reproduces exactly these results (third row).

### 7.2 BAST Metric Results

Figure 5 shows BAST scores for the two caption generation models, we show both results with the manually annotated STs and with the ones automatically predicted by the models. The first observation is that the automatically generated STs are a good proxy for the human evaluation. For all argument combinations, with the exception of locatives (where the differences between the two systems are small) both the BAST computed from automatic and manually annotated STs sort the two systems in the same way. Figure 4 shows some example images and generated captions with the extracted BAST tuples.

Another observation is that overall the numbers are quite low. Despite all the enthusiasm with the latest NN models for sentence generation the F1 of the system for locatives and predicates is quite

[2] We have used the open source project NeuralTalk `https://github.com/karpathy/neuraltalk` which makes it easy to use different pre-trained models for each network.

[3] Evaluation metrics other than BAST have been computed using the tools available at the MsCOCO Challenge website (Lin et al., 2014)

| Dataset test | RNN | CIDEr | Bleu 4 | Bleu 3 | Bleu 2 | Bleu 1 | ROUGE L | METEOR |
|---|---|---|---|---|---|---|---|---|
| MSCOCO* | web ref. | 0.666 | 0.220 | 0.317 | 0.461 | 0.646 | 0.469 | 0.205 |
| MSCOCO* | Model 1 | 0.146 | 0.068 | 0.127 | 0.253 | 0.448 | 0.341 | 0.128 |
| MSCOCO* | Model 2 | 0.666 | 0.220 | 0.317 | 0.461 | 0.646 | 0.469 | 0.205 |
| Flickr-ST | Model 1 | 0.356 | 0.157 | 0.242 | 0.377 | 0.559 | 0.422 | 0.178 |
| Flickr-ST | Model 2 | 0.208 | 0.101 | 0.179 | 0.316 | 0.528 | 0.374 | 0.145 |

Table 2: Results with current metrics for the two models described in the text. MSCOCO* is the subset of MSCOCO used in the NeuralTalk reference experiments. The first row are the results reported in the NeuralTalk project web-site.

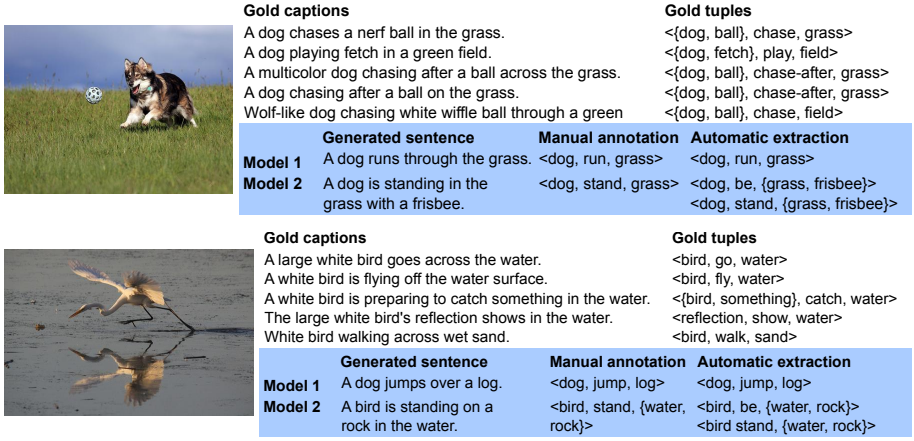


Figure 4: Example results of the two caption generation systems and BAST tuples.

modest, below 25%. Of all the argument types the participants seem to be the easiest to predict for both models, followed by locatives and predicates. This is not surprising since object recognition is probably a more mature research problem in computer vision and state-of-the-art models perform quite well. Overall, however, it seems that caption generation is by no means a solved problem and that there is quite a lot of room for improvement.

## 8 Conclusion

In this paper we have studied the problem of representing the semantics of visually descriptive language. We defined a simple, yet useful, representation and a corresponding evaluation metric. With the proposed metric we can better quantify the agreement between the visual semantics expressed in the gold captions and a generated caption. We show that the metric can be implemented in a fully automatic manner by training models that can accurately predict the semantic representation from sentences. To allow for an objective comparison of caption generation systems we created a new manually annotated dataset of images, captions and underlying visual semantics repre-

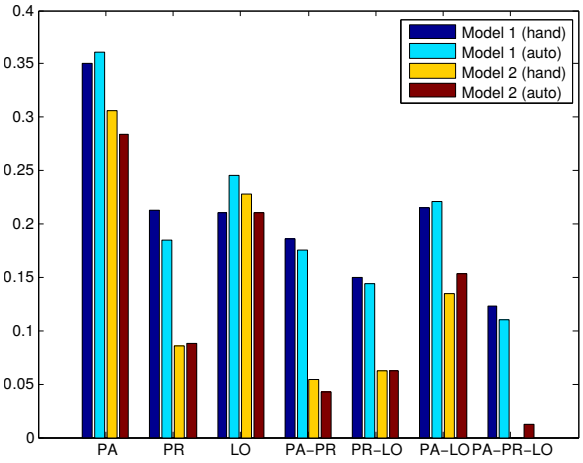


Figure 5: F1 score of the BAST tuples, manually and automatically extracted, from the captions generated by the two evaluated systems for the 50 annotated Flickr-8k validation set images.

sentation by augmenting the widely used Flickr-8K dataset.

Our metric can be used to compare systems but, more importantly, we can use the metric to do a better error analysis. Another nice property of our approach, is that by decoupling the realization of a concept as a lexical item from the underlying visual concept (i.e. the real world entity or event) our annotated corpus can be used to derive different evaluation metrics.

## Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work was partly funded by the Spanish MINECO project RobInstruct TIN2014-58178-R and by the ERA-net CHISTERA project VISEN PCIN-2013-047.

## References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *EACL*, volume 6, pages 249–256.

Xavier Carreras, Isaac Chao, Lluis Padró, and Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. In *LREC*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

David Dowty. 1991. Thematic proto-roles and argument selection. *language*, pages 547–619.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *EMNLP*, pages 1292–1302.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers*, volume 452, page 457.

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 *Computer Vision–ECCV 2010*, pages 15–29. Springer.

Robert Gaizauskas, Josiah Wang, and Arnau Ramisa. 2015. Defining visually descriptive language. In *Proceedings of the 2015 Workshop on Vision and Language (VL'15): Vision and Language Integration Meets Cognitive Systems*.

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*, pages 853–899.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678.

Andrej Karpathy and Li Fei-Fei. 2014. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *LREC*. Citeseer.

Gaurav Kulkarni, Visruth Premraj, Vicente Ordonez, Sudipta Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2891–2903.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

André FT Martins, Miguel Almeida, and Noah A Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *ACL (2)*, pages 617–622. Citeseer.

Stefanie Nowak and Mark J Huiskes. 2010. New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*, volume 1, page 4. Citeseer.

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.

Michael Roth and Kristian Woodsend. 2014. Composition of word representations improves semantic role labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 407–413, Doha, Qatar, October.

Karen Simonyan and Andrew Zisserman. 2014a. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Karen Simonyan and Andrew Zisserman. 2014b. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *arXiv preprint arXiv:1411.5726*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.