Composing Simple Image Descriptions using Web-scale N-grams

Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi
Department of Computer Science
Stony Brook University
NY 11794, USA
{silli, gkulkarni, tlberg, aberg, ychoi}@cs.stonybrook.edu

Abstract

Studying natural language, and especially how people describe the world around them can help us better understand the visual world. In turn, it can also help us in the quest to generate natural language that describes this world in a human manner. We present a simple yet effective approach to automatically compose image descriptions given computer vision based inputs and using web-scale n-grams. Unlike most previous work that summarizes or retrieves pre-existing text relevant to an image, our method *composes* sentences entirely from scratch. Experimental results indicate that it is viable to generate simple textual descriptions that are pertinent to the specific content of an image, while permitting creativity in the description – making for more human-like annotations than previous approaches.

1 Introduction

Gaining a better understanding of natural language, and especially natural language associated with images helps drive research in both computer vision and natural language processing (e.g., Barnard et al. (2003), Pastra et al. (2003), Feng and Lapata (2010b)). In this paper, we look at how to exploit the enormous amount of textual data electronically available today, web-scale n-gram data in particular, in a simple yet highly effective approach to compose image descriptions in natural language. Automatic generation of image descriptions differs from automatic image tagging (e.g., Leong et al. (2010)) in that we aim to generate complex phrases or sentences describing images rather than predicting individual words. These natural language descriptions can be useful for a variety of applications, including image retrieval, automatic video surveillance, and providing image interpretations for visually impaired people.

Our work contrasts to most previous approaches in four key aspects: first, we *compose* fresh sentences from scratch, instead of retrieving (Farhadi et al. (2010)), or summarizing existing text fragments associated with an image (e.g., Aker and Gaizauskas (2010), Feng and Lapata (2010a)). Second, we aim to generate textual descriptions that are truthful to the specific content of the image, whereas related (but subtly different) work in automatic caption generation creates news-worthy text (Feng and Lapata (2010a)) or encyclopedic text (Aker and Gaizauskas (2010)) that is contextually relevant to the image, but not closely pertinent to the specific content of the image. Third, we aim to build a general image description method as compared to work that requires domain specific hand-written grammar rules (Yao et al. (2010)). Last, we allow for some creativity in the generation process which produces more human-like descriptions than a closely related, very recent approach that drives annotation more directly from computer vision inputs (Kulkarni et al., 2011).

In this work, we propose a novel surface realization technique based on web-scale n-gram data. Our approach consists of two steps: (n-gram) phrase selection and (n-gram) phrase fusion. The first step – *phrase selection* – collects candidate phrases that may be potentially useful for generating the description of a given image. This step naturally accommodates uncertainty in image recognition inputs as

Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 220–228,
Portland, Oregon, USA, 23–24 June 2011. ©2011 Association for Computational Linguistics

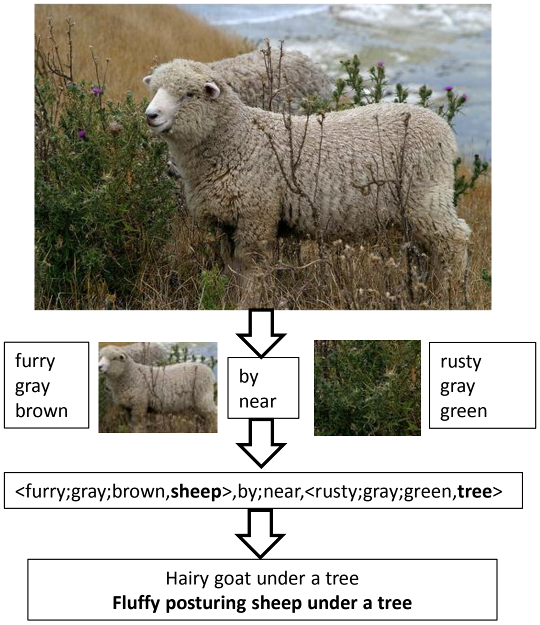

furry
gray
brown
by
near
rusty
gray
green
<furry;gray;brown,**sheep**>,by;near,<rusty;gray;green,**tree**>
Hairy goat under a tree
Fluffy posturing sheep under a tree

Figure 1: The big picture of our task to automatically generate image description.

well as synonymous words and word re-ordering to improve fluency. The second step – *phrase fusion* – finds the optimal compatible set of phrases using dynamic programming to compose a new (and more complex) phrase that describes the image. We compare the performance of our proposed approach to three baselines based on conventional techniques: language models, parsers, and templates.

Despite its simplicity, our approach is highly effective for composing image descriptions: it generates *mostly* appealing and presentable language, while permitting creative writing at times (see Figure 5 for example results). We conclude from our exploration that (1) it is viable to generate simple textual descriptions that are germane to the specific image content, and that (2) world knowledge implicitly encoded in natural language (e.g., web-scale n-gram data) can help enhance image content recognition.

2 Image Recognition

Figure 1 depicts our system flow: a) an image is input into our system, b) image recognition techniques are used to extract visual content information, c) visual content is encoded as a set of triples, d) natural language descriptions are generated.

In this section, we briefly describe the image recognition system that extracts visual information and encodes it as a set of triples. For a given image, the image recognizer extracts *objects*, *attributes* and *spatial relationships* among objects as follows:

1. Objects: including *things* (e.g., bird, bus, car) and *stuff* (e.g., grass, water, sky, road) are detected.
2. Visual attributes (e.g., feathered, black) are predicted for each object.
3. Spatial relationships (e.g., on, near, under) between objects are estimated.

In particular, object detectors are trained using state of the art mixtures of multi-scale deformable parts models (Felzenszwalb et al., 2010). Our set of objects encompasses the 20 PASCAL 2010 object challenge [1] categories as well as 4 additional categories for *flower, laptop, tiger*, and *window* trained on images with associated bounding boxes from Imagenet (Deng et al., 2009). Stuff detectors are trained to detect regions corresponding to non-part based object categories (sky, road, building, tree, water, and grass) using linear SVMs trained on the low level region features of (Farhadi et al., 2009). These are also trained on images with labeled bounding boxes from ImageNet and evaluated at test time on a coarsely sampled grid of overlapping square regions over whole images. Pixels in any region with a classification probability above a fixed threshold are treated as detections.

We select visual attribute characteristics that are relevant to our object and stuff categories. Our attribute terms include 21 visual modifiers – adjectives – related to color (e.g. blue, gray), texture (e.g. striped, furry), material (e.g. wooden, feathered), general appearance (e.g. rusty, dirty, shiny), and shape (e.g. rectangular) characteristics. The attribute classifiers are trained on the low level features of (Farhadi et al., 2009) using RBF kernel SVMs. Preposition functions encoding spatial relationships between objects are hand designed to evaluate the spatial relationships between pairs of regions in an image and provide a score for 16 prepositions (e.g., above, under, against, in etc).

[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/

From these three types of visual output, we construct a meaning representation of an image as a set of triples (one triple for every pair of detected objects). Each triple encodes a spatial relation between two objects in the following format: *<<adj1, obj1>, prep, <adj2, obj2>>*. The generation procedure is elaborated in the following two sections.

3 Baseline Approaches to Surface Realization

This section explores three baseline surface realization approaches: language models (§3.1), randomized local search (§3.2), and template-based (§3.3). Our best approach, *phrase fusion using web-scale n-grams* follows in §4.

3.1 Language Model Based Approach

For each triple, as described in §2, we construct a sentence. For instance, given the triple *<<white, cloud>, in, <blue, sky>>*, we might generate *"There is a white cloud in the blue sky"*.

We begin with a simple decoding scheme based on language models. Let t be a triple, and let V^t be the set of words in t. We perform surface realization by adding function words in-between words in V^t. As a concrete example, suppose we want to determine whether to insert a function word x between a pair of words $\alpha \in V^t$ and $\beta \in V^t$. Then, we need to compare the length-normalized probability $\hat{p}(\alpha x \beta)$ with $\hat{p}(\alpha\beta)$, where $\hat{p}$ takes the n'th root of the probability p for n-word sequences. We insert the new function word x if $\hat{p}(\alpha x \beta) \geq \hat{p}(\alpha\beta)$ using the n-gram models, where the probability of any given sequence $w_1, ..., w_m$ is approximated by

$$p(w_1, ..., w_m) = \prod_{i=1}^{m} p(w_i | w_{i-(n-1)}, ..., w_{i-1})$$

Note that if we wish to reorder words in V^t based on n-gram based language models, then the decoding problem becomes an instance of asymmetric traveler's salesman problem (NP-hard). For brevity, we retain the original order of words in the given triple. We later lift this restriction using the web-scale n-gram based phrase fusion method introduced in §4.

3.2 Randomized Local Search Approach

A much needed extension to the language model based surface realization is incorporating parsers to enforce long distance regularities for more grammatically correct generation. However, optimizing both language-model-based probabilities and parser-based probabilities is intractable. Therefore, we explore a randomized local search approach that makes greedy revisions using both language models and parsers. Randomized local search has been successfully applied to intractable optimization problems in AI (e.g., Chisholm and Tadepalli (2002)) and NLP (e.g., White and Cardie (2002)).

```
Begin Loop (until $T$ iterations or convergence)
    Choose a position $i$ to revise at random
    Choose an edit operation at random
    If the edit yields a better score by LM and PCFG
        Commit the edit
End Loop
```

Table 1: Pseudo code for a randomized local search approach. A possible edit operation includes *insertion, deletion*, and *replacement*. The score of the current sentence is determined by the multiplication LM-based probability and PCFG-based probability.

Table 1 shows the skeleton of the algorithm in our study. Iterating through a loop, it chooses an edit location and an edit operation (insert, delete, or replace) at random. If the edit yields a better score, then we commit the edit, otherwise we jump to the next iteration of the loop. We define the score as

$$score(X) = \hat{p}^{LM}(X)\hat{p}^{PCFG}(X)$$

where X is a given sentence (image description), $\hat{p}^{LM}(X)$ is the length normalized probability of X based on the language model, and $\hat{p}^{PCFG}(X)$ is the length normalized probability of X based on the probabilistic context free grammar (PCFG) model. The loop is repeated until convergence or a fixed number of iterations is reached. Note that this approach can be extended to simulated annealing to allow temporary downward steps to escape from local maxima. We use the PCFG implementation of Klein and Manning (2003).

3.3 Template Based Approach

The third approach is a template-based approach with linguistic constraints, a technique that has often been used for various practical applications such as summarization (Zhou and Hovy, 2004) and dia-

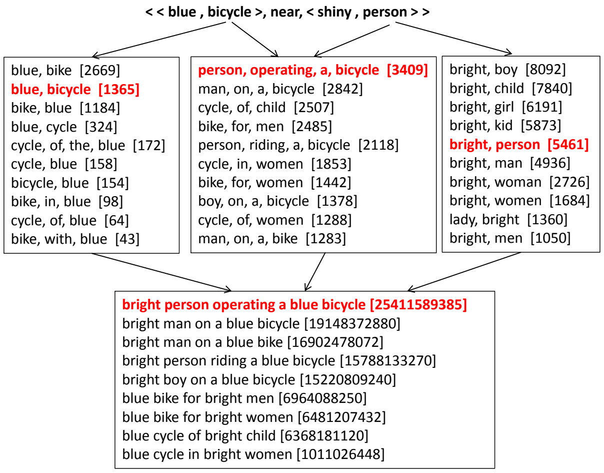

< < blue , bicycle >, near, < shiny , person > >
blue, bike [2669]
blue, bicycle [1365]
bike, blue [1184]
blue, cycle [324]
cycle, of, the, blue [172]
cycle, blue [158]
bicycle, blue [154]
bike, in, blue [98]
cycle, of, blue [64]
bike, with, blue [43]
person, operating, a, bicycle [3409]
man, on, a, bicycle [2842]
cycle, of, child [2507]
bike, for, men [2485]
person, riding, a, bicycle [2118]
cycle, in, women [1853]
bike, for, women [1442]
boy, on, a, bicycle [1378]
cycle, of, women [1288]
man, on, a, bike [1283]
bright, boy [8092]
bright, child [7840]
bright, girl [6191]
bright, kid [5873]
bright, person [5461]
bright, man [4936]
bright, woman [2726]
bright, women [1684]
lady, bright [1360]
bright, men [1050]
bright person operating a blue bicycle [25411589385]
bright man on a blue bicycle [19148372880]
bright man on a blue bike [16902478072]
bright person riding a blue bicycle [15788133270]
bright boy on a blue bicycle [15220809240]
blue bike for bright men [6964088250]
blue bike for bright women [6481207432]
blue cycle of bright child [6368181120]
blue cycle in bright women [1011026448]

Figure 2: Illustration of phrase fusion composition algorithm using web-scale n-grams. Numbers in square brackets are n-gram frequencies.

logue systems (Channarukul et al., 2003). Because the meaning representation produced by the image recognition system has a fixed pattern of *<<adj1, obj1>, prep, <adj2, obj2>>*, it can be templated as *"There is a [adj1] [obj1] [prep] the [adj2] [obj2]."* We also include templates that encode basic discourse constraints. For instance, the template that generated the first sentences in Figure 3 and 4 is: [PREFIX] [#(x_1)] [x_1], [#(x_2)] [x_2], ... and [#(x_k)] [x_k], where x_i is the name of an object (e.g. "cow"), #(x_i) is the number of instances of x_i (e.g. "one"), and PREFIX $\in$ {"This picture shows", "This is a picture of", etc}.

Although this approach can produce good looking sentences in a limited domain, there are many limitations. First, a template-based approach does not allow creative writing and produces somewhat stilted prose. In particular, it cannot add interesting new words, or replace existing content words with better ones. In addition, such an approach does not allow any reordering of words which might be necessary to create a fluent sentence. Finally, hand-written rules are domain-specific, and do not generalize well to new domains.

4 Surface Realization by Phrase Fusion using Web-scale N-gram

We now introduce an entirely different approach that addresses the limitations of the conventional approaches discussed in §3. This approach is based on web-scale n-gram, also known as Google Web 1T data, which provides the frequency count of each possible n-gram sequence for $1 \leq n \leq 5$.

4.1 [Step I] – Candidate Phrase Selection

We first define three different sets of phrases for each given triple *<<adj1, obj1>, prep, <adj2, obj2>>*:

- $\mathcal{O}_1 = \{(x, f) \mid x$ is an n-gram phrase describing the **first** object using the words $adj1$ and $obj1$, and f is the frequency of $x\}$
- $\mathcal{O}_2 = \{(x, f) \mid x$ is an n-gram phrase describing the **second** object using the words $adj2$ and $obj2$, and f is the frequency of $x\}$
- $\mathcal{R} = \{(x, f) \mid x$ is an n-gram describing the **relation** between the two objects using the words $obj1$ and $obj2$, and f is the frequency of $x\}$

We find n-gram phrases for $\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{R}$ from the Google Web 1T data. The search patterns for $\mathcal{O}_1$ is:

- $[adj_1]$ $[\clubsuit]^{n-2}$ $[obj_1]$
- $[obj_1]$ $[\clubsuit]^{n-2}$ $[adj_1]$

where $[\clubsuit]$ is a wildcard word, and $[\clubsuit]^{n-2}$ denotes a sequence of n-2 number of wildcard words in a n-gram sequence. For wildcards, we only allow a limited set of function words, and verbs in the gerund form[2] for reasons that will become clearer in the next step – phrase fusion in §4.2.

Note that it is the second pattern that allows interesting re-ordering of words in the final sentence generation. For instance, suppose $adj1$=green, $obj1$=person. Then it is more natural to generate a phrase using the reverse pattern such as, *"person in green"* or *"person wearing green"* than simply concatenating $adj1$ and $obj1$ to generate *"green person"*. Similarly, given $obj1$=bicycle and $obj2$=man, generating a phrase using the reverse pattern, e.g., *"man with a bicycle"* would be more natural than *"bicycle with a man"*. Our hypothesis is that such ordering preference is implicitly encoded in the web-scale n-grams via frequencies.

It is worthwhile to note that our pattern matching is case sensitive, and we only allow patterns that are

[2] We treat words with suffix "ing" as gerund for simplicity.

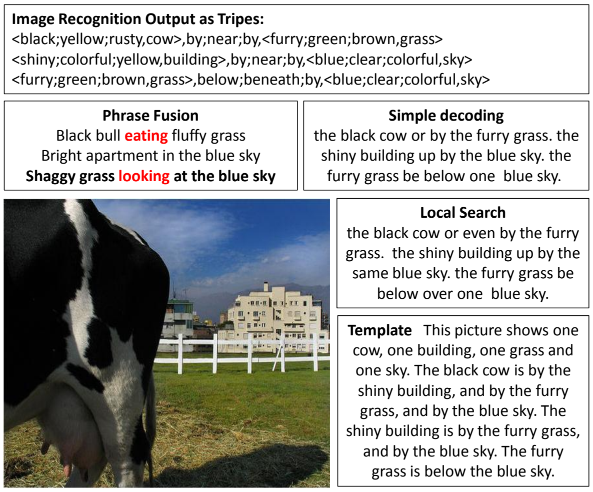


Image Recognition Output as Tripes:
<black;yellow;rusty,cow>,by;near;by,<furry;green;brown,grass>
<shiny;colorful;yellow,building>,by;near;by,<blue;clear;colorful,sky>
<furry;green;brown,grass>,below;beneath;by,<blue;clear;colorful,sky>

Phrase Fusion
Black bull eating fluffy grass
Bright apartment in the blue sky
Shaggy grass looking at the blue sky

Simple decoding
the black cow or by the furry grass. the shiny building up by the blue sky. the furry grass be below one blue sky.

Local Search
the black cow or even by the furry grass. the shiny building up by the same blue sky. the furry grass be below over one blue sky.

Template This picture shows one cow, one building, one grass and one sky. The black cow is by the shiny building, and by the furry grass, and by the blue sky. The shiny building is by the furry grass, and by the blue sky. The furry grass is below the blue sky.

Figure 3: Comparison of image descriptions

all lower-case. From our pilot study, we found that n-grams with upper case characters are likely from named entities, which distort the n-gram frequency distribution that we rely on during the phrase fusion phase. To further reduce noise, we also discard any n-gram that contains a character that is not an alphabet.

Accommodating Uncertainty We extend candidate phrase selection in order to cope with uncertainty from the image recognition. In particular, for each object detection obj_i, we include its top 3 predicted modifiers adj_{i1}, adj_{i2}, adj_{i3} determined by the attribute classifiers (see §2) to expand the set $\mathcal{O}_1$ and $\mathcal{O}_2$ accordingly. For instance, given adj_i =(shiny or white) and obj_i = sheep, we can consider both <shiny,sheep> and <white,sheep> pairs to predict more compatible pairs of words.

Accommodating Synonyms Additionally, we augment each modifier adj_i and each object name obj_i with synonyms to further expand our sets $\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{R}$. These expanded sets of phrases enable resulting generations that are more fluent and creative.

4.2 [Step II] – Phrase Fusion

Given the expanded sets of phrases $\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{R}$ described above, we perform phrase fusion to generate simple image description. In this step, we find the best combination of three phrases, $(\hat{x_1}, \hat{f_1}) \in \mathcal{O}_1$, $(\hat{x_2}, \hat{f_2}) \in \mathcal{O}_2$, and $(\hat{x_R}, \hat{f_R}) \in \mathcal{R}$ as follows:

$$(\hat{x_1}, \hat{x_2}, \hat{x_R}) = \text{argmax}_{x_1,x_2,x_R} score(x_1, x_2, x_R) \quad (1)$$

$$score(x_1, x_2, x_R) = \phi(x_1) \times \phi(x_2) \times \phi(x_R) \quad (2)$$

$$s.t.\ \hat{x_1}\ and\ \hat{x_R} \text{ are compatible}$$

$$\&\quad \hat{x_2}\ and\ \hat{x_R} \text{ are compatible}$$

Two phrases $\hat{x_i}$ and $\hat{x_R}$ are compatible if they share the same object noun obj_i. We define the phrase-level score function $\phi(\cdot)$ as $\phi(x_i) = f_i$ using the Google n-gram frequencies. The equation (2) can be maximized using dynamic programming, by aligning the decision sequence as $\hat{x_1} - \hat{x_R} - \hat{x_2}$.

Once the best combination – $(\hat{x_1}, \hat{x_2}, \hat{x_R})$ is determined, we fuse the phrases by replacing the word obj_1 in the phrase $\hat{x_R}$ with the corresponding phrase $\hat{x_1}$. Similarly, we replace the word obj_2 in the phrase $\hat{x_R}$ with the other corresponding phrase $\hat{x_2}$. Because the wildcard words – [♣] in §4.1 allow only a limited set of function words and gerund, the resulting phrase is highly likely to be grammatically correct.

Computational Efficiency One advantage of our phrase fusion method is its efficiency. If we were to attempt to re-order words with language models in a naive way, we would need to consider all possible permutations of words – an NP-hard problem (§3.1). However, our phrase fusion method is clever in that it probes reordering only on selected pairs of words, where reordering is likely to be useful. In other words, our approach naturally ignores most word pairs that do not require reordering and has a time complexity of only $O(K^2n)$, where K is the maximum number of candidate phrases of any phrase type, and n is the number of phrase types in each sentence. K can be kept as a small constant by selecting K-best candidate phrases of each phrase type. We set $K = 10$ in this paper.

5 Experimental Results

To construct the training corpus for language models, we crawled Wikipedia pages that describe our object set. For evaluation, we use the UIUC PASCAL sentence dataset[3] which contains upto five human-generated sentences that describing 1000 images. Note that all of the approaches presented in

[3] http://vision.cs.uiuc.edu/pascal-sentences/

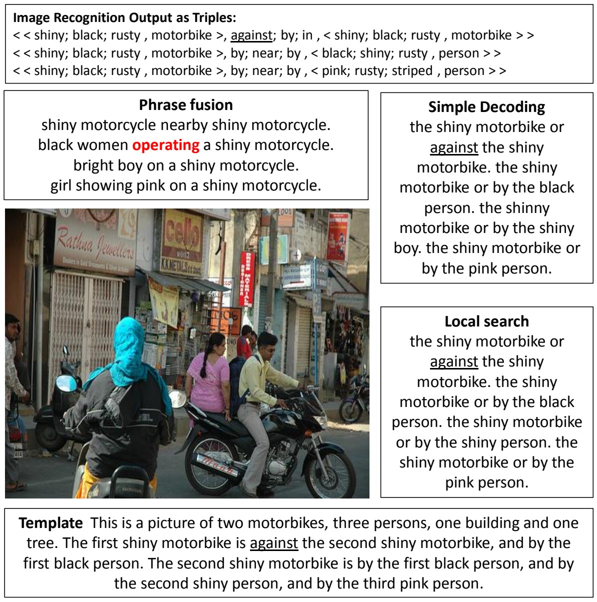


Image Recognition Output as Triples:
< < shiny; black; rusty , motorbike >, against; by; in , < shiny; black; rusty , motorbike > >
< < shiny; black; rusty , motorbike >, by; near; by , < black; shiny; rusty , person > >
< < shiny; black; rusty , motorbike >, by; near; by , < pink; rusty; striped , person > >

Phrase fusion
shiny motorcycle nearby shiny motorcycle.
black women **operating** a shiny motorcycle.
bright boy on a shiny motorcycle.
girl showing pink on a shiny motorcycle.

Simple Decoding
the shiny motorbike or against the shiny motorbike. the shiny motorbike or by the black person. the shinny motorbike or by the shiny boy. the shiny motorbike or by the pink person.

Local search
the shiny motorbike or against the shiny motorbike. the shiny motorbike or by the black person. the shiny motorbike or by the shiny person. the shiny motorbike or by the pink person.

Template This is a picture of two motorbikes, three persons, one building and one tree. The first shiny motorbike is against the second shiny motorbike, and by the first black person. The second shiny motorbike is by the first black person, and by the second shiny person, and by the third pink person.

Figure 4: Comparison of image descriptions

	w/o	w/ syn
LANGUAGE MODEL	0.094	0.106
TEMPLATE	0.087	0.096
LOCAL SEARCH	0.100	0.111
PHRASE FUSION (any best)	0.149	0.153
PHRASE FUSION (best w/ gerund)	0.146	0.149
Human	0.500	0.510

Table 2: Automatic Evaluation: BLEU measured at 1

	Creativ.	Fluency	Relevan.
LANGUAGE MODEL	2.12	1.96	2.09
TEMPLATE	2.04	1.7	1.96
LOCAL SEARCH	2.21	1.96	2.04
PHRASE FUSION	1.86	1.97	2.11

Table 3: Human Evaluation: the scores range over 1 to 3, where 1 is very good, 2 is ok, 3 is bad.

Section 3 and 4 attempt to insert function words for surface realization. In this work, we limit the choice of function words to only those words that are likely to be necessary in the final output.[4] For instance, we disallow function words such as "who" or "or".

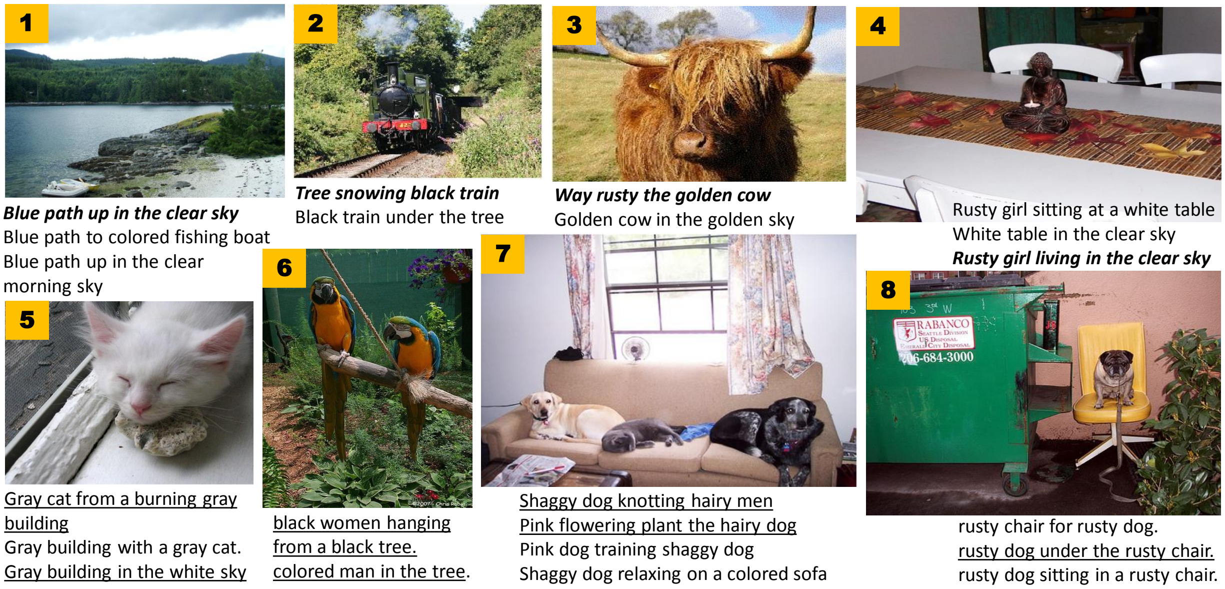
Before presenting evaluation results, we present some samples of image descriptions generated by 4 different approaches in Figure 3 and 4. Notice that only the PHRASE FUSION approach is able to include interesting and adequate verbs, such as *"eating"* or *"looking"* in Figure 3, and *"operating"* in Figure 4. Note that the choice of these action verbs is based only on the co-occurrence statistics encoded in n-grams, without relying on the vision component that specializes in action recognition. These examples therefore demonstrate that world knowledge implicitly encoded in natural language can help enhance image content recognition.

Automatic Evaluation: BLEU (Papineni et al., 2002) is a widely used metric for automatic evaluation of machine translation that measures the n-gram precision of machine generated sentences with respect to human generated sentences. Because our task can be viewed as machine translation from images to text, BLEU (Papineni et al., 2002) may seem like a reasonable choice. However, there is larger inherent variability in generating sentences from images than translating a sentence from one language to another. In fact two people viewing the same picture may produce quite different descriptions. This means BLEU could penalize many correctly generated sentences, and be poorly correlated with human judgment of quality. Nevertheless we report BLEU scores in absence of any other automatic evaluation method that serves our needs perfectly.

The results are shown in Table 2 – first column shows BLEU score considering exact matches, second column shows BLEU with full credit for synonyms. To give a sense of upper bound and to see some limitations of the BLEU score, we also compute the BLEU score between human-generated sentences by computing the BLEU score of the first human sentence with respect to the others.

There is one important factor to consider when interpreting Table 2. The four approaches explored in this paper are purposefully prolific writers in that they generate many more sentences than the number of sentences in the image descriptions written by humans (available in the UIUC PASCAL dataset). In this work, we do not perform sentence selection to reduce the number of sentences in the final output. Rather, we focus on the quality of each generated sentence. The consequence of producing many

[4] This limitation does not apply to TEMPLATE.



Blue path up in the clear sky
Blue path to colored fishing boat
Blue path up in the clear morning sky

Tree snowing black train
Black train under the tree

Way rusty the golden cow
Golden cow in the golden sky

Rusty girl sitting at a white table
White table in the clear sky
Rusty girl living in the clear sky

Gray cat from a burning gray building
Gray building with a gray cat.
Gray building in the white sky

black women hanging from a black tree.
colored man in the tree.

Shaggy dog knotting hairy men
Pink flowering plant the hairy dog
Pink dog training shaggy dog
Shaggy dog relaxing on a colored sofa

rusty chair for rusty dog.
rusty dog under the rusty chair.
rusty dog sitting in a rusty chair.

Figure 5: Sample image descriptions using PHRASE FUSION: some of the unexpected or poetic descriptions are highlighted in **boldface**, and some of the interesting incorrect descriptions are underlined.

more sentences in our output is overall lower BLEU scores, because BLEU precision penalizes spurious repetitions of the same word, which necessarily occurs when generating more sentences. This is not an issue for comparing different approaches however, as we generate the same number of sentences for each method.

From Table 2, we find that our final approach — PHRASE FUSION based on web-scale n-grams performs the best. Notice that there are two different evaluations for PHRASE FUSION: the first one is evaluated for the best combination of phrases (Equation (1)), while the second one is evaluated for the best combination of phrases that contained at least one gerund.

Human Evaluation: As mentioned earlier, BLEU score has some drawbacks including obliviousness to correctness of grammar and inability to evaluate the creativity of a composition. To directly quantify these aspects that could not be addressed by BLEU, we perform human judgments on 120 instances for the four proposed methods. Evaluators do not have any computer vision or natural language generation background.

We consider the following three aspects to evaluate the our image descriptions: *creativity*, *fluency*, and *relevance*. For simplicity, human evaluators assign one set of scores for each aspect per image. The scores range from 1 to 3, where 1 is very good, 2 is ok, and 3 is bad.[5] The definition and guideline for each aspect is:

[Creativity] How creative is the generated sentence?

1 There is creativity either based on unexpected words (in particular, verbs), or describing things in a poetic way.

2 There is minor creativity based on re-ordering words that appeared in the triple

3 None. Looks like a robot talking.

[Fluency] How grammatically correct is the generated sentence?

1 Mostly perfect English phrase or sentence.

2 There are some errors, but mostly comprehensible.

3 Terrible.

[Relevance] How relevant is the generated description to the given image?

1 Very relevant.

2 Reasonably relevant.

3 Totally off.

[5] In our pilot study, human annotations on 160 instances given by two evaluators were identical on 61% of the instances, and close (difference ≤ 1) on 92%.

Table 3 shows the human evaluation results. In terms of *creativity*, PHRASE FUSION achieves the best score as expected. In terms of *fluency* and *relevance* however, TEMPLATE achieves the best scores, while PHRASE FUSION performs the second best. Remember that TEMPLATE is based on hand-engineered rules with discourse constraints, which seems to appeal to evaluators more. It would be straightforward to combine PHRASE FUSION with TEMPLATE to improve the output of PHRASE FUSION with hand-engineered rules. However, our goal in this paper is to investigate statistically motivated approaches for generating image descriptions that can address inherent limitations of hand-written rules discussed in §3.3.

Notice that the relevance score of TEMPLATE is better than that of LANGUAGE MODEL, even though both approaches generate descriptions that consist of an almost identical set of words. This is presumably because the output from LANGUAGE MODEL contains grammatically incorrect sentences that are not comprehendable enough to the evaluators. The relevance score of PHRASE FUSION is also slightly worse than that of TEMPLATE, presumably because PHRASE FUSION often generates poetic or creative expressions, as shown in Figure 5, which can be considered a deviation from the image content.

Error Analysis There are different sources of errors. Some errors are due to mistakes in the original visual recognition input. For example, in the 3rd image in Figure 5, the color of sky is predicted to be "golden". In the 4th image, the wall behind the table is recognized as "sky", and in the 6th image, the parrots are recognized as "person".

Other errors are from surface realization. For instance, in the 8th image, PHRASE FUSION selects the preposition "under", presumably because dogs are typically under the chair rather than on the chair according to Google n-gram statistics. In the 5th image, an unexpected word "burning" is selected to make the resulting output idiosyncratic. Word sense disambiguation sometimes causes a problem in surface realization as well. In the 3rd image, the word "way" is chosen to represent "path" or "street" by the image recognizer. However, a different sense of way – "very" – is being used in the final output.

6 Related Work

There has been relatively limited work on automatically generating natural language image descriptions. Most work related to our study is discussed in §1, hence we highlight only those that are closest to our work here. Yao et al. (2010) present a comprehensive system that generates image descriptions using Head-driven phrase structure (HPSG) grammar, which requires carefully written domain-specific lexicalized grammar rules, and also demands a very specific and complex meaning representation scheme from the image processing. In contrast, our approach handles images in the open-domain more naturally using much simpler techniques.

We use similar vision based inputs – object detectors, modifier classifiers, and prepositional functions – to some very recent work on generating simple descriptions for images (Kulkarni et al., 2011), but focus on improving the sentence generation methodology and produce descriptions that are more true to human generated descriptions. Note that the BLEU scores reported in their work of Kulkarni et al. (2011) are not directly comparable to ours, as the scale of the scores differs depending on the number of sentences generated per image.

7 Conclusion

In this paper, we presented a novel surface realization technique based on web-scale n-gram data to automatically generate image description. Despite its simplicity, our method is highly effective in generating mostly appealing and presentable language, while permitting creative writing at times. We conclude from our study that it is viable to generate simple textual descriptions that are germane to the specific image content while also sometimes producing almost poetic natural language. Furthermore, we demonstrate that world knowledge implicitly encoded in natural language can help enhance image content recognition.

Acknowledgments

This work is supported in part by NSF Faculty Early Career Development (CAREER) Award #1054133.

References

A. Aker and R. Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In *ACL*.

K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. 2003. Matching words and pictures. *JMLR*, 3:1107–1135.

Songsak Channarukul, Susan W. McRoy, and Syed S. Ali. 2003. Doghed: a template-based generator for multimodal dialog systems targeting heterogeneous devices. In *NAACL*.

Michael Chisholm and Prasad Tadepalli. 2002. Learning decision rules by randomized iterative local search. In *ICML*, pages 75–82.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. 2009. Describing objects by their attributes. In *CVPR*.

A. Farhadi, M Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth. 2010. Every picture tells a story: generating sentences for images. In *ECCV*.

P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. 2010. Object detection with discriminatively trained part based models. *tPAMI*, Sept.

Y. Feng and M. Lapata. 2010a. How many words is a picture worth? automatic caption generation for news images. In *ACL*.

Yansong Feng and Mirella Lapata. 2010b. Topic models for image annotation and text illustration. In *HLT*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430. Association for Computational Linguistics.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Babytalk: Understanding and generating simple image descriptions. In *CVPR*.

Chee Wee Leong, Rada Mihalcea, and Samer Hassan. 2010. Text mining for automatic image tagging. In *COLING*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.

Katerina Pastra, Horacio Saggion, and Yorick Wilks. 2003. Nlp for indexing and retrieval of captioned photographs. In *EACL*.

Michael White and Claire Cardie. 2002. Selecting sentences for multidocument summaries using randomized local search. In *ACL Workshop on Automatic Summarization*.

B.Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. 2010. I2t: Image parsing to text description. *Proc. IEEE*, 98(8).

Liang Zhou and Eduard Hovy. 2004. Template-filtered headline summarization. In *Text Summarization Branches Out: Pr ACL-04 Wkshp*, July.