Collective Generation of Natural Image Descriptions

Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg,
Tamara L. Berg and Yejin Choi
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
{pkuznetsova,vordonezroma,aberg,tlberg,ychoi}@cs.stonybrook.edu

Abstract

We present a holistic data-driven approach to image description generation, exploiting the vast amount of (noisy) parallel image data and associated natural language descriptions available on the web. More specifically, given a query image, we retrieve existing human-composed phrases used to describe visually similar images, then selectively combine those phrases to generate a novel description for the query image. We cast the generation process as constraint optimization problems, collectively incorporating multiple interconnected aspects of language composition for *content planning*, *surface realization* and *discourse structure*. Evaluation by human annotators indicates that our final system generates more semantically correct and linguistically appealing descriptions than two nontrivial baselines.

1 Introduction

Automatically describing images in natural language is an intriguing, but complex AI task, requiring accurate computational visual recognition, comprehensive world knowledge, and natural language generation. Some past research has simplified the general image description goal by assuming that relevant text for an image is provided (e.g., Aker and Gaizauskas (2010), Feng and Lapata (2010)). This allows descriptions to be generated using effective summarization techniques with relatively surface level image understanding. However, such text (e.g., news articles or encyclopedic text) is often only loosely related to an image's specific content and many natural images do not come with associated text for summarization.

In contrast, other recent work has focused more on the visual recognition aspect by detecting content elements (e.g., scenes, objects, attributes, actions, etc) and then composing descriptions from scratch (e.g., Yao et al. (2010), Kulkarni et al. (2011), Yang et al. (2011), Li et al. (2011)), or by retrieving existing whole descriptions from visually similar images (e.g., Farhadi et al. (2010), Ordonez et al. (2011)). For the latter approaches, it is unrealistic to expect that there will always exist a *single complete description* for retrieval that is pertinent to a given query image. For the former approaches, visual recognition first generates an intermediate representation of image content using a set of English words, then language generation constructs a full description by adding function words and optionally applying simple re-ordering. Because the generation process sticks relatively closely to the recognized content, the resulting descriptions often lack the kind of coverage, creativity, and complexity typically found in human-written text.

In this paper, we propose a holistic data-driven approach that combines and extends the best aspects of these previous approaches – a) using visual recognition to directly predict individual image content elements, and b) using retrieval from existing human-composed descriptions to generate natural, creative, and inter-

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 359–368,
Jeju, Republic of Korea, 8-14 July 2012.

esting captions. We also lift the restriction of retrieving existing *whole descriptions* by gathering *visually relevant phrases* which we combine to produce novel and query-image specific descriptions. By judiciously exploiting the correspondence between image content elements and phrases, it is possible to generate natural language descriptions that are substantially richer in content and more linguistically interesting than previous work.

At a high level, our approach can be motivated by linguistic theories about the connection between reading activities and writing skills, i.e., substantial reading enriches writing skills, (e.g., Hafiz and Tudor (1989), Tsang (1996)). Analogously, our generation algorithm attains a higher level of linguistic sophistication by *reading* large amounts of descriptive text available online. Our approach is also motivated by language grounding by visual worlds (e.g., Roy (2002), Dindo and Zambuto (2010), Monner and Reggia (2011)), as in our approach the meaning of a phrase in a description is implicitly grounded by the relevant content of the image.

Another important thrust of this work is collective *image-level content-planning*, integrating saliency, content relations, and discourse structure based on statistics drawn from a large image-text parallel corpus. This contrasts with previous approaches that generate multiple sentences without considering discourse flow or redundancy (e.g., Li et al. (2011)). For example, for an image showing a flock of birds, generating a large number of sentences stating the relative position of each bird is probably not useful.

Content planning and phrase synthesis can be naturally viewed as constraint optimization problems. We employ Integer Linear Programming (ILP) as an optimization framework that has been used successfully in other generation tasks (e.g., Clarke and Lapata (2006), Martins and Smith (2009), Woodsend and Lapata (2010)). Our ILP formulation encodes a rich set of linguistically motivated constraints and weights that incorporate multiple aspects of the generation process. Empirical results demonstrate that our final system generates linguistically more appealing and semantically more correct descriptions than two nontrivial baselines.

1.1 System Overview

Our system consists of two parts. For a query image, we first retrieve candidate descriptive phrases from a large image-caption database using measures of visual similarity (§2). We then generate a coherent description from these candidates using ILP formulations for content planning (§4) and surface realization (§5).

2 Vision & Phrase Retrieval

For a query image, we retrieve relevant candidate natural language phrases by visually comparing the query image to database images from the SBU Captioned Photo Collection (Ordonez et al., 2011) (1 million photographs with associated human-composed descriptions). Visual similarity for several kinds of image content are used to compare the query image to images from the database, including: 1) object detections for 89 common object categories (Felzenszwalb et al., 2010), 2) scene classifications for 26 common scene categories (Xiao et al., 2010), and 3) region based detections for *stuff* categories (e.g. grass, road, sky) (Ordonez et al., 2011). All content types are pre-computed on the million database photos, and caption parsing is performed using the Berkeley PCFG parser (Petrov et al., 2006; Petrov and Klein, 2007).

Given a query image, we identify content elements present using the above classifiers and detectors and then retrieve phrases referring to those content elements from the database. For example, if we detect a horse in a query image, then we retrieve phrases referring to visually similar horses in the database by comparing the color, texture (Leung and Malik, 1999), or shape (Dalal and Triggs, 2005; Lowe, 2004) of the detected horse to detected horses in the database images. We collect four types of phrases for each query image as follows:

[1] NPs We retrieve noun phrases for each query object detection (e.g., "the brown cow") from database captions using visual similarity between object detections computed as an equally weighted linear combination of L_2 dis-

tances on histograms of *color*, *texton* (Leung and Malik, 1999), *HoG* (Dalal and Triggs, 2005) and *SIFT* (Lowe, 2004) features.

[2] VPs We retrieve verb phrases for each query object detection (e.g. "boy running") from database captions using the same measure of visual similarity as for NPs, but restricting the search to only those database instances whose captions contain a verb phrase referring to the object category.

[3] Region/Stuff PPs We collect prepositional phrases for each query stuff detection (e.g. "in the sky", "on the road") by measuring visual similarity of appearance (color, texton, HoG) and geometric configuration (object-stuff relative location and distance) between query and database detections.

[4] Scene PPs We also collect prepositonal phrases referring to general image scene context (e.g. "at the market", "on hot summer days", "in Sweden") based on global scene similarity computed using L_2 distance between scene classification score vectors (Xiao et al., 2010) computed on the query and database images.

3 Overview of ILP Formulation

For each image, we aim to generate multiple sentences, each sentence corresponding to a single distinct object detected in the given image. Each sentence comprises of the NP for the main object, and a *subset* of the corresponding VP, region/stuff PP, and scene PP retrieved in §2. We consider four different types of operations to generate the final description for each image:

T1. Selecting the set of objects to describe (one object per sentence).
T2. Re-ordering sentences (i.e., re-ordering objects).
T3. Selecting the set of phrases for each sentence.
T4. Re-ordering phrases within each sentence.

The ILP formulation of §4 addresses T1 & T2, i.e., content-planning, and the ILP of §5 addresses T3 & T4, i.e., surface realization.[1]

[1]It is possible to create one conjoined ILP formulation to address all four operations T1—T4 at once. For computational and implementation efficiency however, we opt for the two-step approach.

4 Image-level Content Planning

First we describe image-level content planning, i.e., *abstract generation*. The goals are to (1) select a subset of the objects based on saliency and semantically compatibility, and (2) order the selected objects based on their content relations.

4.1 Variables and Objective Function

The following set of indicator variables encodes the selection of objects and ordering:

$$y_{sk} = \begin{cases} 1, & \text{if object } s \text{ is selected} \\ & \text{for position } k \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $k = 1, ..., S$ encodes the position (order) of the selected objects, and s indexes one of the objects. In addition, we define a set of variables indicating specific pairs of adjacent objects:

$$y_{skt(k+1)} = \begin{cases} 1, & \text{if } y_{sk} = y_{t(k+1)} = 1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The objective function, F, that we will maximize is a weighted linear combination of these indicator variables and can be optimized using integer linear programming:

$$F = \sum_s F_s \cdot \sum_{k=1}^{S} y_{sk} - \sum_{st} F_{st} \cdot \sum_{k=1}^{S-1} y_{skt(k+1)} \quad (3)$$

where F_s quantifies the salience/confidence of the object s, and F_{st} quantifies the semantic compatibility between the objects s and t. These coefficients (weights) will be described in §4.3 and §4.4. We use IBM CPLEX to optimize this objective function subject to the constraints introduced next in §4.2.

4.2 Constraints

Consistency Constraints: We enforce consistency between indicator variables for indivisual objects (Eq. 1) and consecutive objects (Eq. 2) so that $y_{skt(k+1)} = 1$ *iff* $y_{sk} = 1$ and $y_{t(k+1)} = 1$:

$$\forall_{stk},\ y_{skt(k+1)} \leq y_{sk} \quad (4)$$

$$y_{skt(k+1)} \leq y_{t(k+1)} \quad (5)$$

$$y_{skt(k+1)} + (1 - y_{sk}) + (1 - y_{t(k+1)}) \geq 1 \quad (6)$$

To avoid empty descriptions, we enforce that the result includes at least one object:

$$\sum_s y_{s1} = 1 \quad (7)$$

To enforce contiguous positions be selected:

$$\forall k = 2, ..., S-1, \quad \sum_s y_{s(k+1)} \leq \sum_s y_{sk} \quad (8)$$

Discourse constraints: To avoid spurious descriptions, we allow at most two objects of the same type, where c_s is the type of object s:

$$\forall c \in objTypes, \quad \sum_{\{s:\ c_s=c\}} \sum_{k=1}^{S} y_{sk} \leq 2 \quad (9)$$

4.3 Weight F_s: Object Detection Confidence

In order to quantify the confidence of the object detector for the object s, we define $0 \leq F_s \leq 1$ as the mean of the detector scores for that object type in the image.

4.4 Weight F_{st}: Ordering and Compatibility

The weight $0 \leq F_{st} \leq 1$ quantifies the compatibility of the object pairing (s, t). Note that in the objective function, we subtract this quantity from the function to be maximized. This way, we create a competing tension between the single object selection scores and the pairwise compatibility scores, so that variable number of objects can be selected.

Object Ordering Statistics: People have biases on the order of topic or content flow. We measure these biases by collecting statistics on ordering of object names from the 1 million image descriptions in the SBU Captioned Dataset (Ordonez et al., 2011). Let $f_{\text{ord}}(w_1, w_2)$ be the number of times w_1 appeared before w_2. For instance, $f_{\text{ord}}(window, house) = 2895$ and $f_{\text{ord}}(house, window) = 1250$, suggesting that people are more likely to mention a window before mentioning a house/building[2]. We use these ordering statistics to enhance content flow. We define score for the order of objects using Z-score for normalization as follows:

$$\hat{F}_{st} = \frac{f_{\text{ord}}(c_s, c_t) - mean(f_{\text{ord}})}{std\_dev(f_{\text{ord}})} \quad (10)$$

[2]We take into account synonyms.

We then transform $\hat{F}_{st}$ so that $\hat{F}_{st} \in [0,1]$, and then set $F_{st} = 1 - \hat{F}_{st}$ so that smaller values correspond to better choices.

5 Surface Realization

Recall that for each image, the computer vision system identifies phrases from descriptions of images that are similar in a variety of aspects. The result is a set of phrases representing four different types of information (§2). From this assortment of phrases, we aim to select a subset and glue them together to compose a complete sentence that is linguistically plausible and semantically truthful to the content of the image.

5.1 Variables and Objective Function

The following set of variables encodes the selection of phrases and their ordering in constructing S' sentences.

$$x_{sijk} = \begin{cases} 1, & \text{if phrase } i \text{ of type } j \\ & \text{is selected} \\ & \text{for position } k \\ & \text{in sentence } s \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

where $k = 1, ..., N$ encodes the ordering of the selected phrases, and j indexes one of the four phrases types (object-NPs, action-VPs, region-PPs, scene-PPs), $i = 1, ..., M$ indexes one of the M candidate phrases of each phrase type, and $s = 1, ..., S'$ encodes the sentence (object). In addition, we define indicator variables for adjacent pairs of phrases: $x_{sijkpq(k+1)} = 1$ if $x_{sijk} = x_{spq(k+1)} = 1$ and 0 otherwise. Finally, we define the objective function F as:

$$F = \sum_{sij} F_{sij} \cdot \sum_{k=1}^{N} x_{sijk} - \sum_{sijpq} F_{sijpq} \cdot \sum_{k=1}^{N-1} x_{sijkpq(k+1)} \quad (12)$$

where F_{sij} weights individual phrase goodness and F_{sijpq} adjacent phrase goodness. All coefficients (weights) will be described in Section 5.3 and 5.4.

We optionally prepend the first sentence in a generated description with a *cognitive phrase*.[3]

[3]We collect most frequent 200 phrases of length 1-7 that start a caption from the SBU Captioned Photo Collection.

ILP: Found ***trucks*** parked on first avenue in the east village.

HMM: This is the first ***cellar door*** left back bedroom in center and clothes dryer to the right to the building in the house. This ***HUGE screen*** hanging on the wall outside a burned down building in the house. ***My truck*** parked on first avenue in the east village by the glass buildings in the house.

Human: Flat bed Chisholms truck on display at the vintage vehicle rall y at Astley Green Colliery near Leigh Lancs

ILP: This is a photo of this little flower sprouted up in defiance against grass. Bright yellow flowers growing in a rock garden at Volcan Mombacho.

HMM: These was taken on the flowers growing in a rock garden in the field in two sorts. This little flower sprouted up in defiance in the field in two sorts. A full open flower sprouted up in defiance in the field in gardens. Bright yellow flowers growing in a rock garden in the field.

Human: Yellow flower in my field

ILP: I think this is a boy's bike lied in saltwater for quite a while.

HMM: I liked the way ***bicycles*** leaning against a wall in Copenhagen Denmark in a windy sky in a Singapore bathroom. ***Boy's bike*** lied in saltwater for quite a while in a windy sky in a Singapore bathroom. ***Fruit*** rubbing his face in the encrusted snow in a windy sky in a Singapore bathroom.

Human: You re nobody in Oxford, unless you have a old bike with a basket

Figure 1: ILP & HMM generated captions. In HMM generated captions, underlined phrases show redundancy across different objects (due to lack of discourse constraints), and phrases in **boldface** show awkward topic flow (due to lack of content planning). Note that in the bicycle image, the visual recognizer detected two separate bicycles and some fruits, as can be seen in the HMM result. Via collective image-level content planning (see §4), some of these erroneous detection can be corrected, as shown in the ILP result. Spurious and redundant phrases can be suppressed via discourse constraints (see §5).

These are generic constructs that are often used to start a description about an image, for instance, "This is an image of...". We treat these phrases as an additional type, but omit corresponding variables and constraints for brevity.

5.2 Constraints

Consistency Constraints: First we enforce consistency between the unary variables (Eq. 11) and the pairwise variables so that $x_{sijkpqm} = 1$ *iff* $x_{sijk} = 1$ and $x_{spqm} = 1$:

$$\forall_{ijkpqm},\ x_{sijkpqm} \leq x_{sijk} \tag{13}$$

$$x_{sijkpqm} \leq x_{spqm} \tag{14}$$

$$x_{sijkpqm} + (1 - x_{sijk}) + (1 - x_{spqm}) \geq 1 \tag{15}$$

Next we include constraints similar to Eq. 8 (contiguous slots are filled), but omit them for brevity. Finally, we add constraints to ensure at least two phrases are selected for each sentence, to promote informative descriptions.

Linguistic constraints: We include linguistically motivated constraints to generate syntactically and semantically plausible sentences. First we enforce a noun-phrase to be selected to ensure semantic relevance to the image:

$$\forall s, \quad \sum_{ik} x_{siNPk} = 1 \tag{16}$$

Also, to avoid content redundancy, we allow at most one phrase of each type:

$$\forall sj, \quad \sum_{i} \sum_{k=1}^{N} x_{sijk} \leq 1 \tag{17}$$

Discourse constraints: We allow at most one prepositional scene phrase for the whole description to avoid redundancy:

$$\text{For } j = PPscene, \quad \sum_{sik} x_{sijk} \leq 1 \tag{18}$$

We add constraints that prevent the inclusion of more than one phrase with identical head words: $\forall s, ij, pq$ with the same heads,

$$\sum_{k=1}^{N} x_{sijk} + \sum_{k=1}^{N} x_{spqk} \leq 1 \tag{19}$$

5.3 Unary Phrase Selection

Let M_{sij} be the confidence score for phrase x_{sij} given by the image–phrase matching algorithm (§2). To make the scores across different phrase types comparable, we normalize them using Z-score: $F_{sij} = norm'(M_{sij}) = (M_{sij} - mean_j)/dev_j$, and then transform the values into the range of [0,1].

5.4 Pairwise Phrase Cohesion

In this section, we describe the pairwise phrase cohesion score F_{sijpq} defined for each x_{sijpq} in

ILP: This is a photo of this bird hopping around eating things off of the ground by river.

Human: IMG_6892 Lookn up in the sky its a bird its a plane its ah..... you

ILP: This is a sporty little red convertible made for a great day in Key West FL. This car was in the 4th parade of the apartment buildings.

Human: Hard rock casino exotic car show in June

ILP: Taken in front of my cat sitting in a shoe box. Cat likes hanging around in my recliner.

Human: H happily rests his armpit on a warm Gatorade bottle of water (a small bottle wrapped in a rag)

ILP: I like the way the clouds hanging down by the ground in Dupnitsa of Avikwalal.

Human: Car was raised on the wall over a bridge facing traffic..paramedics were attending the driver on the ground

Figure 2: In some cases (16%), ILP generated captions were preferred over human written ones!

the objective function (Eq. 12). Via F_{sijpq}, we aim to quantify the degree of syntactic and semantic cohesion across two phrases x_{sij} and x_{spq}. Note that we subtract this cohesion score from the objective function. This trick helps the ILP solver to generate sentences with varying number of phrases, rather than always selecting the maximum number of phrases allowed.

N-gram Cohesion Score: We use n-gram statistics from the Google Web 1-T dataset (Brants and Franz., 2006) Let L_{sijpq} be the set of all n-grams ($2 \leq n \leq 5$) across x_{sij} and x_{spq}. Then the n-gram cohesion score is computed as:

$$F^{\text{NGRAM}}_{sijpq} = 1 - \frac{\sum_{l \in L_{sijpq}} NPMI(l)}{size(L_{sijpq})} \quad (20)$$

$$NPMI(ngr) = \frac{PMI(ngr) - PMI_{min}}{PMI_{max} - PMI_{min}} \quad (21)$$

Where NPMI is the normalized point-wise mutual information.[4]

Co-occurrence Cohesion Score: To capture long-distance cohesion, we introduce a co-occurrence-based score, which measures order-preserved co-occurrence statistics between the head words h_{sij} and h_{spq} [5]. Let $f_\Sigma(h_{sij}, h_{spq})$ be the sum frequency of all n-grams that start with h_{sij}, end with h_{spq} and contain a preposition $prep(spq)$ of the phrase spq. Then the co-occurrence cohesion is computed as:

$$F^{\text{CO}}_{sijpq} = \frac{max(f_\Sigma) - f_\Sigma(h_{sij}, h_{spq})}{max(f_\Sigma) - min(f_\Sigma)} \quad (22)$$

Final Cohesion Score: Finally, the pairwise phrase cohesion score F_{ijpq} is a weighted sum of n-gram and co-occurrence cohesion scores:

$$F_{sijpq} = \frac{\alpha \cdot F^{\text{NGRAM}}_{sijpq} + \beta \cdot F^{\text{CO}}_{sijpq}}{\alpha + \beta} \quad (23)$$

where α and β can be tuned via grid search, and F^{NGRAM}_{ijpq} and F^{CO}_{ijpq} are normalized $\in [0, 1]$ for comparability. Notice that F_{sijpq} is in the range [0,1] as well.

6 Evaluation

TestSet: Because computer vision is a challenging and unsolved problem, we restrict our query set to images where we have high confidence that visual recognition algorithms perform well. We collect 1000 test images by running a large number (89) of object detectors on 20,000 images and selecting images that receive confident object detection scores, with some preference for images with multiple object detections to obtain good examples for testing discourse constraints.

Baselines: We compare our ILP approaches with two nontrivial baselines: the first is an HMM approach (comparable to Yang et al. (2011)), which takes as input the same set of candidate phrases described in §2, but for decoding, we fix the ordering of phrases as [NP – VP – Region PP – Scene PP] and find the best combination of phrases using the Viterbi algorithm. We use the same rich set of pairwise

[4]We include the n-gram cohesion for the sentence boundaries as well, by approximating statistics for sentence boundaries with punctuation marks in the Google Web 1-T data.

[5]For simplicity, we use the last word of a phrase as the head word, except VPs where we take the main verb.

cognitive phrases:	HMM with	HMM w/o	ILP with	ILP w/o
	0.111	0.114	0.114	0.116

Table 1: Automatic Evaluation

	ILP selection rate
ILP V.S. HMM (w/o cogn)	67.2%
ILP V.S. HMM (with cogn)	66.3%

Table 2: Human Evaluation (without images)

	ILP selection rate
ILP V.S. HMM (w/o cogn)	53.17%
ILP V.S. HMM (with cogn)	54.5%
ILP V.S. RETRIEVAL	71.8%
ILP V.S. Human	16%

Table 3: Human Evaluation (with images)

phrase cohesion scores (§5.4) used for the ILP formulation, producing a strong baseline[6].

The second baseline is a recent RETRIEVAL based description method (Ordonez et al., 2011), that searches the large parallel corpus of images and captions, and transfers a caption from a visually similar database image to the query. This again is a very strong baseline, as it exploits the vast amount of image-caption data, and produces a description high in linguistic quality (since the captions were written by human annotators).

Automatic Evaluation: Automatically quantifying the quality of machine generated sentences is known to be difficult. BLEU score (Papineni et al., 2002), despite its simplicity and limitations, has been one of the common choices for automatic evaluation of image descriptions (Farhadi et al., 2010; Kulkarni et al., 2011; Li et al., 2011; Ordonez et al., 2011), as it correlates reasonably well with human evaluation (Belz and Reiter, 2006).

Table 1 shows the the BLEU @1 against the original caption of 1000 images. We see that the ILP improves the score over HMM consistently, with or without the use of cognitive phrases.

[6]Including other long-distance scores in HMM decoding would make the problem NP-hard and require more sophisticated decoding, e.g. ILP.

	Grammar	Cognitive	Relevance
HMM	3.40(σ=.82)	3.40(σ=.88)	2.25(σ=1.37)
ILP	3.56(σ=.90)	3.60(σ=.98)	2.37(σ=1.49)
Hum.	4.36(σ=.79)	4.77(σ=.66)	3.86(σ=1.60)

Table 4: Human Evaluation: Multi-Aspect Rating (σ is a standard deviation)

Human Evaluation I – Ranking: We complement the automatic evaluation with Mechanical Turk evaluation. In ranking evaluation, we ask raters to choose a better caption between two choices[7]. We do this rating with and without showing the images, as summarized in Table 2 & 3. When images are shown, raters evaluate content relevance as well as linguistic quality of the captions. Without images, raters evaluate only linguistic quality.

We found that raters generally prefer ILP generated captions over HMM generated ones, twice as much (67.2% ILP V.S. 32.8% HMM), if images are not presented. However the difference is less pronounced when images are shown. There could be two possible reasons. The first is that when images are shown, the Turkers do not try as hard to tell apart the subtle difference between the two imperfect captions. The second is that the relative content relevance of ILP generated captions is negating the superiority in linguistic quality. We explore this question using multi-aspect rating, described below.

Note that ILP generated captions are exceedingly (71.8 %) preferred over the RETRIEVAL baseline (Ordonez et al., 2011), despite the generated captions tendency to be more prone to grammatical and cognitive errors than retrieved ones. This indicates that the generated captions must have substantially better content relevance to the query image, supporting the direction of this research. Finally, notice that as much as 16% of the time, ILP generated captions are preferred over the original human generated ones (examples in Figure 2).

Human Evaluation II – Multi-Aspect Rating: Table 4 presents rating in the 1–5 scale (5: perfect, 4: almost perfect, 3: 70∼80% good, 2:

[7]We present two captions in a randomized order.

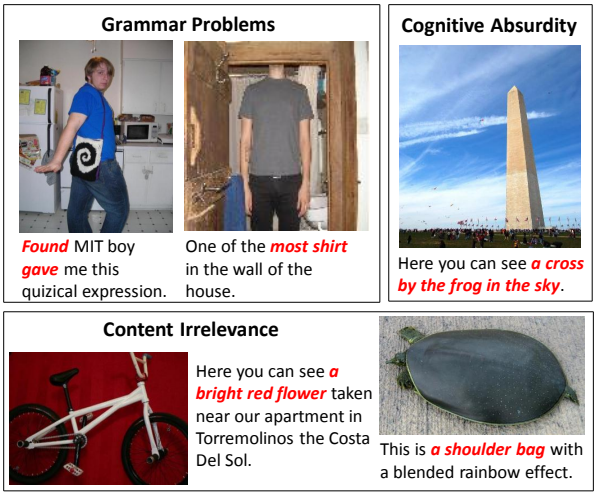

Figure 3: Examples with different aspects of problems in the ILP generated captions.

50∼70% good, 1: totally bad) in three different aspects: *grammar*, *cognitive correctness*,[8] and *relevance*. We find that ILP improves over HMM in all aspects, however, the relevance score is noticeably worse than scores of two other criteria. It turns out human raters are generally more critical against the relevance aspect, as can be seen in the ratings given to the original human generated captions.

Discussion with Examples: Figure 1 shows contrastive examples of HMM vs ILP generated captions. Notice that HMM captions look *robotic*, containing spurious and redundant phrases due to lack of discourse constraints, and often discussing an awkward set of objects due to lack of image-level content planning. Also notice how image-level content planning underpinned by language statistics helps correct some of the erroneous vision detections. Figure 3 shows some example mistakes in the ILP generated captions.

7 Related Work & Discussion

Although not directly focused on image description generation, some previous work in the realm of summarization shares the similar problem of content planning and surface realization. There are subtle, but important differences however. First, sentence compression is hardly the goal of image description generation, as human written descriptions are not necessarily succinct.[9] Second, unlike summarization, we are not given with a set of coherent text snippet to begin with, and the level of noise coming from the visual recognition errors is much higher than that of starting with clean text. As a result, choosing an additional phrase in the image description is much riskier than it is in summarization.

Some recent research proposed very elegant approaches to summarization using ILP for collective content planning and/or surface realization (e.g., Martins and Smith (2009), Woodsend and Lapata (2010), Woodsend et al. (2010)). Perhaps the most important difference in our approach is the use of *negative* weights in the objective function to create the necessary tension between selection (salience) and compatibility, which makes it possible for ILP to generate variable length descriptions, effectively correcting some of the erroneous vision detections. In contrast, all previous work operates with a predefined upper limit in length, hence the ILP was formulated to include as many textual units as possible modulo constraints.

To conclude, we have presented a collective approach to generating natural image descriptions. Our approach is the first to systematically incorporate state of the art computer vision to retrieve visually relevant candidate phrases, then produce images descriptions that are substantially more complex and human-like than previous attempts.

Acknowledgments T. L. Berg is supported in part by NSF CAREER award #1054133; A. C. Berg and Y. Choi are partially supported by the Stony Brook University Office of the Vice President for Research. We thank K. Yamaguchi, X. Han, M. Mitchell, H. Daume III, A. Goyal, K. Stratos, A. Mensch, J. Dodge for data pre-processing and useful initial discussions.

[8] E.g., "A desk on top of a cat" is grammatically correct, but cognitively absurd.

[9] On a related note, the notion of saliency also differs in that human written captions often digress on details that might be tangential to the visible content of the image. E.g., "This is a dress *my mom made*.", where the picture does not show a woman making the dress.

References

Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In *ACL*.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. In *Linguistic Data Consortium*.

James Clarke and Mirella Lapata. 2006. Constraint-based sentence compression: An integer programming approach. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 144–151, Sydney, Australia, July. Association for Computational Linguistics.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA. IEEE Computer Society.

Haris Dindo and Daniele Zambuto. 2010. A probabilistic approach to learning a visually grounded language model through human-robot interaction. In *IROS*, pages 790–796. IEEE.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: generating sentences for images. In *ECCV*.

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. 2010. Object detection with discriminatively trained part based models. *tPAMI*, Sept.

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

Fateh Muhammad Hafiz and Ian Tudor. 1989. Extensive reading and the development of language skills. *ELT Journal*, 43(1):4–13.

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*.

Thomas K. Leung and Jitendra Malik. 1999. Recognizing surfaces using three-dimensional textons. In *ICCV*.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, Portland, Oregon, USA, June. Association for Computational Linguistics.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November.

Andre Martins and Noah A. Smith. 2009. Summarization with a joint model for sentence extraction and compression. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Derek D. Monner and James A. Reggia. 2011. Systematically grounding language through vision in a deep, recurrent neural network. In *Proceedings of the 4th international conference on Artificial general intelligence*, AGI'11, pages 112–121, Berlin, Heidelberg. Springer-Verlag.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

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

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *HLT-NAACL*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *COLING/ACL*.

Deb K. Roy. 2002. Learning visually-grounded words and syntax for a scene description task. Computer Speech and Language, In review.

Wai-King Tsang. 1996. Comparing the effects of reading and writing on writing performance. *Applied Linguistics*, 17(2):210–233.

Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 565–574, Uppsala, Sweden, July. Association for Computational Linguistics.

Kristian Woodsend, Yansong Feng, and Mirella Lapata. 2010. Title generation with quasi-synchronous grammar. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 513–523, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*.

Yezhou Yang, Ching Teo, Hal Daume III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 444–454, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

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