Automatic Caption Generation for News Images

Yansong Feng, *Member*, *IEEE*, and Mirella Lapata, *Member*, *IEEE*

Abstract—This paper is concerned with the task of automatically generating captions for images, which is important for many image-related applications. Examples include video and image retrieval as well as the development of tools that aid visually impaired individuals to access pictorial information. Our approach leverages the vast resource of pictures available on the web and the fact that many of them are captioned and colocated with thematically related documents. Our model learns to create captions from a database of news articles, the pictures embedded in them, and their captions, and consists of two stages. Content selection identifies what the image and accompanying article are about, whereas surface realization determines how to verbalize the chosen content. We approximate content selection with a probabilistic image annotation model that suggests keywords for an image. The model postulates that images and their textual descriptions are generated by a shared set of latent variables (topics) and is trained on a weakly labeled dataset (which treats the captions and associated news articles as image labels). Inspired by recent work in summarization, we propose *extractive* and *abstractive* surface realization models. Experimental results show that it is viable to generate captions that are pertinent to the specific content of an image and its associated article, while permitting creativity in the description. Indeed, the output of our abstractive model compares favorably to handwritten captions and is often superior to extractive methods.

Index Terms—Caption generation, image annotation, summarization, topic models

1 Introduction

Recent years have witnessed an unprecedented growth in the amount of digital information available on the Internet. Flickr, one of the best known photo sharing websites, hosts more than 3 billion images, with approximately 2.5 million images being uploaded every day.[1] Many online news sites like CNN, Yahoo!, and BBC publish images with their stories and even provide photo feeds related to current events. Browsing and finding pictures in large-scale and heterogeneous collections are an important problem that has attracted much interest within information retrieval.

Many of the search engines deployed on the web retrieve images without analyzing their content, simply by matching user queries against collocated textual information. Examples include metadata (e.g., the image's file name and format), user-annotated tags, captions, and, generally, text surrounding the image. As this limits the applicability of search engines (images that do not coincide with textual data cannot be retrieved), a great deal of work has focused on the development of methods that generate description words for a picture *automatically*. The literature is littered with various attempts to learn the associations between image features and words using supervised classification [1], [2], instantiations of the noisy-channel model [3], latent variable models [4], [5], [6], and models inspired by information retrieval [7], [8].

Although keyword-based indexing techniques are popular and the method of choice for image retrieval engines, there are good reasons for using more linguistically meaningful descriptions. A list of keywords is often ambiguous. An image annotated with the words *blue*, *sky*, *car* could depict a blue car or a blue sky, whereas the caption "*car running against the blue sky*" would make the relations between the words explicit. Furthermore, image descriptions tend to be concise, focusing on the most important depicted objects or events. A method that generates such descriptions automatically could therefore improve image retrieval by supporting longer and more targeted queries, by functioning as a short summary of the image's content, and by enabling the use of question-answer interfaces. It could also assist journalists in creating descriptions for the images associated with their articles or in finding images that appropriately illustrate their text. More generally, the ability to link images with textual descriptions would facilitate the retrieval and management of multimedia data (e.g., video and image collections, graphics) as well as increase the accessibility of the web for visually impaired (blind and partially sighted) users who cannot access the content of many sites in the same ways as sighted users can [9].

The standard approach to image description generation adopts a two-stage framework consisting of content selection and surface realization. The former stage analyzes the content of the image and identifies "what to say" (i.e., which events or objects are worth talking about), whereas the second stage determines "how to say it" (i.e., how to render the selected content into natural language text). Both stages are usually manually developed. Content selection makes use of dictionaries that specify a mapping between words and image regions or features [10], [11], [12], [13], [14], and surface realization uses human written templates

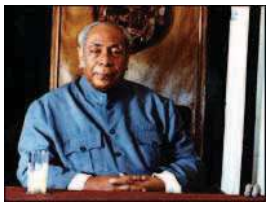
1. http://www.techcrunch.com/2008/11/03/three-billion-photos-at-flickr/.

- *Y. Feng is with the Institute of Computer Science and Technology, Peking University, 128 Zhong Guan Cun North Street, Haidian, Beijing, PO 100871, China. E-mail: fengyansong@pku.edu.cn.*
- *M. Lapata is with the Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh, Room 4.16, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom. E-mail: mlap@inf.ed.ac.uk.*

Manuscript received 6 Oct. 2011; revised 19 Apr. 2012; accepted 12 May 2012; published online 22 May 2012.
Recommended for acceptance by D. Forsyth.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-10-0718.
Digital Object Identifier no. 10.1109/TPAMI.2012.118.

a.

Thousands of Tongans have attended the funeral of King Taufa'ahau Tupou IV, who died last week at the age of 88. Representatives from 30 foreign countries watched as the king's coffin was carried by 1,000 men to the official royal burial ground.



King Tupou, who was 88, died a week ago.

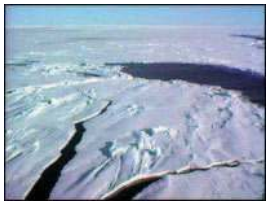
b.

Contaminated Cadbury's chocolate was the most likely cause of an outbreak of salmonella poisoning, the Health Protection Agency has said. About 36 out of a total of 56 cases of the illness reported between March and July could be linked to the product.

Cadbury will increase its contamination testing levels.

c.

A Nasa satellite has documented startling changes in Arctic sea ice cover between 2004 and 2005. The extent of "perennial" ice declined by 14%, losing an area the size of Pakistan or Turkey. The last few decades have seen ice cover shrink by about 0.7% per year.



Satellite instruments can distinguish "old" Arctic ice from "new".

d.

A third of children in the UK use blogs and social network websites but two thirds of parents do not even know what they are, a survey suggests. The children's charity NCH said there was "an alarming gap" in technological knowledge between generations.

Children were found to be far more internet-wise than parents.

Fig. 1. Each entry in the BBC News database contains a document, an image, and its caption (shown in boldface); only the beginning of the documents is shown for the sake of brevity.

or grammars for producing textual output. This approach can create sentences of high quality that are both meaningful and fluent; however, the reliance on manually created resources largely limits the deployment of existing methods to real-world applications. Developing dictionaries that specify exhaustively image-to-text correspondences is a difficult and time-consuming task that must be repeated for new domains and languages. The same is also true for templates and rule-based generation methods; the former are typically specific to the domain in question and not portable to new tasks, whereas the latter can be more general and linguistically sophisticated, albeit with extensive knowledge engineering.

In this paper, we tackle the related problem of generating captions for news images. Our approach leverages the vast resource of pictures available on the web and the fact that many of them naturally co-occur with topically related documents and are captioned. We focus on captioned images embedded in news articles, and learn both models of content selection and surface realization from data without requiring expensive manual annotation. At training time, our models learn from images, their captions, and associated documents, while at test time they are given an image and the document it is embedded in and generate a caption. Compared to most work on image description generation, our approach is shallower, it does not rely on dictionaries specifying image-to-text correspondences, nor does it use a human-authored grammar for the caption creation task. Instead, it uses the document colocated with the image as a proxy for linguistic, visual, and world-knowledge. Our innovation is to exploit this implicit information and treat the surrounding document and caption words as labels for the image, thus reducing the need for human supervision. These labels are weak in the sense that there may be no correspondence between them and regions in the image. However, we argue that the redundancy inherent in such a multimodal dataset allows the development of a fully unsupervised caption generation model, despite noisy input.

Fig. 1 provides example images together with their captions and accompanying documents. Here, document (a) reports on the death of the Tongan King Tupou IV, also depicted in the accompanying image whose caption reads "*King Tupou, who was 88, died a week ago.*" Document (b) discusses a salmonella poisoning outbreak caused by Cadbury's chocolate shown in the image. The caption reads "*Cadbury will increase its contamination testing levels.*" Contrary to image descriptions, captions are contextually relevant to their images but need not describe their specific content in detail. The aim is to create news-worthy text that draws the reader into the accompanying article rather than enumerating the objects in the picture and how they relate to each other. The task itself is challenging even for humans, let alone computers. Along with the title, the lead, and section headings, captions are the most commonly read words in a news article. Due to their prominence, journalists are given explicit instructions on how to write good captions.[2] The latter must be succinct and informative, clearly identify the subject of the picture, establish its relevance to the article, and provide some context for the picture. It is also worth noting that journalists write captions creatively rather than simply cutting and pasting sentences from the document. They do this by relying on general world knowledge and expertise in current affairs that goes beyond what is described in the article or shown in the picture.

Inspired by recent work in summarization, we propose *extractive* and *abstractive* caption generation models. The backbone for both approaches is a probabilistic image

2. A good overview of the caption writing task is given in http://en.wikipedia.org/wiki/Wikipedia:Captions.

annotation model that suggests keywords for an image. The model postulates that images and their textual descriptions are generated by a shared set of latent variables (topics), and is trained on a weakly labeled dataset (consisting of images, their captions, and associated news articles) representative of the scale, diversity, and difficulty of real-world image collections. We consider the keywords output by the image annotation model as a crude approximation of the picture's content. Following an extractive approach, we can then simply identify (and rank) the sentences in the documents that share these keywords. An appealing alternative is to create a new caption that is potentially more concise but also informative and fluent. We propose an abstractive model that operates over image description keywords and document phrases. Their combination gives rise to many caption realizations which we select probabilistically by taking into account dependency and word order constraints.

Our contributions in this work are threefold: We introduce a novel knowledge-lean framework for news image caption generation; we demonstrate that content selection and surface realization models can be learned from weakly labeled data in an unsupervised fashion; and we show that we can compose image captions from scratch without resorting to task-specific templates or image annotations. Experimental results indicate that it is viable to generate captions that are pertinent to the specific content of an image and its associated article, while permitting creativity in the description. Indeed, the output of our abstractive model compares favorably to hand-written captions and is often superior to extractive methods.

2 Related Work

Although image understanding is a popular topic within computer vision, relatively little work has focused on caption generation. As mentioned earlier, a handful of approaches create image descriptions automatically following a two-stage architecture. The picture is first analyzed using image processing techniques into an abstract representation, which is then rendered into a natural language description with a text generation engine. A common theme across different models is domain specificity, the use of hand-labeled data, and reliance on background ontological information.

For example, Héde et al. [13] generate descriptions for images of objects shot in uniform background. Their system relies on a manually created database of objects indexed by an image signature (e.g., color and texture) and two keywords (the object's name and category). Images are first segmented into objects, their signature is retrieved from the database, and a description is generated using templates. Other work (e.g., [11], [12]) creates descriptions for human activities in office scenes. The idea is to extract features of human motion from video keyframes and interleave them with a concept hierarchy of actions to create a case frame from which a natural language sentence is generated. Yao et al. [14] present a general framework for generating text descriptions of image and video content based on image parsing. Specifically, images are hierarchically decomposed into their constituent visual patterns, which are subsequently converted into a semantic representation using WordNet. The image parser is trained on a corpus, manually annotated with graphs representing image structure. A multisentence description is generated using a document planner and a surface realizer.

A notable exception to the use of manually crafted resources is Kulkarni et al. [15], who generate natural language descriptions for images while exploiting state-of-the-art image recognition and generation techniques. Their image recognition system extracts visual information as a set of triples describing the depicted objects, their attributes and spatial relationships (e.g., *<furry, sheep>* against *<green, grass>*). These triples are then used to create descriptive sentences (e.g., *There is a furry sheep against the green grass*), while gluing words (e.g., *there*, *is*, *the*) are provided by a language model or templates. Farhadi et al. [16] describe a related system that can match a descriptive sentence to a given image or to obtain images that illustrate a given sentence. Their approach essentially retrieves sentences rather than composing new ones. Nonetheless, images and sentential descriptions are expressed via a shared meaning representation which also takes the form of triples describing the objects, actions, and scenes (e.g., *<bus, parks, street>*, *<plane, fly, sky>*). More recently, Ordonez et al. [17] demonstrate that this sentence retrieval task scales to a large dataset containing 1 million captioned images.

Much work within computer vision has focused on image annotation,[3] a task related to but distinct from image description generation. The goal is to automatically label an image with keywords relating to its content without however attempting to arrange these into a meaningful sentence or text. Despite differences in application and formulation, all previous methods essentially attempt to learn the correlation between image features and words from examples of images manually annotated with keywords. They are typically developed and evaluated on the Corel database, a collection of stock photographs, divided into themes (e.g., *tigers*, *sunsets*) each of which are associated with keywords (e.g., *sun*, *sea*) that are in turn considered appropriate descriptors for all images belonging to the same theme.

Supervised methods define image annotation as a classification task, e.g., by assuming a one-to-one correspondence between vocabulary words and classes or by grouping several words into a class (see [19] for an overview). Unsupervised approaches attempt to discover the underlying connections between visual features and words, typically by introducing latent variables. Standard latent semantic analysis (LSA) and its probabilistic variant (PLSA) have been applied to this task (e.g., [20], [21], [22]). More sophisticated models estimate the joint distribution of words and regional image features while treating annotation as a problem of statistical inference in a graphical model (e.g., [4], [5], [6]). Relevance models, originally developed for information retrieval, have been also successfully used for image annotation (e.g., [7], [8]). A key idea behind these models is to find the images most similar to the test image and then use their shared keywords for annotation.

Although the bulk of image annotation models learn from images and their corresponding keyword tags, a few

3. Approaches to image annotation are too numerous to describe exhaustively; see [18] for an overview.

approaches exploit full-sentence captions and their structure. Examples include associating names mentioned in the captions to faces depicted in news images (e.g., [23], [24]), verbs to body poses [25], and learning models for recognizing objects [26] and their relative importance [27].

Within natural language processing, most previous efforts have focused on generating captions to accompany complex graphical presentations such as pie charts and bars (e.g., [28], [29]), or on using the captions accompanying information graphics to infer their intended message, e.g., the author's goal to convey ostensible increase or decrease of a quantity of interest [30]. Here, the main emphasis is on how best to describe the data in the graph (e.g., by selecting appropriate sentence templates) rather than image content analysis. There is no image processing involved as it is assumed that the data used to create the graphics are available and the goal is to enable users understand the information expressed in them. More recently, Aker and Gaizauskas [31] generated extended captions for images indexed by GPS coordinates (e.g., a multisentence description of the Eiffel Tower, including where and when it was built and by whom, why it is an important site, and so on). They achieve this by summarizing multiple Web documents that contain information related to an image's location, without, however, taking any visual information into account.

The task of generating captions for news images is novel to our knowledge. Instead of relying on manual annotation or background ontological information we exploit a multimodal database of news articles, images, and their captions. The latter is admittedly noisy, yet can be easily obtained from online sources and contains rich information about the entities and events depicted in the images and their relations. Similarly to previous work, we also follow a two-stage approach. Using an image annotation model, we first describe the picture with keywords, which are subsequently realized into a human readable sentence. Contrary to Kulkarni et al. [15], we do not produce detailed image descriptions; therefore our image analysis is more lightweight (e.g., we do not aim to detect all depicted objects and their relations). The caption generation task bears some resemblance to headline generation [32], where the aim is to create a very short summary for a document. However, we wish to create a caption that not only summarizes the document [31] but is also faithful to the image's content (i.e., the caption should also mention some of the objects or individuals depicted in the image). We therefore explore extractive and abstractive summarization models that rely on visual information to drive the generation process. Our extractive models are close in spirit to Farhadi et al. [16]—the caption generation task can be conceived as that of retrieving the sentence in the document most similar to the image in question. Importantly, we do not make any assumptions regarding the content or structure of the images and as a result our approach is better suited at creating captions for open-domain images. In their work, images are parsed into $<object, action, scene>$ triples and represent 20 categories (taken from the PASCAL 2008 dataset). In contrast, our abstractive models are able to *generate* new sentences (e.g., by reusing and recombining phrases and words from the news article) as opposed to *retrieving* the best matching ones.

3 Problem Formulation

We formulate the image caption generation task as follows: Given a news image I and its associated document $\mathcal{D}$, create a natural language caption C that captures the image's content given $\mathcal{D}$. The training data thus consists of document-image-caption tuples like the ones shown in Fig. 1. During testing, we are given a document and an associated image for which we must generate a caption.

3.1 BBC News Database

Our experiments used news articles accompanied by captioned images. Most image-related datasets used in computer vision and image retrieval are not suitable for caption generation since they have been developed with different tasks in mind. Examples include image annotation and segmentation, object recognition, scene analysis, or image parsing [14], [33], [34], [35], [36], [37], [38], [39]. Existing datasets often contain images (depicting one or two prominent objects against a relatively simple background) and annotation keywords (in the range of $[20, 300]$ words) rather than captions. The datasets created by Farhadi et al. [16] and Hodosh et al. [40] contain image descriptions; however, as mentioned above, they are limited to specific object categories and scene types (e.g., actions).

For these reasons, we created our own dataset[4] by downloading articles (3,361 in total) from the BBC News[5] website. The dataset covers a wide range of topics including national and international politics, technology, sports, education, and so on. News articles normally use color images which are around 200 pixels wide and 150 pixels high. The average caption length is 9.5 words, the average sentence length is 20.5 words, and the average document length 421.5 words. The caption vocabulary is 6,180 words and the document vocabulary is 26,795. The vocabulary shared between captions and documents is 5,921 words. The captions tend to use half as many words as the document sentences, and more than 50 percent of the time contain words that are not attested in the document (even though they may be attested in the collection).

This dataset differs from more typical image collections both in form and content. Besides images and words describing them, it also contains documents whose importance in our case is twofold: First, the document contains the necessary background information which the image depicts or supplements. Second, we can exploit the rich linguistic information inherent in the text and address caption generation with methods akin to text summarization without extensive knowledge engineering. The images and captions in this collection also deviate from more traditional datasets. News images are occasionally cluttered, they display several objects (not only a few prominent ones) and complex scenes, and are often rendered in low resolution. As explained earlier, captions also differ from canonical image descriptions; although they can be denotative (describe some of the objects the image depicts), they can also express connotative meanings (i.e., describe sociological, political, or economic attitudes reflected in the image, or the accompanying document). Image captions

4. Available from http://homepages.inf.ed.ac.uk/s0677528/data.html.
5. http://news.bbc.co.uk/.

may be easy to obtain and cost free, but are admittedly noisy and far from ideal. Although performed by experts (i.e., journalists), the caption generation task is not constrained in any way—words and syntactic structures are chosen with the aim of creating a good caption rather than rendering the task amenable to current vision and language generation techniques. Luckily, the images are accompanied with collateral text, which we argue can be informative and make up for the noise.

3.2 Data Validation

The dataset just described will serve two purposes. First, we will use the images, captions, and associated articles as training data to learn an image annotation model that will provide description keywords for the picture. These keywords will be then used to guide our caption generation model. Second, the human authored captions will function as a gold standard for the image annotation model and for the end-to-end caption generation task. In the former case, we will remove stopwords and treat the caption as a bag of content words (i.e., nouns, adjectives, and verbs). As the image annotation model plays a key role in our generation process, it is important to assess the quality of the captions as labels and whether they do indeed capture some of the image's content. There is no point in learning an image annotation model on labels that are extremely noisy or plainly wrong.

To assess the level of noise in the dataset, we therefore randomly selected 240 image-caption pairs and manually examined whether the content words (nouns, verbs, and adjectives) present in the captions could in principle describe the image. We found out that the captions expressed the picture's content 90 percent of the time. Furthermore, approximately 88 percent of the nouns in subject or object position described salient objects in the pictures. We also conducted a larger scale study to assess the quality of the caption words as annotation labels. In our experiment, participants were presented with a news picture followed by a set of annotation keywords and an associated news document. They were asked to rate these keywords by how well they described the news image given the accompanying document. Participants used a seven-point rating scale where high ratings indicate that words are closely related to both the image and document. We randomly selected from our dataset 30 documents, together with their images and captions. We collected ratings from 26 unpaid volunteers, all self-reported native English speakers. The experiment was conducted remotely over the Internet using WebExp [41], an interactive software package for administering Web-based experiments. Our experimental instructions are given in Appendix A, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.118.

Fig. 2 shows the proportion of caption words given a rating of 1, 2, 3, and so on. As can be seen, the majority of words (more than 71.5 percent) were given a rating of 4 or higher. The mean rating was 4.41, suggesting that participants perceived the caption words as a reasonable approximation of the image's content (given the associated document). We further assessed how well participants agreed in their judgments. We calculated intersubject agreement as the mean pairwise correlation (Pearson's r) between the ratings they produced. The mean correlation coefficient was 0.514 (with a standard deviation of 0.053, a maximum of 0.598, and a minimum of 0.412). This indicates that there is a fair amount of agreement with respect to the keywords in question, i.e., participants agree that these are mostly appropriate descriptors of the images. In the following sections, we present our modeling approach, focusing first on content selection (Section 4.1) and then moving on to discuss surface realization (Sections 4.2 and 4.3).

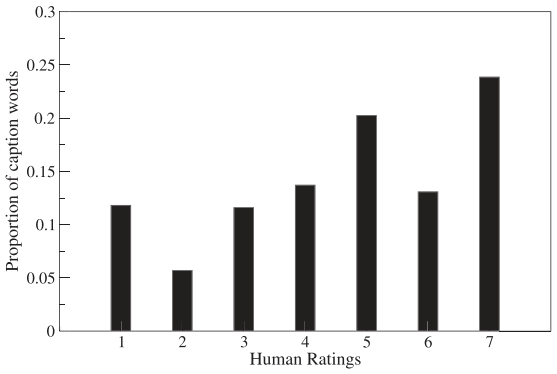


Fig. 2. Proportion of caption words given a [1-7] rating by human judges. The rating task involved assessing how well the caption words captured the image's content in relation to the accompanying article.

4 Modeling

Our model consists of two stages. Content selection identifies what the image and accompanying article are about, whereas surface realization determines how to verbalize the chosen content. Before describing our model in detail, we summarize our assumptions regarding the caption generation task and the nature of the data on which it is being modeled.

1. The caption describes the content of the image directly or indirectly. Unlike traditional image annotation where keywords describe salient objects, captions supply more detailed information, not only about objects and their attributes, but also events. For example, in Fig. 1a the caption mentions *King Tupou* shown in the picture but also his age and the fact that he died a week ago.
2. The accompanying document describes the content of the image. This is trivially true for news documents where the images conventionally depict events, objects or people mentioned in the article.
3. Since our images are implicitly rather than explicitly labeled, we do not assume that we can enumerate *all* objects present in the image nor that we can create a detailed description of them. Instead, we hope to model event-related information such as "what happened," "who did it," and "where" with the help of the news document.

4.1 Image Content Selection

We define a probabilistic image annotation model based on the assumption that images and their surrounding text are generated by a shared set of latent variables or topics.

Specifically, we describe documents and images by a common multimodal vocabulary consisting of textual words and visual terms. Due to polysemy and synonymy, many words in this vocabulary will refer to the same underlying concept. Using Latent Dirichlet Allocation (LDA [4]), a probabilistic model of text generation, we represent visual and textual meaning jointly as a probability distribution over a set of topics. Our annotation model takes these topic distributions into account while finding the most likely keywords for an image and its associated document.

4.1.1 Image and Document Representation

Words and images represent distinct modalities; images live in a continuous feature space, whereas words are discrete. Yet, both modalities on some level capture the same underlying concepts as they are used to describe the same objects. A common first step in previous image annotation methods is the segmentation of the picture into regions, using either a fixed-grid layout or an image segmentation algorithm. Regions are then described by a standard set of features, including color, texture, and shape, and subsequently treated as continuous vectors (e.g., [5], [42]) or in quantized form (e.g., [3], [22]). Through this process, the low-level image features are made to resemble word-like units.

In our work, visual features receive a discrete representation and each image is treated as a bag of visual words. In order to do this we use the Scale Invariant Feature Transform (SIFT) algorithm [43], [44]. The general idea behind the algorithm is to first sample an image with the difference-of-Gaussians point detector at different scales and locations. Each detected region is represented with the SIFT descriptor, which is a histogram of directions at different locations in the detected region and scale. Importantly, this descriptor is, to some extent, invariant to translation, scale, rotation, and illumination changes. SIFT features have been shown to be superior to other descriptors [45] and are considered state of the art in object recognition [46]. We further quantize the SIFT descriptors using the K-means clustering algorithm to obtain a discrete set of visual terms which form our visual vocabulary Voc_v. Each entry in this vocabulary represents a group of image regions which are similar in content or appearance and assumed to originate from similar objects. More formally, each image I is expressed in a bag-of-words format vector, $[w_{v_1}, w_{v_2}, \ldots, w_{v_L}]$, where $w_{v_i} = n$ only if I has n regions labeled with v_i.

Since visual and textual modalities now have the same status—they are both represented as bags of words—we can also represent any image-caption-document tuple *jointly* as a mixed document d_{Mix}. The underlying assumption is that the two modalities express the same meaning, which, as we explain below, can be operationalized as a probability distribution over a set of topics. For ease of exposition, we first describe the basics of LDA and then move on to discuss our image annotation model which makes use of probabilities estimated by LDA.

4.1.2 Latent Dirichlet Allocation

LDA can be represented as a three-level hierarchical Bayesian model, shown graphically in Fig. 3. Given a corpus consisting of D documents, each document is modeled using a mixture over K topics (assumed to follow a multinomial distribution θ with a Dirichlet prior), which are in turn characterized as distributions over words. The words in the document are generated by repeatedly sampling a topic according to the topic distribution, and selecting a word given the chosen topic. Blei et al. [47] describe the generative process for a document d as follows:

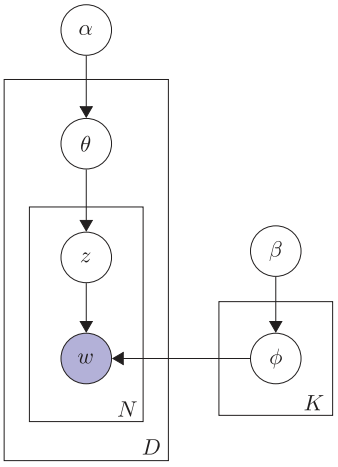


Fig. 3. The LDA topic model: Shaded nodes represent observed variables, unshaded nodes indicate latent variables. Arrows indicate conditional dependencies between variables, whereas plates (the rectangles in the figure) refer to repetitions of sampling steps. The variables in the lower right corner refer to the number of samples.

1. choose $\theta|\alpha \sim Dir(\alpha)$,
2. for $n \in 1, 2, \ldots, N$:
   a. choose topic $z_n|\theta \sim Mult(\theta)$,
   b. choose a word $w_n|z_n, \beta_{1:K} \sim Mult(\beta_{z_n})$,

where each entry of $\beta_{1:K}$ is a distribution over words, indicating a topic definition.

The mixing proportion over topics θ is drawn from a Dirichlet prior with parameters α whose role is to create a smoothed topic distribution. Once α and β are sampled, then each document is generated according to the topic proportions $z_{1:K}$ and word probabilities over topics β. The probability of a document d in a corpus can be obtained as

$$P(d|\alpha,\beta) = \int_{\theta} P(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_k} P(z_k|\theta)P(w_n|z_k,\beta)\right)\mathrm{d}\theta. \quad (1)$$

The central computational problem in LDA is to estimate $P(\theta, z_{1:k}|d,\alpha,\beta)$, the posterior distribution of the hidden variables given a document. Although it is generally intractable to compute this distribution directly, a variety of approximate inference algorithms have been proposed in literature. We follow the convexity-based variational inference procedure described in Blei et al. [47], which involves two steps: 1) introducing variational parameters in order to find the tightest lower bound for the target posterior distribution, and 2) obtaining the tight lower bound through minimizing the Kullback-Leibler (KL) divergence between the introduced variational distribution and the true posterior distribution (we refer interested readers to Blei et al. [47] and Blei [4] for more details on their inference procedure).

An LDA model trained on a document collection yields two sets of parameters, $P(w|z_{1:K})$, the word probabilities given topics, and $P(z_{1:K}|d)$, the topic proportions for each document. The latter are document-specific, whereas the former represent the set of topics (in the form of word conditional probabilities) learned from the document collection. Given a trained model, it is also possible to perform inference on an unseen document d_{new}, and obtain the approximate topic proportions as

$$p(z|d_{new}) \approx \frac{\gamma}{\sum_{j=1}^{K}\gamma_j}, \tag{2}$$

where $\gamma_{1:K}$ are variational Dirichlet parameters obtained during inference on the new document. We can further compute word predictive probabilities given an unseen document as

$$p(w|d_{new}) \approx \sum_{k=1}^{K} P(w|z_k)\frac{\gamma_k}{\sum_{j=1}^{K}\gamma_j}, \tag{3}$$

where $P(w|z_{1:K})$ are word probabilities over topics $z_{1:K}$ learned during model training.

4.1.3 Image Annotation Model

In a standard image annotation setting, a hypothetical model is given an image I and a set of keywords W, and must find the subset W_I $(W_I \subseteq W)$ which appropriately describes the image I:

$$W_I^* = \arg\max_{W} P(W|I). \tag{4}$$

The keywords are usually assumed to be conditionally independent of each other, so the above equation can be simplified as

$$W_I^* = \arg\max_{W} \prod_{w\in W} P(w|I). \tag{5}$$

Recall that each entry in our dataset is an image-caption-document tuple $(I, C, \mathcal{D})$. In this setting, our model must find a subset of keywords W_I that appropriately describe image I with the help of the accompanying document $\mathcal{D}$:

$$W_I^* = \arg\max_{W_t} P(W_t|I, \mathcal{D}). \tag{6}$$

Here, W_t denotes a set of textual words (we use the subscript t to discriminate from the visual words which are not part of the model's output). We also assume that the keywords are conditionally independent of each other:

$$W_I^* = \arg\max_{W_t} P(W_t|I, \mathcal{D}) = \arg\max_{W_t} \prod_{w_t\in W_t} P(w_t|I, \mathcal{D}). \tag{7}$$

Since I and $\mathcal{D}$ are represented jointly as the concatenation of textual and visual terms, we may intuitively simplify the problem and use the mixed document representation d_{Mix} directly in estimating the conditional probabilities $P(w_t|I, \mathcal{D})$:

$$P(w_t|I, \mathcal{D}) \approx P(w_t|d_{Mix}). \tag{8}$$

Substituting (8) into (7) yields

$$W_I^* = \arg\max_{W_t} P(W_t|I, \mathcal{D}) \approx \arg\max_{W_t} \prod_{w_t\in W_t} P(w_t|d_{Mix}). \tag{9}$$

As mentioned earlier, we assume that the image and its associated text are generated by a mixture of latent topics which we infer using LDA. Specifically, we select the number of topics K and apply the LDA algorithm to a corpus consisting of documents $\{d_{Mix}\}$ in order to obtain the multimodal word distributions over topics $P(w|z_{1:K})$, and the estimated posterior of the topic proportions over documents $P(z_{1:K}|d_{Mix})$.

Given an unseen image-document pair, it is also possible to approximately infer the topic proportions $P(z_{1:K}|d_{Mix_{new}})$ on the new document $d_{Mix_{new}}$ using (2). We then substitute (3) into (9)[6]:

$$\begin{aligned} W_I^* &\approx \arg\max_{W_t} \prod_{w_t\in W_t} P(w_t|d_{Mix}) \\ &= \arg\max_{W_t} \prod_{w_t\in W_t} \sum_{k=1}^{K} P(w_t|z_k)P(z_k|d_{Mix}) \\ &\approx \arg\max_{W_t} \prod_{w_t\in W_t} \sum_{k=1}^{K} P(w_t|z_k)\frac{\gamma_k}{\sum_{j=1}^{K}\gamma_j}, \end{aligned} \tag{10}$$

where $P(w_t|z_k)$ are obtained during training and $\gamma_{1:K}$ are inferred on the image-document test pair.

However, note that for an unseen image d_I and accompanying document $d_{\mathcal{D}}$, the estimated topic proportions are solely based on variational inference, which is an approximate algorithm. In order to render the model more robust, we further smooth the topic proportions $P(z_{1:K}|d_{Mix})$ with probabilities based on each modality:

$$\begin{aligned} P^*(z_{1:K}|d_{Mix}) \approx\ & q_1 P(z_{1:K}|d_{Mix}) \\ & + q_2 P(z_{1:K}|d_{\mathcal{D}}) \\ & + q_3 P(z_{1:K}|d_I), \end{aligned} \tag{11}$$

where $P(z_{1:K}|d_{\mathcal{D}})$ and $P(z_{1:K}|d_I)$ are inferred on $d_{\mathcal{D}}$ and d_I, respectively, and q_1, q_2, q_3 are smoothing parameters (which we tune experimentally on held-out data); q_3 is a shorthand for $(1 - q_1 - q_2)$. We do not train three separate models for each probability term in (11), but use the mixture document d_{Mix}, the text document $d_{\mathcal{D}}$, and the image d_I to perform inference on the same topic model.

In sum, calculating $P(W_t|I, \mathcal{D})$ boils down to estimating the probabilities $P(w_t|d_{Mix})$ according to (10) and (11), which we obtain using the LDA topic model. We first train an LDA model on the multimodal document collection $\{d_{Mix}\}$ to learn the multimodal topic representations and use inference to obtain the topic distributions of unseen image-document pairs. In the end, for each unseen image-document pair, we obtain the probability over all textual words $\{w_t\}$, the n-best of which we consider as the annotations for image I. Note that the annotation model just described outputs a distribution over the whole vocabulary which can be naturally treated as a ranked word list, but also as a unigram language model. This probabilistic

6. During training, the model has access to all three elements $(I, C, \mathcal{D})$, so the mixed document d_{Mix} is a concatenation of the visual terms and words present in the caption and document. During testing, the model is given an image and its accompanying document, so d_{Mix} will contain words based on I and $\mathcal{D}$, but not C.

formulation is advantageous when generating captions for an image as our generation model is also probabilistic and thus the two components of content selection and surface realization can be easily integrated.

4.2 Extractive Caption Generation

Our extractive caption generator draws inspiration from previous work on automatic summarization, most of which focuses on sentence extraction (see [48] and [49] for comprehensive overviews). The idea is to create a summary simply by identifying and subsequently concatenating the most important sentences in a document. Without a great deal of linguistic analysis, it is possible to create summaries for a wide range of documents, independently of style, text type, and subject matter. For our caption generation task, we need only extract a single sentence. And our guiding hypothesis is that this sentence must be maximally similar to the description keywords generated by the annotation model. Given the probabilistic nature of our image annotation model, we are able to represent the content of an image in two ways, i.e., as a ranked list of keywords and as a distribution of topics. We discuss below different ways of operationalizing the similarity between a sentence and each of these representations.

4.2.1 Word Overlap-Based Sentence Selection

Perhaps the most intuitive way of measuring the similarity between image keywords and document sentences is word overlap:

$$Overlap(W_I, S_d) = \frac{|W_I \cap S_d|}{|W_I \cup S_d|}, \tag{12}$$

where W_I is the set of keywords suggested by our image annotation model and S_d a sentence in the document. The selected caption is then the sentence that has the highest overlap with the image keywords.

4.2.2 Vector Space-Based Sentence Selection

Word overlap is admittedly a naive measure of similarity, based on lexical identity. We can overcome this by representing keywords and sentences in vector space [50] and computing the similarity between the two vectors representing the image keywords and document sentences, respectively. We create a word-sentence co-occurrence matrix where each row represents a word, each column a sentence, and each entry the frequency with which the word appears within the sentence (we are assuming that image keywords also form a sentence). More precisely, matrix cells are weighted by their $tf * idf$ values. The similarity of the vectors representing the keywords $\overrightarrow{W_I}$ and document sentence $\overrightarrow{S_d}$ can be quantified by measuring the cosine of their angle:

$$sim(\overrightarrow{W_I}, \overrightarrow{S_d}) = \frac{\overrightarrow{W_I} \cdot \overrightarrow{S_d}}{|\overrightarrow{W_I}||\overrightarrow{S_d}|}. \tag{13}$$

4.2.3 Topic-Based Sentence Selection

Recall that the backbone of our image annotation model is a probabilistic topic model with images and documents rendered into a bag of visual and textual words and represented as a probability distribution over a set of latent topics. Under this framework, the similarity between an image and a sentence can be broadly measured by the extent to which they share the same topic distributions [51]. For example, we may use the Kullback-Leibler divergence to measure the difference between two distributions p and q:

$$KL(p, q) = \sum_{j=1}^{K} p_j \log_2 \frac{p_j}{q_j}, \tag{14}$$

where p and q are shorthand for the image topic distribution $P_{d_{Mix}}$ and sentence topic distribution P_{S_d}, respectively. We infer the image topic distribution according to the mixed document (using both the image and the document). When doing inference on the document sentence, we also take its neighboring sentences into account to avoid estimating the topic proportions on short sentences inaccurately.

The KL divergence is asymmetric and, in many applications, it is preferable to apply a symmetric measure such as the Jensen Shannon (JS) divergence. The latter measures the "distance" between p and q through $\frac{(p+q)}{2}$; the average of p and q are as follows:

$$JS(p, q) = \frac{1}{2}\left[KL\left(p, \frac{(p+q)}{2}\right) + KL\left(q, \frac{(p+q)}{2}\right)\right]. \tag{15}$$

4.3 Abstractive Caption Generation

Although extractive methods yield naturally grammatical captions and require relatively little linguistic analysis, there are a few caveats to consider. As discussed before, there is often no single sentence in the document that uniquely describes the image's content. In most cases the keywords are found in the document but interspersed across multiple sentences. Second, the selected sentences make for long captions (sometimes longer than the average document sentence), which are not concise and overall not as catchy as human-written captions. For these reasons, we turn to abstractive caption generation and present models based on single words but also phrases.

4.3.1 Word-Based Caption Generation

Banko et al. [32] (see also [52]) propose a bag-of-words model for headline generation. Following the traditional natural language generation paradigm, their model consists of a content selection and surface realization component. Content selection is modeled as the probability of a word appearing in the headline given that the same word appears in the corresponding document and is independent of other words in the headline. The likelihood of different surface realizations is estimated using a bigram model. They also take the distribution of the length of the headlines into account in an attempt to bias the model toward generating output of reasonable length (around five words):

$$\begin{aligned} P(w_1, w_2, \ldots, w_n) = &\prod_{i=1}^{n} P(w_i \in H | w_i \in \mathcal{D}) \\ &\cdot P(len(H) = n) \\ &\cdot \prod_{i=2}^{n} P(w_i | w_{i-1}), \end{aligned} \tag{16}$$

where w_i is a word that may appear in headline H, $\mathcal{D}$ the document being summarized, and $P(len(H) = n)$ is a

headline length distribution model. Specifically, Banko et al. [32] assume that headline length follows a normal distribution which they learn from a training corpus (the 1997 Reuters News Stories).

The above model can be easily adapted to our caption generation task. Content selection is now the probability of a word appearing in the caption given the image and its associated document, which we obtain from the output of our image annotation model. In addition, we replace the bigram language model with a trigram one:

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i \in C|I, \mathcal{D}) \cdot P(len(C) = n) \cdot \prod_{i=3}^{n} P(w_i|w_{i-1}, w_{i-2}), \quad (17)$$

where C is the caption, I the image, $\mathcal{D}$ the accompanying document, and $P(w_i \in C|I, \mathcal{D})$ the image annotation probability.

Despite its simplicity, the caption generation model in (17) has a major drawback. As the image annotation model does not take function words into account, content selection will ignore them too, at the expense of the grammaticality of the generated captions. In other words, there will be no function words to glue the content words together. One way to remedy this is to revert to a content selection model that ignores the image and simply estimates the probability of a word appearing in the caption given the same word appearing in the document, while at the same time, the model takes note of the image annotation probabilities during surface realization. We do this by modifying the language model responsible for surface realization so that it prefers words that have high image annotation probabilities and are likely to appear in a sentence according to a background language model. We use an *adaptive* language model [53] that modifies an n-gram model with local unigram probabilities:

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i \in C|w_i \in \mathcal{D}) \cdot P(len(C) = n) \cdot \prod_{i=3}^{n} P_{adap}(w_i|w_{i-1}, w_{i-2}), \quad (18)$$

where $P(w_i \in C|w_i \in \mathcal{D})$ is the probability of w_i appearing in the caption given that it appears in the document $\mathcal{D}$, and $P_{adap}(w_i|w_{i-1}, w_{i-2})$ is the language model adapted with probabilities from our image annotation model:

$$P_{adap}(w|h) = \frac{\alpha(w)}{z(h)} P_{word}(w|h), \quad (19)$$

$$\alpha(w) \approx \left(\frac{P_{image}(w)}{P_{word}(w)}\right)^{\beta}, \quad (20)$$

$$z(h) = \sum_{w} \alpha(w) \cdot P_{word}(w|h), \quad (21)$$

where $P_{word}(w|h)$ is the probability of w given the history h of preceding words (i.e., the original trigram model), $P_{image}(w)$ is the probability of w according to the image annotation model, $P_{word}(w)$ is the probability of w according to the original background language model, and β is a scaling parameter.

The model in (18) has three components. The conditional probability $P(w_i \in C|w_i \in \mathcal{D})$ captures the gist of the article; the adapted language model $P_{adap}(w_i|w_{i-1}, w_{i-2})$ ensures that the output is grammatical and consistent with the associated image. The length component $P(len(C) = n)$, modeled as a normal distribution, modulates the caption length.

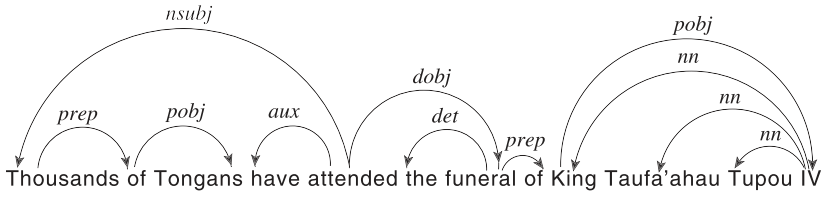
4.3.2 Phrase-Based Caption Generation

The model outlined in (18) will generate captions with function words. However, there is no guarantee that these will be compatible with their surrounding context or that the captions will be globally coherent beyond the trigram horizon. To avoid these problems, we turn our attention to phrases which are naturally associated with function words and may potentially capture long-range dependencies. Phrases have been previously used in abstractive summarization. For example, Zhou and Hovy [54] first identify a list of keywords which are then used to extract phrases from the document. The phrases are linked together to create headlines using a set of handwritten rules. Building on this approach, Soricut and Marcu [55] identify a list of keywords but also use syntactic information (extracted from parse trees) to build syntactically driven phrases around the extracted keywords. Finally, Wan et al. [56] extract dependencies from the input document and glue them together using n-grams.

Although it is relatively straightforward to extend content selection from individual words to phrases, this poses additional difficulties for surface realization. Realizers based on language models are typically built from individual words rather than phrases and as a result they do not take phrase adjacency constraints into account. Our model relies on phrases which we obtain from the output of a dependency parser. A phrase is simply a head and its dependents (or modifiers), with the exception of verbs, where we record only the head (otherwise, an entire sentence could be a phrase). Fig. 4 shows the dependency representation for the sentence "*Thousands of Tongans have attended the funeral of King Taufa'ahau Tupou IV*" and the set of phrases extracted from it. We only consider dependencies whose heads are nouns, verbs, and prepositions, as these constitute 80 percent of all dependencies attested in our caption data.

We define a bag-of-phrases model for caption generation by modifying the content selection and caption length components in (18) as follows:

$$P(\rho_1, \rho_2, \ldots, \rho_m) \approx \prod_{j=1}^{m} P(\rho_j \in C|\rho_j \in \mathcal{D}) \cdot P\left(len(C) = \sum_{j=1}^{m} len(\rho_j)\right) \cdot \prod_{i=3}^{L} P_{adap}(w_i|w_{i-1}, w_{i-2}), \quad (22)$$



Extracted Phrases
thousands, of Tongans, thousands attended, have attended, the funeral, attended funeral, funeral of, of King Taufa'ahau Tupou IV, Tupou IV, Taufa'ahau Tupou IV, King Taufa'ahau Tupou IV, of IV

Fig. 4. Example of a dependency graph for the sentence "*Thousands of Tongans have attended the funeral of King Tuafa'ahau.*" Directed edges represent heads and their dependents. Phrases extracted from this graph are shown in the adjacent table.

where $L = \sum_{j=1}^{m} len(\rho_j)$. The term $P(\rho_j \in C|\rho_j \in \mathcal{D})$ models the probability of phrase ρ_j appearing in the caption given that it also appears in the document and is estimated as

$$P(\rho_j \in C|\rho_j \in \mathcal{D}) = \prod_{w_j \in \rho_j} P(w_j \in C|w_j \in \mathcal{D}), \tag{23}$$

where w_j is a word in the phrase ρ_j.

One problem with the models discussed thus far is that words or phrases are independent of each other. It is up to the trigram model to enforce coarse ordering constraints. These may be sufficient when considering isolated words, but phrases are longer and their combinations are subject to structural constraints that are not captured by sequence models. We therefore attempt to take phrase *adjacency* constraints into account by estimating the probability of phrase ρ_j attaching to the right of phrase ρ_i as

$$\begin{aligned} P(\rho_j|\rho_i) &= \sum_{w_i \in \rho_i} \sum_{w_j \in \rho_j} p(w_j|w_i) \\ &= \frac{1}{2} \sum_{w_i \in \rho_i} \sum_{w_j \in \rho_j} \left\{ \frac{f(w_i, w_j)}{f(w_i, -)} + \frac{f(w_i, w_j)}{f(-, w_j)} \right\}, \end{aligned} \tag{24}$$

where $p(w_j|w_i)$ is the probability of a phrase containing word w_j appearing to the right of a phrase containing word w_i, $f(w_i, w_j)$ indicates the number of times two phrases containing w_i and w_j are adjacent, $f(w_i, -)$ is the number of times w_i appears on the left of any phrase, and $f(-, w_i)$ the number of times it appears on the right.[7]

After integrating the adjacency probabilities into (22), the caption generation model becomes

$$\begin{aligned} P(\rho_1, \rho_2, ..., \rho_m) \approx & \prod_{j=1}^{m} P(\rho_j \in C|\rho_j \in \mathcal{D}) \\ & \cdot \prod_{j=2}^{m} P(\rho_j|\rho_{j-1}) \\ & \cdot P(len(C) = \sum_{j=1}^{m} len(\rho_j)) \\ & \cdot \prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i|w_{i-1}, w_{i-2}). \end{aligned} \tag{25}$$

The model in (25) takes long distance dependency constraints into account and has some notion of syntactic structure through the use of attachment probabilities. As it has a primitive notion of caption length estimated by $P(len(C) = \sum_{j=1}^{m} len(\rho_j))$, it will invariably generate captions of similar (phrase) length. Ideally, we would like the model to modulate the length of its output depending on the chosen content. However, we leave this to future work.

7. Equation (24) is smoothed to avoid zero probabilities.

4.3.3 Search

To generate a caption, it is necessary to find the sequence of words that maximizes $P(w_1, w_2, \ldots, w_n)$ for the word-based model (18) and $P(\rho_1, \rho_2, \ldots, \rho_m)$ for the phrase-based model (25). We rewrite both probabilities as the weighted sum of their log form components and use beam search to find a near-optimal sequence. Note that we can make search more efficient by reducing the size of the document $\mathcal{D}$. Using one of our extractive generation models from Section 4.2, we may rank its sentences in terms of their relevance to the image content and consider only the n-best ones. Alternatively, we could consider the single most relevant sentence together with its surrounding context under the assumption that neighboring sentences are about the same or similar topics.

5 Evaluation

In this section, we evaluate the caption generation models presented above. We give details on our training procedure, parameter estimation, and present the baseline methods used for comparison with our models. We first discuss results on the performance of the image annotation model (Section 4.1) and then evaluate the caption generation task as a whole. The image annotation model is used as a proxy to content selection; it highlights important objects or events depicted in the image (and mentioned in the document) that should also figure in the generated caption.

5.1 Image Annotation

5.1.1 Data

All our annotation experiments were conducted on the BBC dataset described in Section 3.1. We used 2,881 image-caption-document tuples for training, 240 tuples for development, and 240 for testing. All documents and captions were part-of-speech tagged and lemmatized with Tree Tagger [57]. We excluded from the vocabulary low frequency words (appearing fewer than five times) and words other than nouns, verbs, and adjectives. We preprocessed the images as follows: We first extracted SIFT keypoints with descriptors from each image (150 on average) and then used K-means to quantize these features into a discrete set of visual terms. We varied K experimentally (from 100 to 2,000).

5.1.2 Parameter Tuning

We trained our LDA topic model on the multimodal document collection $\{d_{Mix}\}$, varying the number of topics from 15 to 1,000. The hyperparameter α was set to 0.1; β, the

word-topic probability table, was initialized randomly. The maximum number of iterations for variational inference was set to 1,000. We tuned the smoothing parameters q_1, q_2, and q_3 (see (12)) on the development set. The best values were $q_1 = 0.84$, $q_2 = 0.12$, and $q_3 = 0.04$.[8] As the number of visual terms and topics are interrelated, we exhaustively examined all possible combinations on the development set. We obtained the best results on image annotation with 1,000 topics and 750 visual terms.

5.1.3 Comparison Models

We compared our topic model against several baselines. First, we trained a vanilla LDA model (TxtLDA) on the document collection without taking the images into account. This model estimates $P(w_t|\mathcal{D}) = \sum_{k=1}^{K} P(w_t|z_k)P(z_k|\mathcal{D})$, the probability of word w_t given text document $\mathcal{D}$, and assumes that the most probable words are the best keywords for the accompanying image. To assess the individual contribution of the visual information, we also trained an LDA model on image-caption keyword pairs without taking the news articles into account. This model (ImgLDA) estimates $P(w_t|I) = \sum_{k=1}^{K} P(w_t|z_k)P(z_k|I)$, the probability of word w_t given image I. For both models, we tuned the number of topics on the development set; we obtained the best performance with 500 topics for TxtLDA; the combination of 1,000 visual words with 500 topics performed best for ImgLDA. Our third baseline is an extension of the continuous relevance annotation model (ContRel, [7]). Unlike other approaches where a set of latent variables is introduced, each defining a joint distribution on the space of keywords and image features, this model captures the joint probability of images and annotated words *directly*, without requiring an intermediate clustering stage (i.e., each annotated image in the training set is treated as a latent variable). In Feng and Lapata [58], we modified this model so as to exploit the information present in the document with improved results. Our extensions were twofold. First, in estimating the conditional probability of a keyword given an image, we also considered its likelihood in the collateral document. Second, we used an LDA model (trained on the document collection) to prune from the model's output words that are not representative of the document's topics.

We also compared our approach with two closely related latent variable models (originally developed for image-caption pairs), a PLSA-based model [22], and CorrLDA [42]. Following Monay and Gatica-Perez [22], we experimented with three variants of PLSA, namely, PLSA-Words, PLSA-Mixed, and PLSA-Features. These models vary in how they obtain the topic proportions $P(z_{1:K}|d)$ on the training data. PLSA-Words and PLSA-Features are both asymmetric, estimating the topic proportions from a single modality, i.e., the text or the images. In both cases, the estimated topic structure is kept fixed and the other modality (visual or textual) is *folded-in* [21]. PLSA-Mixed uses both the images and annotation keywords to infer the topic space. CorrLDA first generates image regions from a Gaussian multinomial distribution parameterized with Dirichlet priors. Then, for each annotation word, it uniformly selects a region from the image and generates a word according to the topic used to generate that region. Although PLSA and CorrLDA were initially developed for standard image annotation (and thus trained on image-keyword pairs), it is straightforward to adapt them to our setting and train on image-caption-document tuples. We therefore report two sets of results, with or without the document. Parameters for these models were optimized on the development set. For CorrLDA, we followed the mean-field variational inference strategy proposed in Blei [4]. The optimal number of topics for PLSA was 200 and for CorrLDA was 50. In both cases, the optimal number of visual terms was 2,000.

Finally, as a sanity check, we compared our model against two simple baselines. The first baseline ranks the content words (i.e., nouns, verbs, and adjectives) appearing in each document according to their $tf * idf$ weight [50] and selects the top m ones to be the final image keywords. Our second baseline (DocTitle) simply annotates the image with the document's title (excluding stopwords).

5.1.4 Evaluation Method

Our evaluation followed the experimental methodology proposed in Lavrenko et al. [7]. We are given an unannotated image I with its associated document and asked to automatically produce suitable keywords for it. We consider the 10-best output keywords as the annotations for image I and compare them against the original (gold standard) captions. Model performance is evaluated using precision, recall, and F1. In the image annotation task, precision is the percentage of correctly annotated words over all annotations that the system suggested. Recall is the percentage of correctly annotated words over the number of genuine annotations in the test data. F1 is the harmonic mean of precision and recall. These measures are averaged over all items in the test set. In addition to F1, we also report Mean Average Precision (mAP), an evaluation measure commonly used in information retrieval. Mean average precision is the mean of the Average Precision (AP) of a set of queries. The AP of a query is the average of the precision scores at the rank locations of each relevant document (or image in our case). Intuitively, the higher the mAP value a hypothetical query-retrieval system has, the earlier it finds the relevant images. We refer the interested reader to Monay and Gatica-Perez [22] and Buckley and Voorhees [59] for more details. Note that we cannot give mAP scores for models that do not produce a ranking over the entire vocabulary (i.e., $tf * idf$, DocTitle, and ContRel).

5.1.5 Results

Our results are summarized in Table 1. As can be seen, our model (MixLDA) outperforms all comparison models that consider the document alone, without the image. This is true for $tf * idf$ and the baseline based on document titles. MixLDA also outperforms TxtLDA by a large margin in terms of precision (9 percent) and recall (16.2 percent). F1 improves by 11.4 percent and the difference is statistically significant ($p < 0.01$) using a stratified shuffling-based randomization test [60]. MixLDA also achieves a significant increase in F1 over ImgLDA. A similar pattern emerges

8. Note that the contribution of the image term q_3 is minimal. This suggests that the estimation of the topic proportions based on images alone is noisy. The fact that most weight is given to the topic proportions based on d_{Mix} (the term q_1) indicates that visual and textual information are most effective when combined, rather than when being individually considered.

TABLE 1
Automatic Image Annotation Results on the BBC News Dataset

Model	Precision	Recall	F1	mAP
$tf * idf$	4.37	7.09	5.41	NA
DocTitle	9.22	7.03	7.20	NA
TxtLDA	7.30	16.90	10.20	22.76
PLSA-Features	8.80	18.50	12.00	24.72
PLSA-Words	8.99	20.10	12.60	28.54
PLSA-Mixed	8.37	15.90	11.10	19.84
ImgLDA	7.92	17.40	10.60	24.04
CorrLDA	5.33	11.80	7.36	2.27
PLSA-Features$_{\mathcal{D}}$	10.20	21.80	13.80	26.12
PLSA-Words$_{\mathcal{D}}$	10.26	22.60	14.04	26.26
PLSA-Mixed$_{\mathcal{D}}$	10.30	22.60	14.16	26.26
CorrLDA$_{\mathcal{D}}$	3.87	8.74	5.36	2.72
ContRel	14.70	27.90	19.80	NA
MixLDA	16.30	33.10	21.60	35.01

All scores are reported as percentages. PLSA and CorrLDA models trained on image-caption-document tuples are indicated with the subscript $\mathcal{D}$.

when considering the mAP value, which is substantially higher for MixLDA compared to TxtLDA and ImgLDA. Interestingly, TxtLDA and ImgLDA obtain comparable precision, recall, and F1. It seems that visual and textual information are complementary, and on their own are not rich enough to represent the semantics of the images. MixLDA achieves this to a greater extent as it uses a concatenated representation of words and visual features d_{Mix}. It thus assumes that the two modalities have equal importance in defining the latent space, which, as our results suggest, is beneficial.

The continuous relevance model (ContRel) improves considerably upon TxtLDA, ImgLDA, CorrLDA, and PLSA, but is significantly worse ($p < 0.01$) than MixLDA. On the surface, MixLDA seems similar to ContRel; both models take advantage of visual and textual information. ContRel smooths the conditional probability of a word given an image with the conditional probability of the same word given the document and uses an LDA model (trained on the news document collection) to remove nontopical keywords from the model's output. MixLDA is conceptually simpler, but, importantly, LDA is the actual model rather than a postprocessing step, thus exploiting the synergy between visual and textual information more directly. Topics are created based on both modalities, which are treated on an equal footing. Compared to ContRel, MixLDA improves precision by 1.6 percent, recall by 5.2 percent, and the overall F1 by 1.8 percent.

All variants of PLSA significantly ($p < 0.01$) improve upon TxtLDA and ImgLDA in terms of F1. Models trained on image-caption pairs tend to perform worse compared to those additionally using the accompanying document. Note that F1 variations in performance among the different PLSA models are small. Interestingly, we observe the opposite when considering mAP. PLSA-words trained on image-caption tuples is by far the best model. A similar result is reported in Monay and Gatica-Perez [22], who use mAP to evaluate their PLSA models on the Corel dataset. CorrLDA performs significantly ($p < 0.01$) worse than PLSA, TxtLDA, and ImgLDA using both F1 and mAP.

TABLE 2
Image Annotations Generated by TxtLDA, ImgLDA, PLSA-Words, and MixLDA

TxtLDA	come, **king**, man, family, royal, crown, ground, leave, join, new
ImgLDA	carry, man, **die**, arrest, break, rule, include, nation, face, public
PLSA-Words	**die**, flower, leave, mount, **king**, capital, pressure, mark, official, come
MixLDA	**king**, **die**, Tuesday, succeed, man, pressure, public, minister, carry, **week**
TxtLDA	case, agency, milk, service, health, firm, product, report, spokesman, outbreak
ImgLDA	cause, work, national, public, report, drop, company, spokesman, follow, eat
PLSA-Words	operation, wales, **testing**, public, remain, problem, follow, protection, **level**, blame
MixLDA	agency, milk, work, bar, product, service, health, measure, recall, brand
TxtLDA	**ice**, change, rise, cover, use, datum, sea, **satellite**, global, research
ImgLDA	look, open, term, **ice**, **satellite**, day, project, December, late, **arctic**
PLSA-Words	**old**, sign, remain, west, melt, time, rise, study, space, **satellite**
MixLDA	**ice**, **arctic**, time, change, cover, extent, summer, wind, tell, area
TxtLDA	**child**, home, survey, use, new, understand, stay, want, know, **parent**
ImgLDA	**child**, access, suggest, good, **parent**, home, young, technology, risk, family
PLSA-Words	**child**, **parent**, game, mobile, gap, want, know, prevent, charity, need
MixLDA	**child**, **parent**, technology, use, survey, family, new, drive, know, mobile

Words in boldface indicate an exact match with the gold standard caption shown in Fig. 1.

Recall that in CorrLDA word topic assignments are drawn from the image regions which are in turn drawn from a Gaussian distribution. Although this modeling choice delivers better results on the simpler Corel dataset, it does not seem able to capture the characteristics of our images which are noisier and more complex. Moreover, CorrLDA assumes that annotation keywords must correspond to image regions. This assumption is too restrictive in our setting, where a single keyword may refer to many objects or people in an image (e.g., the word *badminton* is used to collectively describe an image depicting *players*, *shuttlecocks*, and *rackets*).

In sum, we observe that the proposed annotation model (MixLDA) is robust to inherent noise and improves upon competitive image annotation approaches. Table 2 shows examples of keywords generated by TxtLDA, ImgLDA, PLSA-Words, and MixLDA for the images in Fig. 1 (the first row corresponds to Fig. 1a, the second row to Fig. 1b, and so on).

5.2 Image Caption Generation

5.2.1 Data

Our caption generation experiments were conducted on the same BBC News dataset used for image annotation and using the same training, development, and test set partitions. In addition, documents and captions were parsed with the

Stanford parser [61] in order to obtain dependencies for the phrase-based abstractive model.

5.2.2 Parameter Tuning

Both extractive and abstractive generation models used MixLDA for content selection with the set of parameters found optimal on the image annotation task (750 visual terms, 1,000 topics). The abstractive models rely on a trigram model for creating coherent output which we trained using the SRI language modeling toolkit[9] on a newswire corpus consisting of BBC and Yahoo! news documents (6.9M words). The attachment probabilities (see (24)) were estimated from the same corpus. We tuned the caption length parameter on the development set using a range of $[5, 14]$ tokens for the word-based model and $[2, 5]$ phrases for the phrase-based model. Following Banko et al. [32], we approximated the length distribution with Gaussian distribution. The scaling parameter β for the adaptive language model was also tuned on the development set using a range of $[0.5, 0.9]$. We report results with β set to 0.5.

For the abstractive models the beam size was set to 500 (with at least 50 states for the word-based model). For the phrase-based model, we also experimented with reducing the search scope, either by considering only the n most similar sentences to the keywords (range $[2, 10]$), or simply the single most similar sentence and its neighbors. The former method delivered better results with five sentences (and the KL divergence similarity function).

5.2.3 Evaluation Method

We evaluated the performance of our captions automatically and also by eliciting human judgments. Our automatic evaluation was based on Translation Edit Rate (TER, [62]), a measure commonly used to evaluate the quality of machine translation output. TER is defined as the minimum number of edits a human would have to perform to change the system output so that it exactly matches a reference translation. In our case, the original captions written by the BBC journalists were used as reference:

$$\mathrm{TER}(E, E_r) = \frac{\mathrm{Ins} + \mathrm{Del} + \mathrm{Sub} + \mathrm{Shft}}{N_r}, \qquad (26)$$

where E is the hypothetical system output, E_r the reference caption, and N_r the reference length. The number of possible edits include insertions (Ins), deletions (Del), substitutions (Sub), and shifts (Shft). TER is similar to word error rate, the only difference being that it allows shifts. A shift moves a contiguous sequence to a different location within the same system output and is counted as a single edit. The perfect TER score is 0; however, note that it can be higher than 1 due to insertions. The minimum translation edit alignment is usually found through beam search. We used TER to compare the output of our extractive and abstractive models with the original captions and also for parameter tuning (see the discussion above).

In our human evaluation study, participants were presented with a document, an associated image, and its caption, and asked to rate the latter on two dimensions: grammaticality (Is the sentence fluent or word salad?) and relevance (Does it succinctly describe the content of the image and document?). We used a seven-point rating scale; participants were encouraged to give high ratings to captions that were grammatical and appropriate descriptions of the image given the accompanying document. Our experimental instructions are given in Appendix B, which is available in the online supplemental material. We randomly selected 12 document-image pairs from the test set and generated captions for them using the best extractive system (based on KL-divergence) and two abstractive systems (word based and phrase based). We also included the original human-authored caption as an upper bound. We collected ratings from 23 unpaid volunteers, all self-reported native English speakers. The study was conducted over the Internet using the WebExp [41] experimental software.

TABLE 3
TER Results for Extractive, Abstractive Models, and Lead Sentence Baseline

Model	TER	AvgLen
Lead sentence	$2.12^{\dagger\triangleright}$	21.0
Word Overlap	$2.46^{*\dagger\triangleright}$	24.3
Cosine	$2.26^{\dagger\triangleright}$	22.0
KL Divergence	1.77^{*}	18.4
JS Divergence	1.77^{*}	18.6
Abstract Words	$1.11^{*\dagger\triangleright}$	10.0
Abstract Phrases	$1.06^{*\dagger\triangleright}$	10.1

*: significantly different from lead sentence baseline; †: significantly different from KL divergence; ▷: significantly different from JS

5.2.4 Results

Table 3 reports our results on the test set using TER. We compare four extractive models based on word overlap, cosine similarity, and two probabilistic similarity measures, namely, KL and JS divergence, and two abstractive models based on words (see (18)) and phrases (see (25)). We also include a simple baseline that selects the first document sentence as a caption and show the average caption length (AvgLen) for each model. We examined whether performance differences among models are statistically significant, using the Wilcoxon test.

As can be seen, the probabilistic extractive models (KL and JS divergence) outperform word overlap and cosine similarity (all differences are statistically significant, $p < 0.01$).[10] They make use of the same topic model as the image annotation model, and are thus able to select sentences that cover common content. They are also significantly better than the lead sentence, which is a competitive baseline. It is well known that news articles are written so that the lead contains the most important information in a story.[11] This is an encouraging result as it highlights the importance of the visual information for the caption generation task. In general, word overlap is the worst performing model, which is not unexpected as it does not take any lexical variation into account. Cosine is slightly better but not significantly different from the lead sentence. The abstractive models obtain the best TER scores overall;

9. http://www.speech.sri.com/projects/srilm/.

10. We also note that mean length differences are not significant among these models.

11. As a rule of thumb, the lead should answer most or all of the five Ws (who, what, when, where, why).

TABLE 4
Mean Ratings on Caption Output Elicited by Humans

Model	Grammaticality	Relevance
KL Divergence	$6.42^{*\dagger}$	$4.10^{*\dagger}$
Abstract Words	$2.08^{\dagger}$	$3.20^{\dagger}$
Abstract Phrases	4.80^{*}	4.96^{*}
Gold Standard	$6.39^{*\dagger}$	5.55^{*}

*: *significantly different from the word-based abstractive system*, †: *significantly different from the phrase-based abstractive system.*

however, they generate shorter captions in comparison to the other models (closer to the length of the gold standard) and as a result TER treats them favorably simply because the number of edits is less. For this reason, we turn to the results of our judgment elicitation study, which assesses in more detail the quality of the generated captions.

Recall that participants judge the system output on two dimensions, grammaticality and relevance. Table 4 reports mean ratings for the output of the extractive system (based on the KL divergence), the two abstractive systems, and the human-authored gold standard caption. We performed an Analysis of Variance (ANOVA) to examine the effect of system type on the generation task. Post-hot Tukey tests were carried out on the mean of the ratings shown in Table 4 (for grammaticality and relevance).

The word-based system yields the least grammatical output. It is significantly worse than the phrase-based abstractive system ($\alpha < 0.01$), the extractive system ($\alpha < 0.01$), and the gold standard ($\alpha < 0.01$). Unsurprisingly, the phrase-based system is significantly less grammatical than the gold standard and the extractive system, whereas the latter is perceived as equally grammatical as the gold standard (the difference in the means is not significant). With regard to relevance, the word-based system is significantly worse than the phrase-based system, the extractive system, and the gold standard. Interestingly, the phrase-based system performs on the same level as the human gold standard (the difference in the means is not significant) and significantly better than the extractive system. Overall, the captions generated by the phrase-based system capture almost the same content as the human-authored captions, even though they tend to be less grammatical. Table 5 shows examples of system output for the image-document pairs in Fig. 1 (the first row corresponds to Fig. 1a, the second row to Fig. 1b, and so on). More detailed analysis and examples of system output are provided in Appendix C, which is available in the online supplemental material.

TABLE 5
Captions Written by Humans (G) and Generated by Our Systems

KL	Last year, thousands of Tongans took part in unprecedented demonstrations to demand greater democracy and public ownership of key national assets.
A_W	King Toupou IV died at the age of Tongans last week.
A_P	King Toupou IV died at the age of 88 last week.
G	King Tupou, who was 88, died a week ago.
KL	Contaminated Cadbury's chocolate was the most likely cause of an outbreak of salmonella poisoning, the Health Protection Agency has said.
A_W	Purely dairy milk buttons Easter had agreed to work has caused.
A_P	The 105g dairy milk buttons Easter egg affected by the recall.
G	Cadbury will increase its contamination testing levels.
KL	So a planet with less ice warms faster, potentially turning the projected impacts of global warming into reality sooner than anticipated.
A_W	Dr less winds through ice cover all over long time when.
A_P	The area of the Arctic covered in Arctic sea ice cover.
G	Satellite instruments can distinguish "old" Arctic ice from "new".
KL	That's where parents come in.
A_W	The survey found a third of children are about mobile phones.
A_P	The survey found a third of children in the driving seat.
G	Children were found to be far more internet-wise than parents.

KL: Extractive Model Based on KL-Divergence A_W: word-based abstractive model; A_P phrase-based abstractive model.

6 Conclusions

In this paper, we introduced the novel task of automatic caption generation for news images. The task fuses insights from computer vision and natural language processing and holds promise for various multimedia applications, such as image and video retrieval, development of tools supporting news media management, and for individuals with visual impairment. As a departure from previous work, we have approached this task in a knowledge-lean fashion by leveraging the vast resource of images available on the Internet and exploiting the fact that many of these co-occur with textual information (i.e., captions and associated documents). Our results show that it is possible to learn a caption generation model from weakly labeled data without costly manual involvement.

The dataset we employed contains real-world images and exhibits a large vocabulary including both concrete object names and abstract keywords; instead of manually creating annotations, image captions are treated as labels for the image. Although the caption words are admittedly noisy compared to traditional human-created keywords, we show that they can be used to learn the correspondences between visual and textual modalities, and also serve as a gold standard for the caption generation task. Moreover, this news dataset contains a unique component, the news document, which provides both information regarding to the image's content and rich linguistic information required for the generation procedure.

Inspired by recent work in summarization, we have presented extractive and abstractive caption generation models. A key aspect of our approach is to allow both the visual and textual modalities to influence the generation task. This is achieved through an image annotation model that characterizes pictures in terms of description keywords that are subsequently used to guide the caption generation process. Our results show that the visual information plays an important role in content selection. Simply extracting a sentence from the document often yields an inferior caption. Our experiments also show that a probabilistic abstractive model defined over phrases yields promising results. It generates captions that are more grammatical than a closely related word-based system and manages to capture the gist of the image (and document) as well as the captions written by journalists.

We have primarily explored the feasibility of caption generation in the news domain. However, the proposed

framework can be applied to other types of data, including photo sharing sites and life-science publications, which conventionally contain graphical illustrations with detailed textual descriptions [63]. The uses of the image annotation model discussed in this paper are many and varied. An interesting future direction concerns the application of the proposed model in a semi-supervised setting where the annotation output is iteratively refined with some manual intervention [64]. We also believe that the annotation model can be usefully employed in an information retrieval setting where the goal is to find the image most relevant for a given query or document.

The model presented here could be further improved in several ways. First, we could allow an infinite number of topics and develop a nonparametric version that learns how many topics are optimal. Second, our model is based on word unigrams, and in this sense takes very little linguistic knowledge into account. Recent developments in topic modeling could potentially rectify this, e.g., by assuming that each word is generated by a distribution that combines document-specific topics and parse-tree specific syntactic transitions [65]. Third, our model considers mostly local features for representing the images. A better representation would also take global feature dependencies into account (e.g., the spatial relationships among different image regions).

Our caption generation model adopts a two-stage approach where the image processing and surface realization are carried out sequentially. A more general model should integrate the two steps in a unified framework. Indeed, an avenue for future work would be to define a phrase-based model for both image annotation and caption generation, e.g., by exploiting recent work in detecting visual phrases (e.g., [66]). We also believe that our approach would benefit from more detailed linguistic and nonlinguistic information. For instance, we could experiment with features related to document structure such as titles, headings, and sections of articles, and also exploit syntactic information more directly. The latter is currently used in the phrase-based model by taking attachment probabilities into account. We could, however, improve grammaticality more globally by generating a well-formed tree (or dependency graph).

REFERENCES

[1] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Image Classification for Content-Based Indexing," *IEEE Trans. Image Processing,* vol. 10, no. 1, pp. 117-130, 2001.

[2] A.W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

[3] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," *Proc. Seventh European Conf. Computer Vision,* pp. 97-112, 2002.

[4] D. Blei, "Probabilistic Models of Text and Images," PhD dissertation, Univ. of Massachusetts, Amherst, Sept. 2004.

[5] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan, "Matching Words and Pictures," *J. Machine Learning Research,* vol. 3, pp. 1107-1135, 2002.

[6] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous Image Classification and Annotation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1903-1910, 2009.

[7] V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the Semantics of Pictures," *Proc. 16th Conf. Advances in Neural Information Processing Systems,* 2003.

[8] S. Feng, V. Lavrenko, and R. Manmatha, "Multiple Bernoulli Relevance Models for Image and Video Annotation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1002-1009, 2004.

[9] L. Ferres, A. Parush, S. Roberts, and G. Lindgaard, "Helping People with Visual Impairments Gain Access to Graphical Information through Natural Language: The *igraph* System," *Proc. 11th Int'l Conf. Computers Helping People with Special Needs,* pp. 1122-1130, 2006.

[10] A. Abella, J.R. Kender, and J. Starren, "Description Generation of Abnormal Densities Found in Radiographs," *Proc. Symp. Computer Applications in Medical Care,* Am. Medical Informatics Assoc., pp. 542-546, 1995.

[11] A. Kojima, T. Tamura, and K. Fukunaga, "Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions," *Int'l J. Computer Vision,* vol. 50, no. 2, pp. 171-184, 2002.

[12] A. Kojima, M. Takaya, S. Aoki, T. Miyamoto, and K. Fukunaga, "Recognition and Textual Description of Human Activities by Mobile Robot," *Proc. Third Int'l Conf. Innovative Computing Information and Control,* pp. 53-56, 2008.

[13] P. Héde, P.A. Moëllic, J. Bourgeoys, M. Joint, and C. Thomas, "Automatic Generation of Natural Language Descriptions for Images," *Proc. Recherche d'Information Assistée par Ordinateur,* 2004.

[14] B. Yao, X. Yang, L. Lin, M.W. Lee, and S. Chun Zhu, "I2T: Image Parsing to Text Description," *Proc. IEEE,* vol. 98, no. 8, pp. 1485-1508, 2009.

[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L. Berg, "Baby Talk: Understanding and Generating Image Descriptions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1601-1608, 2011.

[16] A. Farhadi, M. Hejrati, A. Sadeghi, P. Yong, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every Picture Tells a Story: Generating Sentences from Images," *Proc. 11th European Conf. Computer Vision,* pp. 15-29, 2010.

[17] V. Ordonez, G. Kulkarni, and T.L. Berg, "Im2Text: Describing Images Using 1 Million Captioned Photographs," *Advances in Neural Information Processing Systems,* vol. 24, pp. 1143-1151, 2011.

[18] A. Makadia, V. Pavlovic, and S. Kumar, "Baselines for Image Annotation," *Int'l J. Computer Vision,* vol. 90, no. 1, pp. 88-105, 2010.

[19] C.-F. Chai and C. Hung, "Automatically Annotating Images with Keywords: A Review of Image Annotation Systems," *Recent Patents on Computer Science,* vol. 1, pp. 55-68, 2008.

[20] J.-Y. Pan, H.-J. Yang, and C. Faloutsos, "MMSS: Multi-Modal Story-Oriented Video Summarization," *Proc. Fourth IEEE Conf. Data Mining,* pp. 491-494, 2004.

[21] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning,* vol. 41, no. 2, pp. 177-196, 2001.

[22] F. Monay and D. Gatica-Perez, "Modeling Semantic Aspects for Cross-Media Image Indexing," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 10, pp. 1802-1817, Oct. 2007.

[23] T.L. Berg, A.C. Berg, J. Edwards, and D. Forsyth, "Who's in the Picture," *Advances in Neural Information Processing Systems,* vol. 17, pp. 137-144, 2005.

[24] M. Özcan, L. Jie, V. Ferrari, and B. Caputo, "A Large-Scale Database of Images and Captions for Automatic Face Naming," *Proc. British Machine Vision Conf.,* pp. 1-11, 2011.

[25] J. Luo, B. Caputo, and V. Ferrari, "Who's Doing What: Joint Modeling of Names and Verbs for Simultaneous Face and Pose Annotation," *Advances in Neural Information Processing Systems,* vol. 22, pp. 1168-1176, 2009.

[26] J. Wang, K. Markert, and M. Everingham, "Learning Models for Object Recognition from Natural Language Descriptions," *Proc. British Machine Vision Conf.,* 2009.

[27] S. Ju Hwang and K. Grauman, "Learning the Relative Importance of Objects from Tagged Images for Retrieval and Cross-Modal Search," *Int'l J. Computer Vision,* pp. 1-20, 2011.

[28] V.O. Mittal, J.D. Moore, G. Carenini, and S. Roth, "Describing Complex Charts in Natural Language: A Caption Generation System," *Computational Linguistics,* vol. 24, pp. 431-468, 1998.

[29] M. Corio and G. Lapalme, "Generation of Texts for Information Graphics," *Proc. Seventh European Workshop Natural Language Generation,* pp. 49-58, 1999.

[30] S. Elzer, S. Carberry, I. Zukerman, D. Chester, N. Green, and S. Demir, "A Probabilistic Framework for Recognizing Intention in Information Graphics," *Proc. 19th Int'l Conf. Artificial Intelligence,* pp. 1042-1047, 2005.
[31] A. Aker and R. Gaizauskas, "Generating Image Descriptions Using Dependency Relational Patterns," *Proc. 48th Ann. Meeting Assoc. for Computational Linguistics,* pp. 1250-1258, 2010.
[32] M. Banko, V. Mittal, and M. Witbrock, "Headline Generation Based on Statistical Translation," *Proc. 38th Ann. Meeting Assoc. for Computational Linguistics,* pp. 318-325, 2000.
[33] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," *Proc. Eighth IEEE Int'l Conf. Computer Vision,* pp. 416-423, 2001.
[34] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," *Proc. Workshop Generative-Model Based Vision,* pp. 59-70, 2004.
[35] G. Griffin, A. Holub, and P. Perona, "Caltech 256 Object Category Data Set," Technical Report 7694, California Inst. of Technology, http://authors.library.caltech.edu/7694, 2007.
[36] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting Image Databases from the Web," *Proc. 11th IEEE Int'l Conf. Computer Vision,* pp. 1-8, 2007.
[37] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," *Int'l J. Computer Vision,* vol. 77, nos. 1-3, pp. 157-173, 2008.
[38] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, and J. Kaufhold, "Evaluation of Localized Semantics: Data, Methodology, and Experiments," *Int'l J. Computer Vision,* vol. 77, nos. 1-3, pp. 199-217, 2008.
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 248-255, 2009.
[40] M. Hodosh, P. Young, C. Rashtchian, and J. Hockenmaier, "Cross-Caption Coreference Resolution for Automatic Image Understanding," *Proc. 14th Conf. Computational Natural Language Learning,* pp. 162-171, 2010.
[41] F. Keller, S. Gunasekharan, N. Mayo, and M. Corley, "Timing Accuracy of Web Experiments: A Case Study Using the WebExp Software Package," *Behavior Research Methods,* vol. 41, no. 1, pp. 1-12, 2009.
[42] D. Blei and M. Jordan, "Modeling Annotated Data," *Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* pp. 127-134, 2003.
[43] D. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1150-1157, 1999.
[44] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.
[45] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 257-263, 2003.
[46] A. Bosch, "Image Classification for a Large Number of Object Categories," PhD dissertation, Universitat de Girona, Sept. 2007.
[47] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.
[48] K. Sparck Jones, "Automatic Summarizing: Factors and Directions," *Advances in Automatic Text Summarization,* I. Mani and M.T. Maybury, eds., pp. 1-33, MIT Press, 1999.
[49] I. Mani, *Automatic Summarization.* John Benjamins Publishing Co., 2001.
[50] G. Salton and M. McGill, *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.
[51] M. Steyvers and T. Griffiths, "Probabilistic Topic Models," *A Handbook of Latent Semantic Analysis,* T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, eds. Psychology Press, 2007.
[52] M. Witbrock and V. Mittal, "Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries," *Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* pp. 315-316, 1999.
[53] R. Kneser, J. Peters, and D. Klakow, "Language Model Adaptation Using Dynamic Marginals," *Proc. Fifth European Conf. Speech Comm. and Technology,* vol. 4, pp. 1971-1974, 1997.
[54] L. Zhou and E. Hovy, "Headline Summarization at ISI," *Proc. HLT-NAACL Text Summarization Workshop and Document Understanding Conf.,* pp. 174-178, 2003.
[55] R. Soricut and D. Marcu, "Stochastic Language Generation Using WIDL-Expressions and Its Application in Machine Translation and Summarization," *Proc. 21st Int'l Conf. Computational Linguistics and the 44th Ann. Meeting Assoc. for Computational Linguistics,* pp. 1105-1112, 2006.
[56] S. Wan, R. Dale, M. Dras, and C. Paris, "Statistically Generated Summary Sentences: A Preliminary Evaluation of Verisimilitude Using Precision of Dependency Relations," *Proc. Workshop Using Corpora for Natural Language Generation,* 2005.
[57] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," *Proc. Int'l Conf. New Methods in Language Processing,* 1994.
[58] Y. Feng and M. Lapata, "Automatic Image Annotation Using Auxiliary Text Information," *Proc. 46th Ann. Meeting Assoc. of Computational Linguistics: Human Language Technologies,* pp. 272-280, 2008.
[59] C. Buckley and E.M. Voorhees, "Retrieval System Evaluation," *TREC: Experiment and Evaluation in Information Retrieval,* E.M. Voorhees and D.K. Harman, eds., pp. 53-78, MIT Press, 2005.
[60] E.W. Noreen, *Computer-Intensive Methods for Testing Hypotheses: An Introduction.* John Wiley & Sons, Inc., 1989.
[61] D. Klein and C.D. Manning, "Accurate Unlexicalized Parsing," *Proc. 41st Ann. Meeting Assoc. of Computational Linguistics,* pp. 423-430, 2003.
[62] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," *Proc. Seventh Conf. Assoc. for Machine Translation in the Americas,* pp. 223-231, 2006.
[63] A. Ahmed, E.P. Xing, W.W. Cohen, and R.F. Murphy, "Structured Correspondence Topic Models for Mining Captioned Figures in Biological Literature," *Proc. ACM SIGKDD 15th Int'l Conf. Knowledge Discovery and Data Mining,* pp. 39-48, 2009.
[64] R. Socher and L. Fei-Fei, "Connecting Modalities: Semi-Supervised Segmentation and Annotation of Images Using Unaligned Text Corpora," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 966-973, 2010.
[65] J. Boyd-Graber and D. Blei, "Syntactic Topic Models," *Proc. 22nd Conf. Advances in Neural Information Processing Systems,* 2009.
[66] A. Sadeghi and A. Farhadi, "Recognition Using Visual Phrases," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1745-1752, 2011.

Yansong Feng is a lecturer at the Institute of Computer Science and Technology of Peking University. His research interests include machine learning techniques for natural language processing and computer vision (e.g., information extraction, document summarization, multimedia analysis, and retrieval). He is a member of the IEEE.

Mirella Lapata is a professor in the School of Informatics, University of Edinburgh. Her research interests include probabilistic learning techniques for natural language understanding and generation. Examples include document simplification and summarization, discourse modeling, and word sense disambiguation. She is a member of the IEEE.