How Many Words is a Picture Worth?
Automatic Caption Generation for News Images

Yansong Feng and **Mirella Lapata**
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
Y.Feng-4@sms.ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

In this paper we tackle the problem of automatic caption generation for news images. Our approach leverages the vast resource of pictures available on the web and the fact that many of them are captioned. Inspired by recent work in summarization, we propose *extractive* and *abstractive* caption generation models. They both operate over the output of a probabilistic image annotation model that preprocesses the pictures and suggests keywords to describe their content. Experimental results show that an abstractive model defined over phrases is superior to extractive methods.

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 three billion images, with approximately 2.5 million images being uploaded every day.[1] Many on-line 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 is 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 meta-data (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 (Vailaya et al., 2001; Smeulders et al., 2000), instantiations of the noisy-channel model (Duygulu et al., 2002), latent variable models (Blei and Jordan, 2003; Barnard et al., 2002; Wang et al., 2009), and models inspired by information retrieval (Lavrenko et al., 2003; Feng et al., 2004).

In this paper we go one step further and generate captions for images rather than individual keywords. Although image indexing techniques based on keywords 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 under the blue sky*" would make the relations between the words explicit. Automatic caption generation could improve image retrieval by supporting longer and more targeted queries. It could also assist journalists in creating descriptions for the images associated with their articles. Beyond image retrieval, it could 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 (Ferres et al., 2006).

We explore the feasibility of automatic caption generation in the news domain, and create descriptions for images associated with on-line articles. Obtaining training data in this setting does not require expensive manual annotation as many articles are published together with captioned images. Inspired by recent work in summarization, we propose *extractive* and *abstractive* caption gen-

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

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1239–1249,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

eration models. The backbone for both approaches is a probabilistic image annotation model that suggests keywords for an image. We can then simply identify (and rank) the sentences in the documents that share these keywords or create a new caption that is potentially more concise but also informative and fluent. Our abstractive model 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. Experimental results show that the model's output compares favorably to handwritten 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 the interplay between visual and linguistic information. A handful of approaches generate 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. (2004) 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. Kojima et al. (2002, 2008) create descriptions for human activities in office scenes. They extract features of human motion 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. (2009) 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 multi-sentence description is generated using a document planner and a surface realizer.

Within natural language processing most previous efforts have focused on generating captions to accompany complex graphical presentations (Mittal et al., 1998; Corio and Lapalme, 1999; Fasciano and Lapalme, 2000; Feiner and McKeown, 1990) 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 (Elzer et al., 2005). Little emphasis is placed on image processing; 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.

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 on-line sources, and contains rich information about the entities and events depicted in the images and their relations. Similar 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. The caption generation task bears some resemblance to headline generation (Dorr et al., 2003; Banko et al., 2000; Jin and Hauptmann, 2002) where the aim is to create a very short summary for a document. Importantly, we aim to create a caption that not only summarizes the document but is also a 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 models that rely on visual information to drive the generation process. Our approach thus differs from most work in summarization which is solely text-based.

3 Problem Formulation

We formulate image caption generation as follows. Given an image I, and a related knowledge database κ, create a natural language description C which captures the main content of the image under κ. Specifically, in the news story scenario, we will generate a caption C for an image I and its accompanying document D. The training data thus consists of document-image-caption tu-

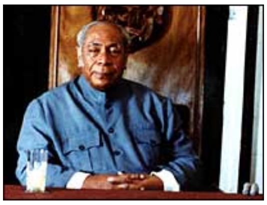
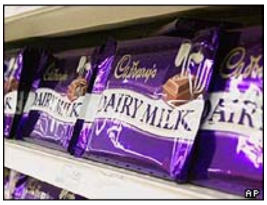
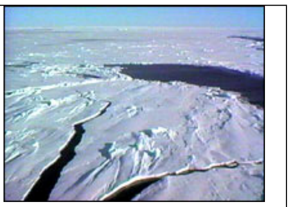
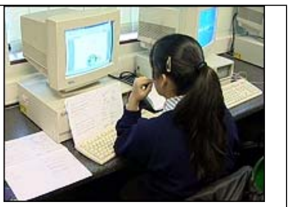
Thousands of Tongans have attended the funeral of King Taufa'ahau Tupou IV, who died last week at the age of 88. Representatives from 30 foreign countries watched as the king's coffin was carried by 1,000 men to the official royal burial ground.	 **King Tupou, who was 88, died a week ago.**	Contaminated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

Table 1: Each entry in the BBC News database contains a document an image, and its caption.

ples like the ones shown in Table 1. During testing, we are given a document and an associated image for which we must generate a caption.

Our experiments used the dataset created by Feng and Lapata (2008).[2] It contains 3,361 articles downloaded from the BBC News website[3] each of which is associated with a captioned news image. The latter is usually 203 pixels wide and 152 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% of the time contain words that are not attested in the document (even though they may be attested in the collection).

Generating image captions is a challenging task even for humans, let alone computers. Journalists are given explicit instructions on how to write captions[4] and laypersons do not always agree on what a picture depicts (von Ahn and Dabbish, 2004). Along with the title, the lead, and section headings, captions are the most commonly read words in an article. A good caption must be succinct and informative, clearly identify the subject of the picture, establish the picture's relevance to the article, provide context for the picture, and ultimately draw the reader into the article. It is also worth noting that journalists often write their own captions rather than simply extract sentences from the document. In doing so they rely on general world knowledge but also expertise in current affairs that goes beyond what is described in the article or shown in the picture.

4 Image Annotation

As mentioned earlier, our approach relies on an image annotation model to provide description keywords for the picture. Our experiments made use of the probabilistic model presented in Feng and Lapata (2010). The latter is well-suited to our task as it has been developed with noisy, multimodal data sets in mind. The model is based on the assumption that images and their surrounding text are generated by mixtures of latent topics which are inferred from a concatenated representation of words and visual features.

Specifically, images are preprocessed so that they are represented by word-like units. Local image descriptors are computed using the Scale Invariant Feature Transform (SIFT) algorithm (Lowe, 1999). The general idea behind the algorithm is to first sample an image with the difference-of-Gaussians point detector at different

[2] Available from http://homepages.inf.ed.ac.uk/s677528/data/

[3] http://news.bbc.co.uk/

[4] See http://www.theslot.com/captions.html and http://www.thenewsmanual.net/ for tips on how to write good captions.

scales and locations. Importantly, this detector is, to some extent, invariant to translation, scale, rotation and illumination changes. Each detected region is represented with a SIFT descriptor which is a histogram of edge directions at different locations. Subsequently SIFT descriptors are quantized into a discrete set of visual terms via a clustering algorithm such as K-means.

The model thus works with a bag-of-words representation and treats each article-image-caption tuple as a single document d_{Mix} consisting of textual and visual words. Latent Dirichlet Allocation (LDA, Blei et al. 2003) is used to infer the latent topics assumed to have generated d_{Mix}. The basic idea underlying LDA, and topic models in general, is that each document is composed of a probability distribution over topics, where each topic represents a probability distribution over words. The document-topic and topic-word distributions are learned automatically from the data and provide information about the semantic themes covered in each document and the words associated with each semantic theme. The image annotation model takes the topic distributions into account when finding the most likely keywords for an image and its associated document.

More formally, given an image-caption-document tuple (I,C,D) the model finds the subset of keywords W_I ($W_I \subseteq W$) which appropriately describe I. Assuming that keywords are conditionally independent, and I, D are represented jointly by d_{Mix}, the model estimates:

$$\begin{aligned} W_I^* &\approx \arg\max_{W_t} \prod_{w_t \in W_t} P(w_t|d_{Mix}) \qquad (1) \\ &= \arg\max_{W_t} \prod_{w_t \in W_t} \sum_{k=1}^{K} P(w_t|z_k)P(z_k|d_{Mix}) \end{aligned}$$

W_t denotes a set of description keywords (the subscript t is used to discriminate from the visual words which are not part of the model's output), K the number of topics, $P(w_t|z_k)$ the multimodal word distributions over topics, and $P(z_k|d_{Mix})$ the estimated posterior of the topic proportions over documents. Given an unseen image-document pair and trained multimodal word distributions over topics, it is possible to infer the posterior of topic proportions over the new data by maximizing the likelihood. The model delivers a ranked list of textual words w_t, the n-best of which are used as annotations for image I.

It is important to note that the caption generation models we propose are not especially tied to the above annotation model. Any probabilistic model with broadly similar properties could serve our purpose. Examples include PLSA-based approaches to image annotation (e.g., Monay and Gatica-Perez 2007) and correspondence LDA (Blei and Jordan, 2003).

5 Extractive Caption Generation

Much work in summarization to date focuses on sentence extraction where a summary is created 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. We discuss below different ways of operationalizing similarity.

Word Overlap Perhaps the simplest way of measuring the similarity between image keywords and document sentences is word overlap:

$$Overlap(W_I,S_d) = \frac{|W_I \cap S_d|}{|W_I \cup S_d|} \qquad (2)$$

where W_I is the set of keywords and S_d a sentence in the document. The caption is then the sentence that has the highest overlap with the keywords.

Cosine Similarity 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 (Salton and McGill, 1983). The latter is 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 appeared within the 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:

$$sim(\overrightarrow{W_I},\overrightarrow{S_d}) = \frac{\overrightarrow{W_I} \cdot \overrightarrow{S_d}}{|\overrightarrow{W_I}||\overrightarrow{S_d}|} \qquad (3)$$

Probabilistic Similarity Recall that the backbone of our image annotation model is a topic model with images and documents represented as a probability distribution over 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 (Steyvers and Griffiths, 2007). For example, we may use the KL divergence to measure the difference between the distributions p and q:

$$D(p,q) = \sum_{j=1}^{K} p_j \log_2 \frac{p_j}{q_j} \tag{4}$$

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

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:

$$JS(p,q) = \frac{1}{2}\left[D(p,\frac{(p+q)}{2}) + D(q,\frac{(p+q)}{2})\right] \tag{5}$$

6 Abstractive Caption Generation

Although extractive methods yield grammatical captions and require relatively little linguistic analysis, there are a few caveats to consider. Firstly, 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. Secondly, the selected sentences make for long captions (sometimes longer than the average document sentence), 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.

Word-based Model Our first abstractive model builds on and extends a well-known probabilistic model of headline generation (Banko et al., 2000). The task is related to caption generation, the aim is to create a short, title-like headline for a given document, without however taking visual information into account. Like captions, headlines have to be catchy to attract the reader's attention.

Banko et al. (2000) propose a bag-of-words model for headline generation. It consists of content selection and surface realization components. Content selection is modeled as the probability of a word appearing in the headline given the same word appearing in the corresponding document and is independent from 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 towards generating concise output:

$$\begin{aligned} P(w_1,w_2,...,w_n) = & \prod_{i=1}^{n} P(w_i \in H | w_i \in D) \\ & \cdot P(len(H) = n) \\ & \cdot \prod_{i=2}^{n} P(w_i | w_{i-1}) \end{aligned} \tag{6}$$

where w_i is a word that may appear in headline H, D the document being summarized, and $P(len(H) = n)$ a headline length distribution model.

The above model can be easily adapted to the 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 (see Section 4). In addition we replace the bigram surface realizer with a trigram:

$$\begin{aligned} P(w_1,w_2,...,w_n) = & \prod_{i=1}^{n} P(w_i \in C | I, D) \\ & \cdot P(len(C) = n) \\ & \cdot \prod_{i=3}^{n} P(w_i | w_{i-1}, w_{i-2}) \end{aligned} \tag{7}$$

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

Despite its simplicity, the caption generation model in (7) has a major drawback. The content selection component will naturally tend to ignore function words, as they are not descriptive of the image's content. This will seriously impact the grammaticality of the generated captions, as there will be no appropriate 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. At the same time we modify our surface realization component so that it takes note of the image annotation probabilities. Specifically, we use an *adaptive* language model (Kneser et al., 1997) that modifies an *n*-gram model with local unigram probabilities:

$$\begin{aligned} P(w_1, w_2, \ldots, w_n) = & \prod_{i=1}^{n} P(w_i \in C | w_i \in D) \quad (8) \\ & \cdot P(len(C) = n) \\ & \cdot \prod_{i=3}^{n} P_{adap}(w_i | w_{i-1}, w_{i-2}) \end{aligned}$$

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

$$P_{adap}(w|h) = \frac{\alpha(w)}{z(h)} P_{back}(w|h) \quad (9)$$

$$\alpha(w) \approx (\frac{P_{adap}(w)}{P_{back}(w)})^{\beta} \quad (10)$$

$$z(h) = \sum_{w} \alpha(w) \cdot P_{back}(w|h) \quad (11)$$

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

Phrase-based Model The model outlined in equation (8) will generate captions with function words. However, there is no guarantee that these will be compatible with their surrounding context or that the caption 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 can potentially capture long-range dependencies.

Specifically, we obtain phrases from the output of a dependency parser. A phrase is simply a head and its dependents with the exception of verbs, where we record only the head (otherwise, an entire sentence could be a phrase). For example, from the first sentence in Table 1 (first row, left document) we would extract the phrases: *thousands of Tongans*, *attended*, *the funeral*, *King Taufa'ahau Tupou IV*, *last week*, *at the age*, *died*, and so on. We only consider dependencies whose heads are nouns, verbs, and prepositions, as these constitute 80% 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 equation (8) as follows:

$$\begin{aligned} P(\rho_1, \rho_2, \ldots, \rho_m) \approx & \prod_{j=1}^{m} P(\rho_j \in C | \rho_j \in D) \quad (12) \\ & \cdot P(len(C) = \sum_{j=1}^{m} len(\rho_j)) \\ & \cdot \prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i | w_{i-1}, w_{i-2}) \end{aligned}$$

Here, $P(\rho_j \in C | \rho_j \in D)$ models the probability of phrase ρ_j appearing in the caption given that it also appears in the document and is estimated as:

$$P(\rho_j \in C | \rho_j \in D) = \prod_{w_j \in \rho_j} P(w_j \in C | w_j \in D) \quad (13)$$

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 *attachment* constraints into account by estimating the probability of phrase ρ_j attaching to the right of phrase ρ_i as:

$$\begin{aligned} P(\rho_j | \rho_i) = & \sum_{w_i \in \rho_i} \sum_{w_j \in \rho_j} p(w_j | w_i) \quad (14) \\ = & \frac{1}{2} \sum_{w_i \in \rho_i} \sum_{w_j \in \rho_j} \{ \frac{f(w_i, w_j)}{f(w_i, -)} + \frac{f(w_i, w_j)}{f(-, w_j)} \} \end{aligned}$$

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 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.[5]

After integrating the attachment probabilities into equation (12), the caption generation model becomes:

$$\begin{aligned} P(\rho_1, \rho_2, \ldots, \rho_m) \approx & \prod_{j=1}^{m} P(\rho_j \in C | \rho_j \in D) \quad (15) \\ & \cdot \prod_{j=2}^{m} P(\rho_j | \rho_{j-1}) \\ & \cdot P(len(C) = \textstyle\sum_{j=1}^{m} len(\rho_j)) \\ & \cdot \textstyle\prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i | w_{i-1}, w_{i-2}) \end{aligned}$$

[5] Equation (14) is smoothed to avoid zero probabilities.

On the one hand, the model in equation (15) takes long distance dependency constraints into account, and has some notion of syntactic structure through the use of attachment probabilities. On the other hand, it has a primitive notion of caption length estimated by $P(len(C) = \sum_{j=1}^{m} len(\rho_j))$ and will therefore generate captions of the same (phrase) length. Ideally, we would like the model to vary the length of its output depending on the chosen context. However, we leave this to future work.

Search To generate a caption it is necessary to find the sequence of words that maximizes $P(w_1, w_2, ..., w_n)$ for the word-based model (equation (8)) and $P(\rho_1, \rho_2, ..., \rho_m)$ for the phrase-based model (equation (15)). 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 D. Using one of the models from Section 5, we may rank its sentences in terms of their relevance to the image keywords 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.

7 Experimental Setup

In this section we discuss our experimental design for assessing the performance of 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.

Data All our experiments were conducted on the corpus created by Feng and Lapata (2008), following their original partition of the data (2,881 image-caption-document tuples for training, 240 tuples for development and 240 for testing). Documents and captions were parsed with the Stanford parser (Klein and Manning, 2003) in order to obtain dependencies for the phrase-based abstractive model.

Model Parameters For the image annotation model we extracted 150 (on average) SIFT features which were quantized into 750 visual terms. The underlying topic model was trained with 1,000 topics using only content words (i.e., nouns, verbs, and adjectives) that appeared no less than five times in the corpus. For all models discussed here (extractive and abstractive) we report results with the 15 best annotation keywords. For the abstractive models, we used a trigram model trained with the SRI toolkit on a newswire corpus consisting of BBC and Yahoo! news documents (6.9 M words). The attachment probabilities (see equation (14)) 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. (2000), we approximated the length distribution with a Gaussian. 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 (range $[2, 5]$). The former method delivered better results with 10 sentences (and the KL divergence similarity function).

Evaluation We evaluated the performance of our models automatically, and also by eliciting human judgments. Our automatic evaluation was based on Translation Edit Rate (TER, Snover et al. 2006), 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:

$$\mathrm{TER}(E, E_r) = \frac{\mathrm{Ins} + \mathrm{Del} + \mathrm{Sub} + \mathrm{Shft}}{N_r} \qquad (16)$$

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 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 align-

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

Table 2: TER results for extractive, abstractive models, and lead sentence baseline; *: sig. different from lead sentence; †: sig. different from KL and JS divergence.

ment is usually found through beam search. We used TER to compare the output of our extractive and abstractive models 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 describe succinctly the content of the image and document?). We used a 1–7 rating scale, participants were encouraged to give high ratings to captions that were grammatical and appropriate descriptions of the image given the accompanying document. We randomly selected 12 document-image pairs from the test set and generated captions for them using the best extractive system, 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.

8 Results

Table 2 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 equation (8)) and phrases (see equation (15)). 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.

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^{*}

Table 3: Mean ratings on caption output elicited by humans; *: sig. different from word-based abstractive system; †: sig. different from phrase-based abstractive system.

As can be seen the probabilistic models (KL and JS divergence) outperform word overlap and cosine similarity (all differences are statistically significant, $p < 0.01$).[6] 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.[7] 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, 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 3 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 3 (for grammaticality and relevance).

[6] We also note that mean length differences are not significant among these models.

[7] As a rule of thumb the lead should answer most or all of the five W's (who, what, when, where, why).

G:	King Tupou, who was 88, died a week ago.
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:	Cadbury will increase its contamination testing levels.
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:	Satellite instruments can distinguish "old" Arctic ice from "new".
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:	Children were found to be far more internet-wise than parents.
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.

Table 4: Captions written by humans (G) and generated by extractive (KL), word-based abstractive (A_W), and phrase-based extractive (A_P systems).

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 with 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 the same content as the human-authored captions, even though they tend to be less grammatical. Examples of system output for the image-document pairs shown in Table 1 are given in Table 4 (the first row corresponds to the left picture (top row) in Table 1, the second row to the right picture, and so on).

9 Conclusions

We have presented extractive and abstractive models that generate image captions for news articles. 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.

Future extensions are many and varied. Rather than adopting a two-stage approach, where the image processing and caption generation 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. We also believe that our approach would benefit from more detailed linguistic and non-linguistic 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

Banko, Michel, Vibhu O. Mittal, and Micheael J. Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Hong Kong, pages 318–325.

Barnard, Kobus, Pinar Duygulu, David Forsyth, Nando de Freitas, David Blei, and Michael Jordan. 2002. Matching words and pictures. *Journal of Machine Learning Research* 3:1107–1135.

Blei, David and Michael Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th An-*

nual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, ON, pages 127–134.

Blei, David, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Corio, Marc and Guy Lapalme. 1999. Generation of texts for information graphics. In *Proceedings of the 7th European Workshop on Natural Language Generation*. Toulouse, France, pages 49–58.

Dorr, Bonnie, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Summarization*. Edmonton, Canada, pages 1–8.

Duygulu, Pinar, Kobus Barnard, Nando de Freitas, and David Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision*. Copenhagen, Denmark, pages 97–112.

Elzer, Stephanie, Sandra Carberry, Ingrid Zukerman, Daniel Chester, Nancy Green, , and Seniz Demir. 2005. A probabilistic framework for recognizing intention in information graphics. In *Proceedings of the 19th International Conference on Artificial Intelligence*. Edinburgh, Scotland, pages 1042–1047.

Fasciano, Massimo and Guy Lapalme. 2000. Intentions in the coordinated generation of graphics and text from tabular data. *Knowledge Information Systems* 2(3):310–339.

Feiner, Steven and Kathleen McKeown. 1990. Coordinating text and graphics in explanation generation. In *Proceedings of National Conference on Artificial Intelligence*. Boston, MA, pages 442–449.

Feng, Shaolei Feng, Victor Lavrenko, and R Manmatha. 2004. Multiple Bernoulli relevance models for image and video annotation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. Washington, DC, pages 1002–1009.

Feng, Yansong and Mirella Lapata. 2008. Automatic image annotation using auxiliary text information. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies*. Columbus, OH, pages 272–280.

Feng, Yansong and Mirella Lapata. 2010. Topic models for image annotation and text illustration. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, LA.

Ferres, Leo, Avi Parush, Shelley Roberts, and Gitte Lindgaard. 2006. Helping people with visual impairments gain access to graphical information through natural language: The *graph* system. In *Proceedings of 11th International Conference on Computers Helping People with Special Needs*. Linz, Austria, pages 1122–1130.

Héde, Patrick, Pierre Allain Moëllic, Joël Bourgeoys, Magali Joint, and Corinne Thomas. 2004. Automatic generation of natural language descriptions for images. In *Proceedings of Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications Ordinateur) (RIAO)*. Avignon, France.

Jin, Rong and Alexander G. Hauptmann. 2002. A new probabilistic model for title generation. In *Proceedings of the 19th International Conference on Computational linguistics*. Taipei, Taiwan, pages 1–7.

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

Kneser, Reinhard, Jochen Peters, and Dietrich Klakow. 1997. Language model adaptation using dynamic marginals. In *Proceedings of 5th European Conference on Speech Communication and Technology*. Rhodes, Greece, volume 4, pages 1971–1974.

Kojima, Atsuhiro, Mamoru Takaya, Shigeki Aoki, Takao Miyamoto, and Kunio Fukunaga. 2008. Recognition and textual description of human activities by mobile robot. In *Proceedings of the 3rd International Conference on Innovative Computing Information and Control*. IEEE Computer Society, Washington, DC, pages 53–56.

Kojima, Atsuhiro, Takeshi Tamura, and Kunio Fukunaga. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision* 50(2):171–184.

Lavrenko, Victor, R. Manmatha, and Jiwoon Jeon. 2003. A model for learning the semantics of pictures. In *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems*. Vancouver, BC.

Lowe, David G. 1999. Object recognition from local scale-invariant features. In *Proceedings of International Conference on Computer Vision*. IEEE Computer Society, pages 1150–1157.

Mittal, Vibhu O., Johanna D. Moore, Giuseppe Carenini, and Steven Roth. 1998. Describing complex charts in natural language: A caption generation system. *Computational Linguistics* 24:431–468.

Monay, Florent and Daniel Gatica-Perez. 2007. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(10):1802–1817.

Salton, Gerard and M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Smeulders, Arnols W.M., Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12):1349–1380.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*. Cambridge, pages 223–231.

Steyvers, Mark and Tom Griffiths. 2007. Probabilistic topic models. In T. Landauer, D. McNamara, S Dennis, and W Kintsch, editors, *A Handbook of Latent Semantic Analysis*, Psychology Press.

Vailaya, Aditya, Mário A. T. Figueiredo, Anil K. Jain, and Hong-Jiang Zhang. 2001. Image classification for content-based indexing. *IEEE Transactions on Image Processing* 10:117–130.

von Ahn, Luis and Laura Dabbish. 2004. Labeling images with a computer game. In *ACM Conference on Human Factors in Computing Systems*. New York, NY, pages 319–326.

Wang, Chong, David Blei, and Li Fei-Fei. 2009. Simultaneous image classification and annotation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. Miami, FL, pages 1903–1910.

Yao, Benjamin, Xiong Yang, Liang Lin, Mun Wai Lee, and Song chun Zhu. 2009. I2t: Image parsing to text description. *Proceedings of IEEE (invited for the special issue on Internet Vision)* .