# Automatic Caption Generation for News Images

*Yansong Feng*

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2011

# Abstract

This thesis is concerned with the task of automatically generating captions for images, which is important for many image-related applications. Automatic description generation for video frames would help security authorities manage more efficiently and utilize large volumes of monitoring data. Image search engines could potentially benefit from image description in supporting more accurate and targeted queries for end users. Importantly, generating image descriptions would aid blind or partially sighted people who cannot access visual information in the same way as sighted people can. However, previous work has relied on fine-gained resources, manually created for specific domains and applications In this thesis, we explore the feasibility of automatic caption generation for news images in a knowledge-lean way. We depart from previous work, as we learn a model of caption generation from publicly available data that has not been explicitly labelled for our task. The model consists of two components, namely extracting image content and rendering it in natural language.

Specifically, we exploit data resources where images and their textual descriptions co-occur naturally. We present a new dataset consisting of news articles, images, and their captions that we required from the BBC News website. Rather than laboriously annotating images with keywords, we simply treat the captions as the labels. We show that it is possible to learn the visual and textual correspondence under such noisy conditions by extending an existing generative annotation model (Lavrenko et al., 2003). We also find that the accompanying news documents substantially complements the extraction of the image content. In order to provide a better modelling and representation of image content, We propose a probabilistic image annotation model that exploits the synergy between visual and textual modalities under the assumption that images and their textual descriptions are generated by a shared set of latent variables (topics). Using Latent Dirichlet Allocation (Blei and Jordan, 2003), we represent visual and textual modalities *jointly* as a probability distribution over a set of topics. Our model takes these topic distributions into account while finding the most likely keywords for an image and its associated document.

The availability of news documents in our dataset allows us to perform the caption generation task in a fashion akin to text summarization; save one important difference that our model is not solely based on text but uses the image in order to select content from the document that should be present in the caption. We propose both *extractive* and *abstractive* caption generation models to render the extracted image content in natural language without relying on rich knowledge resources, sentence-templates or grammars. The backbone for both approaches is our topic-based image annotation model. Our *extractive* models examine how to best select sentences that overlap in content with our image annotation model. We modify an existing *abstractive* headline generation model to our scenario by incorporating visual information. Our own model operates over image description keywords and document phrases by taking dependency and word order constraints into account. Experimental results show that both approaches can generate human-readable captions for news images. Our phrase-based abstractive model manages to yield as informative captions as those written by the BBC journalists.

# Acknowledgements

First and foremost, I would like to thank my supervisor, Mirella Lapata, for her valuable guidance and continuous support throughout my life and study in Edinburgh. Mirella's deep insight to science and exceptional patience encourages me to attempt various interesting topics and her skillful supervision helps me shape this thesis in the right direction. It is really my pleasure to work with her during the past years.

Many thanks to my second supervisor, Victor Lavrenko, for his helpful and valuable advice. I really benefit from the thoughtful discussions with Victor, especially topics about math and information retrieval. I am also grateful to Steve Renals and Miles Osborne who have served on my committee and provided insightful comments and advice.

I would like to thank all folks in the statNLP group: Frank, Sharon, Joel, Moreno, Ben, Tom... I really enjoy the interesting weekly discussions and especially, thank you for the very detailed, helpful comments and suggestions for my every talk.

I also want to thank my office mates and friends, from Buccleuch Place to Forum: Markus, Lexi, Hieu, Neil, Sharon, Katya, Jeff, Muhammad and Ioannis. Thank you for making our office such a comfortable and easy-going place. My thanks also go to Jenny, Avril, Heather and David, whose kind supports make my study much easier and smoother.

Many thanks to my Chinese colleagues and friends in Edinburgh: Xingkun, Songfang, Xionghu, Zhao Xu, Wang Zheng, Fei, Wang Dong, Liu Zhe, Guoting, Wensheng, Danyi, Ma Shuai, Yinghui, Wang Xin, Lin Kuang and Zhang Le, I quite enjoy the time with you guys in Edinburgh. I am also grateful to my friends in China: Yang Guang, Huge, Molan, Pu Ming, Chongde, and Xiaoguang who keep in touch, and provide an ear when needed.

Finally, thank you to my family, especially my dear parents for whom I should do more. Your love, constant support and encouragement makes me what I am today.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Yansong Feng*)

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This thesis is concerned with the task of automatically generating captions for news images. Although previous work has focused on generating descriptions for domain specific images, the task of caption generation is novel to our knowledge. Our caption generation model comprises two steps, namely content selection, which we operationalize as image annotation, and surface realization. In this chapter, we motivate why this work is important and discuss its potential for applications. We also present the main contributions and the structure of the thesis.

## 1.1 Image Description Generation

Computer vision and natural language processing (NLP) have been previously treated in isolation. The former usually deals with how to make machines *see* the world, while the latter mainly focuses on how to make machines *understand* human language. However, there is an increasing demand for bringing the two together. Consider video surveillance as an example. Most existing systems rely on manually created visual-textual correspondences to analyze video clips and then generate description sentences by filling predefined sentence-templates (Kojima et al., 2002). Due to their reliance on hand-crafted knowledge, most systems can only handle limited number of concepts (e.g., *turn*, *stop*, *enter*, *exit*, *person*, *car*, *truck*, etc) and scenes in specific domains (e.g., traffic and office monitoring). However, a smarter surveillance system could automatically inspect scenes of interest, handle more general concepts, summarize what it has *seen*, and then periodically *write* a human readable report. Generally, in this application, a computer system is expected to automatically extract the content of key frames from streams of video data and generate natural language descriptions expressing this content. This is an instantiation of the task of automatically generating natural language text for images.

This task is important for many reasons. An image description system can help people better manage the increasing volumes of multimedia data ranging from daily life entertainment (e.g., a sports video collection) to military security (e.g., images or videos of data collected from battle fields). Such a system would save much human labor, and provide people with easier access to large scale multimedia resources. Generating image descriptions would be of great importance for the blind or partially sighted people who cannot access visual information, such as pictures on the internet, in the same way as sighted people can. Besides web resources, graphs are widely used in many financial and stock news articles to supply better representations for numerical information. The detailed information conveyed from lines, charts, or plots is usually not fully presented in the article and can be easily understood by sighted users. An automatic image description generation model could be used to interpret the graph into several natural language sentences, indicating the shape of graphics, extreme points, changing trends of values and so on, and therefore allow visually impaired people to access as much information as sighted ones.

Besides helping people with special needs, an automatic image description generation component would also create more detailed and complete summaries for documents with multimedia representations. Current document summarization systems focus solely on textual information while ignoring pictures, graphical figures, or tables that are embedded in documents. These representations usually convey complementary information that is only implicitly described in the main text. Furthermore, these graphical representations could play an important role in determining what information is crucial for the document and should therefore be included in the summary. An image description generation module could help decide what to say in the summary and automatically render the missing information into natural language, thus enabling text summarization systems to produce more comprehensive summaries.

As far as image retrieval is concerned, automatic image description generation could help improve system accuracy and end-user experience. Although image indexing techniques based on keywords are popular and the method of choice for practical image retrieval engines, there are good reasons for using more linguistically meaningful descriptions. A list of isolated keywords is often ambiguous. An image annotated with the words "*car*, *blue*, *sky*" could depict a blue car or a blue sky, whereas the caption "*a car running under the blue sky*" makes the relations between the words explicit (e.g., *sky* is modified by *blue*, *car* is *under* the *sky*), and supplies richer underlying information usually absent from keyword lists, such as actions (e.g., *running*), who did what to whom, name entities and so on.

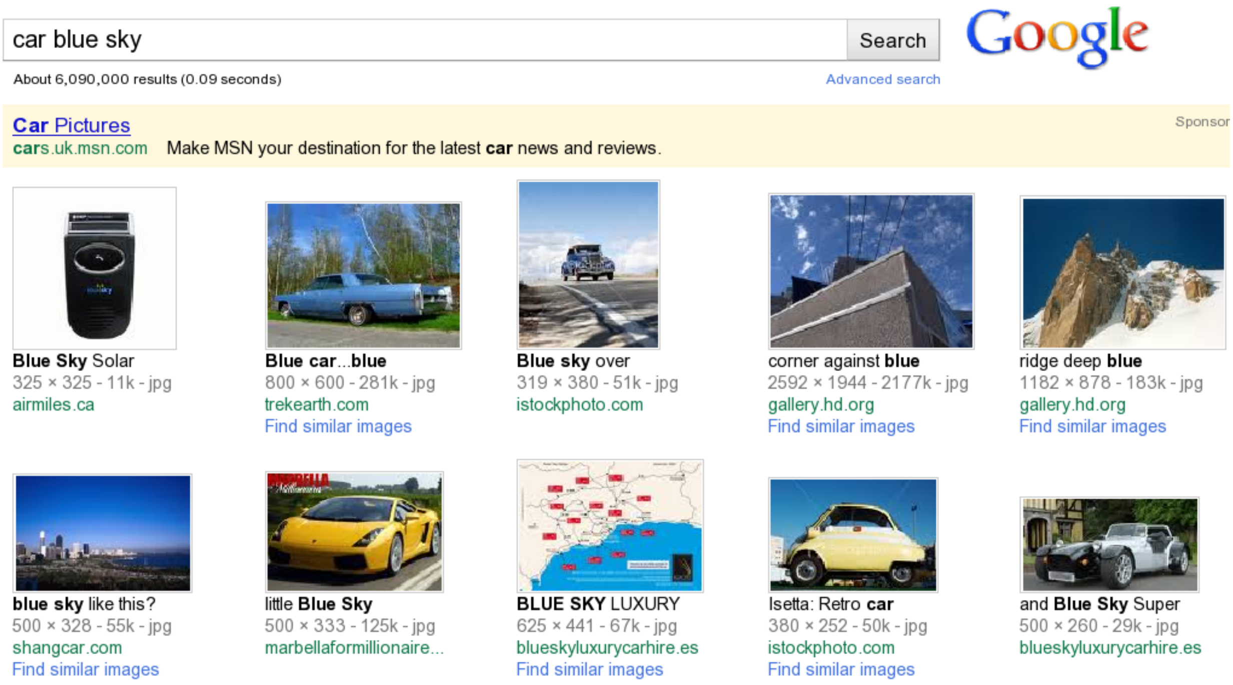
Figure 1.1 shows the output from Google Images[1], a popular image retrieval engine. Given a query, search engines usually retrieve relevant pictures by analyzing the image caption (if it exists), textual descriptions found adjacent to the image, and other text-related factors such as the file name of the image and clickthrough data (Weston et al., 2010). However, to our best knowledge, since they do not analyze the actual content of the images, search engines cannot be used to retrieve pictures from unannotated collections. As an example, we submitted the query "*car*, *blue*, *sky*" in the hope of finding pictures describing "*a car running under the blue sky*". The search engine returned the images shown in Figure 1.1, ranked based on their relevance to the query. Only three images capture the scene, "*a car under the blue sky*", while the rest are about *blue sky*, or just a *car*. The example illustrates that existing image retrieval engines could benefit from captioned image databases, which would provide a more natural and accurate search experience for end-users, e.g., by supporting longer and more targeted queries and enabling the use of question-answer interfaces.

An automatic image caption generation module could also assist journalists in creating descriptions for the news images or videos associated with their articles. Many on-line news sites like CNN, Reuters, and BBC publish images and videos with their stories and even provide photo feeds related to current events. Journalists and editors have to manually create captions for these images. The latter must be informative, clearly identify the subject of the pictures, provide context for them, establish their relevance to the news articles, or sometimes establish their relation with previous events. This task is difficult even for humans as it requires both general real-world knowledge and awareness of the specific news events being depicted. An automatic image description generation model can help produce sentences that describe the news image itself or relate the image to current or previous relevant news events. Journalists could then select suitable sentences from this output according to specific requirements.

Although the image description generation task is promising for many real-world applications, it has so far received little attention from both computer vision and natural language processing. Generally speaking, most previous work (see Chapter 2 for a detailed overview of related work) follows a two-step framework consisting of content selection and surface realization. The former step involves analyzing the image con-

---

[1] `www.google.com/images`



car blue sky

About 6,090,000 results (0.09 seconds)

**Car Pictures**
cars.uk.msn.com Make MSN your destination for the latest **car** news and reviews.

Figure 1.1: Output of Google Images for the query "*car*, *blue*, *sky*" (2010-08-07). Images are ranked based on their relevance to the query.

tent which is subsequently rendered into human readable sentences. A major obstacle here is the reliance on detailed and fine-gained world knowledge representing both the textual and visual modalities needed throughout the process. For instance, when extracting content from the images, most approaches assume that the correspondence between visual and textual information is known, i.e., that there exists a mapping between objects depicted in the image and their names. It is also common to rely on human written sentence templates or grammars to produce readable sentences. The reliance on manually created data largely limits the deployment of existing approaches to real-world applications.

Therefore, the main aim in this thesis is to develop a knowledge-lean approach[2] to automatically generating descriptions for images that requires minimal supervision and does not rely on manually created resources. Specifically, we will aim to generate captions for news images. News data is abundant and publicly available, although noisy. The task of generating captions for news images is novel to our knowledge, yet poses the same challenges with image description generation. In what follows, we

---

[2] We use the term *knowledge-lean* to refer to approaches that minimize the denpendence on fine-gained training data, external rules or knowledge sources (Pedersen and Bruce, 1998), such as predefined grammars or sentence templates.

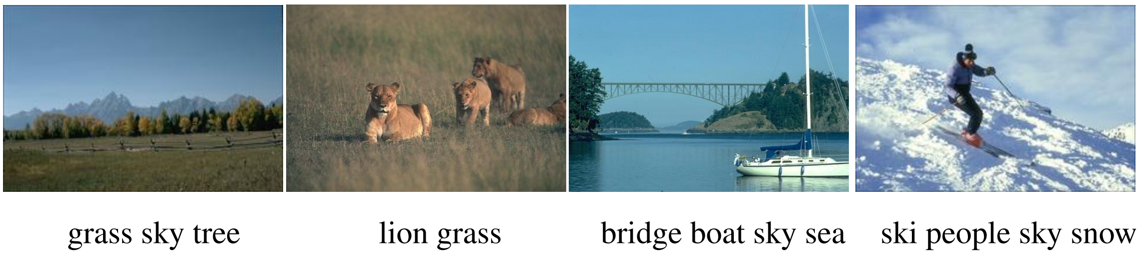
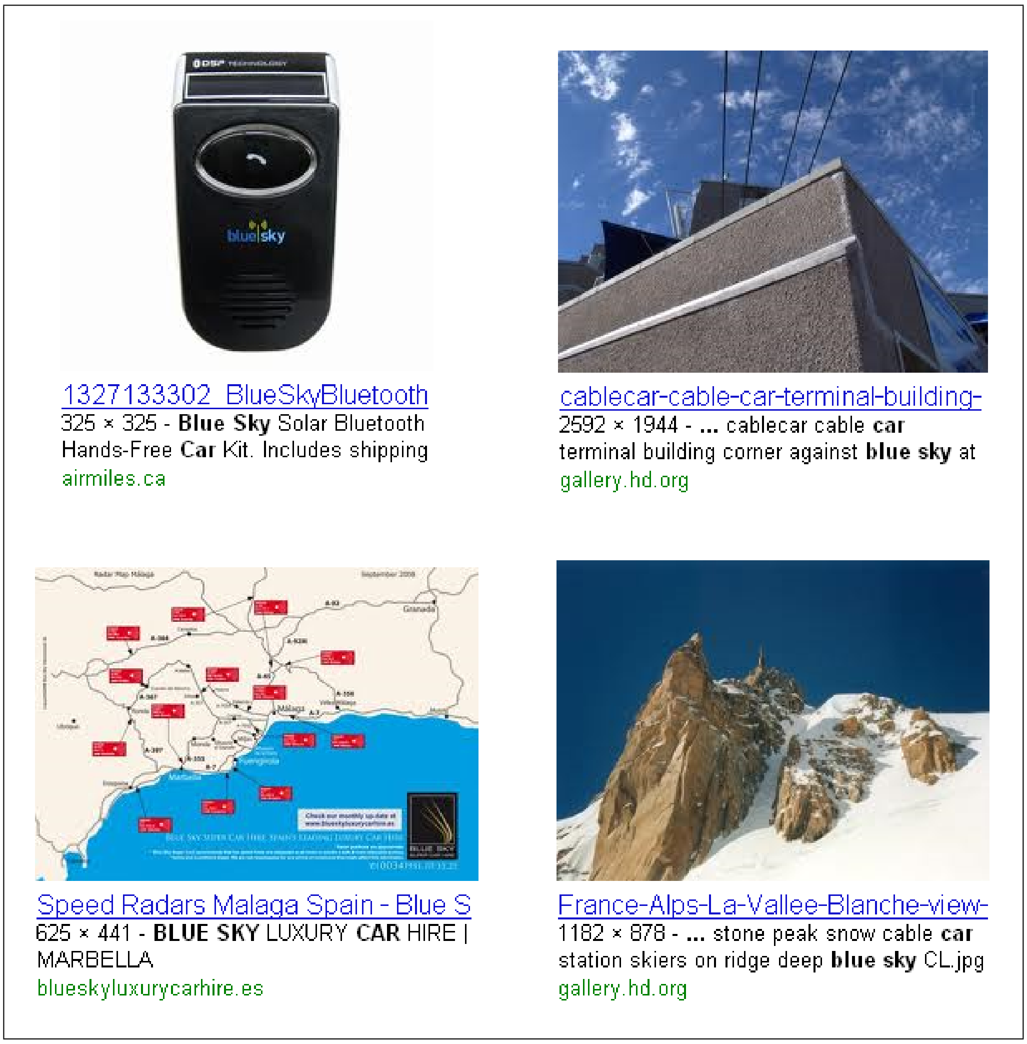


Figure 1.2: Images with abbreviated annotations from the Corel database, which has been frequently used as a test bed in automatic image annotation research.

outline these challenges and motivate the approach taken in this thesis.

**Extracting Image Content** The first challenge concerns identifying what the image is about (i.e., extracting its content). Given an image, an ideal image understanding system would reliably identify the depicted scene, its objects, which objects are important or prominent, and their relations etc. However, full image understanding is beyond the capability of current computer vision research. For example, most previous work (Kojima et al., 2002; Héde et al., 2004) adopts a knowledge-rich approach, where the cross-modality correspondence is made explicit through human annotation. More recent research has placed emphasis on a relatively simpler approach, namely automatic image annotation, which can be considered as an approximation of the full image understanding problem by addressing the main objects or events instead of every objects in the image (Datta et al., 2008) (see Figure 1.2). Given an image, a hypothecal image annotation system is expected to automatically label it with description keywords. This task, on its own, is of significant importance for many image-based applications, such as image retrieval, picture browsing support, and story picturing (Lavrenko et al., 2003; Jeon et al., 2003; Blei and Jordan, 2003; Joshi et al., 2006; Li and Wang, 2006). Especially, since manually annotating images for a large database is a labor intensive and time consuming task. In the long run, it can also be expensive since the work has to be repeated with every new collection.

In practice, existing image retrieval systems annotate their image databases mainly by analyzing image captions (if they exist), textual descriptions found adjacent to the images, and other text-related information such as the file name of the image, metadata of the image, or user click information. For example, consider the images and their surrounding text in Figure 1.3. The short description "*Blue Sky Solar Bluetooth Hands-free Car Kit, include shipping ...*" found around the first image is further used as an



1327133302 BlueSkyBluetooth
325 × 325 - **Blue Sky** Solar Bluetooth Hands-Free **Car** Kit. Includes shipping
airmiles.ca

cablecar-cable-car-terminal-building-
2592 × 1944 - ... cablecar cable **car** terminal building corner against **blue sky** at
gallery.hd.org

Speed Radars Malaga Spain - Blue S
625 × 441 - **BLUE SKY** LUXURY **CAR** HIRE | MARBELLA
blueskyluxurycarhire.es

France-Alps-La-Vallee-Blanche-view-
1182 × 878 - ... stone peak snow cable **car** station skiers on ridge deep **blue sky** CL.jpg
gallery.hd.org

Figure 1.3: This figure shows annotation details (mainly, the text surrounding an image or its filename) for some pictures from Figure 1.1 (by Google Images).

annotation for it even though there are many words (e.g., "*solar, car, shipping*") that are not directly related to the image's content. Search engine takes textual queries as input and return images with annotations most similar to them. As they do not analyze the actual content of the images, image search engines will perform poorly when retrieving pictures from unannotated collections, or with low quality annotations. The latter is common in web applications as texts found near the images are often irrelevant to their content.

To remedy this, a large number of image annotation models have been proposed recently that exploit the synergy between visual and textual modalities by learning the correspondence between image regions (or features) and keywords. These approaches follow many distinct learning paradigms, ranging from supervised classification (Smeulders et al., 2000; Vailaya et al., 2001; Gupta et al., 2008) to instantiations of the noisy-channel model (Duygulu et al., 2002) and methods inspired by information retrieval (Lavrenko et al., 2003; Jeon et al., 2003; Feng et al., 2004). Despite their differences, all these methods essentially attempt to learn the correlation between image features and words from examples of annotated images.

The Corel database has been extensively used as a testbed for the development and evaluation of image annotation models. It is 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 considered appropriate descriptors for all images belonging to the same theme. Unfortunately, the Corel database is not representative of real-world image collections. It has a small number of themes with many closely related images which in turn share keyword descriptions. It is therefore relatively easy to learn the associations between images and keywords and do well on annotation and retrieval tasks (Tang and Lewis, 2007).

Beyond the data requirements, specific learning paradiagms also limit the applicability of automatic image annotation. For example, most discriminative models, especially classification-based ones, typically achieve better performance given adequate training data than generative models, however, they are limited to a predefined vocabulary, which is usually hard to expand and embed in large vocabulary real-world applications since the classifiers should be re-trained for new keywords.

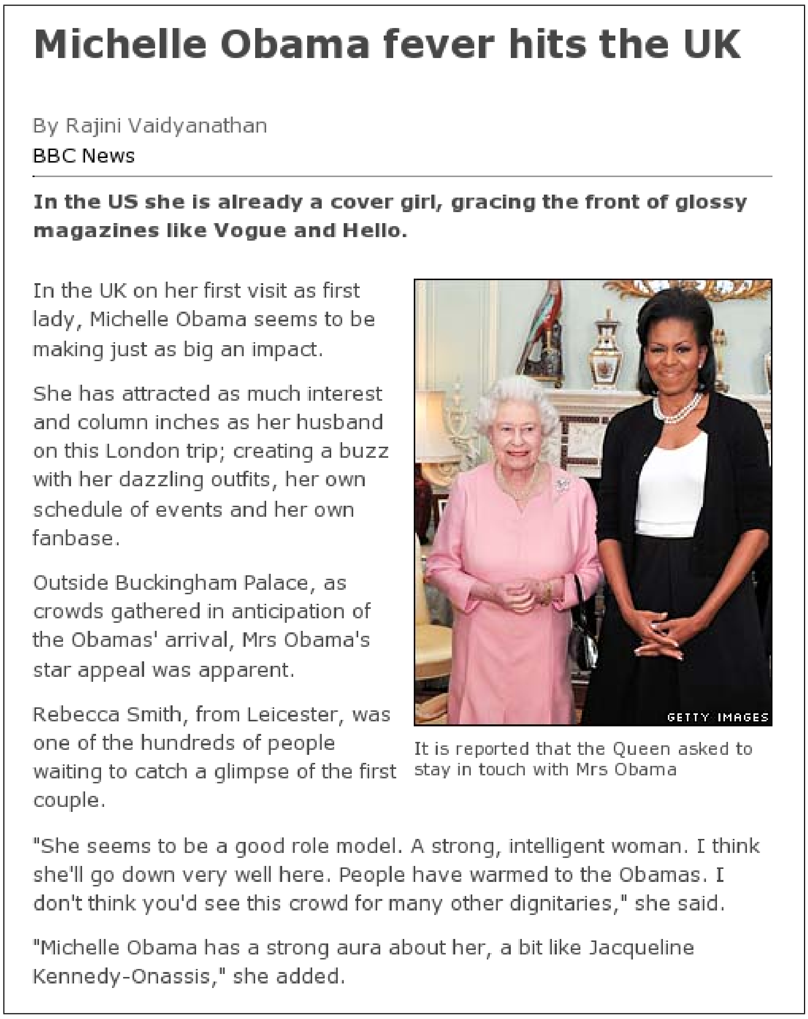
**Rendering Image Content in Natural Language** Even if we assume that we can reliably describe the image content in terms of keywords, rendering these keywords into human-readable output is far from trivial. A common framework across different image description generation methods is to rely on a domain specific background knowledge base to organize the extracted image content into a structured representation with pre-specified semantic relations, and then, to use a template-based or grammar-based surface realizer to produce sentences for this structured image content (Kojima et al., 2002; Héde et al., 2004; Yao et al., 2009; Kojima et al., 2008).

Although this framework can output grammatical sentences, the reliance on manually created knowledge bases restricts its applicability in wider domains. For instance, in an office-scene video surveillance application (Kojima et al., 2000), a human action concept ontology is manually constructed to map a sequence of human positions and postures into abstract actions (e.g., a trajectory of head motions passing a door is interpreted as the action *enter*). This knowledge base is highly related to the specific application and can not be expected to work well out-of-domain, for instance, when applied to traffic scenes. Furthermore, manually obtaining such a background knowledge base is time consuming, costly and has to be repeated for new domains. Yao et al. (2009) state that the LHI database, containing around 1 million deep segmented images together with information denoting the functional relationships among objects, is annotated by a team of 23 annotators aided by a software development team with two years full-time work[3]. Besides the creation of knowledge bases, this framework is further limited by the substantial human involvement required in the surface realization process. The predefined sentence templates or grammars are essential parts in most realizers, but most of them are not reusable across domains. The template-filling approaches often generate repetitive and stilted text due to the limited number of predefined templates. Moreover, neither template-filling or grammar based models are flexible enough to express the image content in different contexts.

**The Synergy between the Visual and Textual Modalities** Being multi-modal, the image description generation task must unavoidably exploit the synergy between visual and textual modalities. Many experimental studies in language acquisition suggest that word meaning arises not only from exposure to the linguistic environment but also from our interaction with the physical world. For example, infants, from an early age, are able to form perceptually-based category representations (Quinn et al., 1993). Perhaps unsurprisingly, words that refer to concrete entities and actions are among the first words being learned as these are directly observable in the environment (Bornstein et al., 2004). Experimental evidence also shows that children respond to categories on the basis of visual features, e.g., they generalize object names to new objects often on the basis of similarity in shape (Landau et al., 1998) and texture (Jones et al., 1991). Humans can describe images effortlessly, probably because they have a common underlying representation for the two modalities (Feng and Lapata, 2010c). Although the textual and visual modalities have been extensively studied in isolation, there is little work addressing their interaction, i.e., whether an NLP problem will benefit from a co-occurring computer vision task, or if a related NLP task benefits a computer vision problem. It is no doubt challenging but of great interest to look at this interaction and find proper representations to accommodate the synergy, which could also be useful for other multimedia applications.

---

[3] `http://www.imageparsing.com`



**Michelle Obama fever hits the UK**

By Rajini Vaidyanathan
BBC News

**In the US she is already a cover girl, gracing the front of glossy magazines like Vogue and Hello.**

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact.

She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase.

Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

Rebecca Smith, from Leicester, was one of the hundreds of people waiting to catch a glimpse of the first couple.

"She seems to be a good role model. A strong, intelligent woman. I think she'll go down very well here. People have warmed to the Obamas. I don't think you'd see this crowd for many other dignitaries," she said.

"Michelle Obama has a strong aura about her, a bit like Jacqueline Kennedy-Onassis," she added.

GETTY IMAGES

It is reported that the Queen asked to stay in touch with Mrs Obama

Figure 1.4: An extract from the BBC News website (screenshot on 2010-08-07). It contains a news image, the caption for this image, and a news document together with its title.

## 1.2 Thesis Contributions

In this thesis, we exploit data resources where images and their textual descriptions co-occur naturally. Specifically, we focus on news images, their captions, and associated articles, which are publicly available on news websites. An example is given in Figure 1.4 from the BBC News website[4]. Here the image shows Michelle Obama and the Queen of England; it is also accompanied with a caption and an article reporting on Michelle Obama's trip as a first lady to London. We explore the feasibility of automatic caption generation in the news domain, and create descriptions for such news images associated with on-line articles. Obtaining training data in this setting does

---

[4] `http://news.bbc.co.uk`

not require expensive manual annotation as many articles are published together with captioned images. Instead of relying on manual annotation or background ontological information, we exploit on-line resources which are admittedly noisy yet can be obtained easily, are abundant and contain rich linguistic information and background knowledge. We follow a two-stage modeling framework comprising of a content selection module and a surface realization component. Our approach thus first employs an image annotation model to describe the picture with keywords which are then subsequently realized into a human readable caption. For example, our phrase-based caption generation model can generate a caption "*The first lady is an impact in the UK.*" for the news image and its story shown in Figure 1.4. The main contributions of this thesis are summarized below:

1. The task of generating captions for news images is novel to our knowledge. Our work departs from previous research in image and graphics caption generation (Héde et al., 2004; Kojima et al., 2002, 2008; Yao et al., 2009; Mittal et al., 1998; Corio and Lapalme, 1999; Ferres et al., 2006), in that it analyzes the image content and renders it into a human-readable sentence in a knowledge-lean way. We utilize annotation free data which is widely available on the internet and do not rely on hand-crafted training sets, or other fine-gained knowledge bases during modeling. Essentially, our models work in a learning-from-data fashion. We learn the visual-textual correspondence from data that has not been explicitly labeled by human annotators, and then rendering the extracted image content in natural language without relying on manually created sentence-templates or grammars. Our models operate on a multimodal dataset and exploit the synergy between visual and textual modalities. Our experimental results show that the accompanying news documents helps extract more accurate image content whilst the incorporation of visual information helps create more targeted and informative image captions.

2. We focus on the data acquisition bottleneck associated with image related applications, such as image annotation, image retrieval and image description generation. We exploit data resources where images and their textual descriptions co-occur naturally. We present a new database consisting of articles, images, and their captions which we collected from on-line news sources (e.g., BBC News)[5]. Rather than laboriously annotating images with their keywords, we simply treat

---

[5] Available from `http://homepages.inf.ed.ac.uk/mlap/resources/index.html`

captions as labels. These annotations are admittedly noisy and far from ideal. We then propose an image annotation model which can learn from such annotations and their auxiliary documents. Specifically, we extend and modify the continuous relevance model (Lavrenko et al., 2003) to suit our task. Our experimental results show that it is possible to learn an image annotation model from caption-picture pairs even if these are not explicitly annotated in any way. We also show that the annotation model benefits substantially from the associated news document, beyond the caption or image.

3. We propose a probabilistic image annotation model that learns to automatically label images under the assumption that images and their surrounding text are generated by a shared set of latent variables or topics. Specifically, we describe texts and images by a common multimodal vocabulary consisting of textual words and *visual terms* (visiterms). Using Latent Dirichlet Allocation (LDA, Blei and Jordan 2003), 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. Our experimental results show that our model is robust to the noise inherent in such data and is useful on its own right, not limited to the caption generation task. It improves upon competitive approaches that prioritize one modality over the other or exploit them indirectly. We also show how the model can be straightforwardly modified to perform automatic text illustration. Experimental results on both tasks bring improvements over competitive models.

4. Inspired by recent advances in text summarization, we propose both *extractive* and *abstractive* caption generation models to render the extracted image content into natural language without fine-gained sentence-templates and grammars. The backbone for both approaches is our topic-based probabilistic image annotation model that suggests content for an image with the help of its associated document. We propose several *extractive* models and examine how to best select sentences that overlap in content with our image annotation model. We also show how an existing *abstractive* headline generation model can be modified to fit to our image caption generation scenario by incorporating visual information. Our own models operate over image description keywords and document phrases, and we also take dependency and word order constraints into account.

Experimental results show that both approaches are possible to generate human-readable image captions without relying on manually created sentence-templates or grammars. Our abstractive model defined over phrases yields more grammatical output than word-based models and manages to yield as informative captions as human author.

## 1.3 Thesis Overview

The remainder of this thesis is structured as follows:

- Chapter 2 surveys previous work on image description generation, which usually contains two modules, content selection and surface realization. We summarize endeavors in image description generation from both the computer vision and natural language processing communities. We review image annotation models, including discriminative and generative ones. The latter will serve as our content selection module and are discussed in more detail. We also review aspects of text summarization, both extractive and abstractive, that are relevant to our image caption generation task.

- Chapter 3 is concerned with the dataset we use throughout this thesis. We discuss popular image databases used in previous computer vision research and their shortcomings. We show that the rich resources available on the internet are a good place to harvest freely annotated data. Specifically, we introduce a news image dataset and argue that it is suitable as a testbed for the task of image caption generation. This new dataset differs from traditional image databases as it has not been explicitly annotated by human annotators, and is thus noisy in nature, it contains low resolution images covering various topics, and has a unique component—the associated news document.

- Chapter 4 details our efforts to adapt the continuous relevance model (CRM) (Lavrenko et al., 2003), a state-of-the-art image annotation model, to our news dataset. We extend CRM by taking into account the associated news documents. We use the extra document information to smooth the conditional probabilities of keywords given the news image, and further prune the model's output by assuming that image content words should be strong topic indicators of the associated document. Our experimental results provide evidence that it is possible to create an annotation model from noisy data that has not been explicitly hand labeled and show that the extended CRM works better than either the original CRM or solely text-based models.

- Chapter 5 introduces a generative image annotation model which improves upon the extended CRM model. We survey existing latent variable image annotation models and highlight the importance of balancing the contributions of visual and textual information in a topic model. We show how the image and its associated textual document can be *jointly* rendered into a mixture document which is used to build a topic model where the two different modalities (in the form of a mixture document) are deemed to be generated by a set of latent topics. Our annotation model takes these topic distributions into account while finding the most likely keywords for an image and its associated document. We also show how the model can be straightforwardly modified to perform automatic text illustration.

- Chapter 6 discusses several approaches we develop to produce a caption for a news image given its associated document. Specifically, we formulate the caption generation task as a summarization problem. We first attempt extractive approaches and investigate different criterions to select a description sentence from the document as the image caption. We also propose abstractive approaches akin to headline generation and introduce both word-based and phrase-based caption generation models. We evaluate our models based on automatic and manual evaluation. We 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.

- Chapter 7 summarizes the major findings of the thesis and suggests future research directions.

## 1.4 Published Work

This thesis is mainly based on three publications:

- Chapters 3 and 4 are extended version of the paper "Automatic Image Annotation Using Auxiliary Text Information" (Feng and Lapata, 2008). Specifically, Chapter 3 describes the dataset we use throughout and Chapter 4 presents the Extended Relevance Model.

- Chapter 5 elaborates on the paper "Topic Models for Image Annotation and Text Illustration" (Feng and Lapata, 2010b), where we demonstrate how a classic topic model can be extended to perform the automatic image annotation task in our news dataset.

- Chapter 6 expends the paper "How Many Words Is a Picture Worth? Automatic Caption Generation for News Images" (Feng and Lapata, 2010a), where we propose both extractive and abstractive models to render the image content (extracted by our probabilistic annotation model introduced in Chapter 5) into human-readable sentences.

# Chapter 2

# Related Work

In this chapter, we will broadly review the work related to the field of automatic image description generation. We first discuss previous methods on automatic description generation for images and videos, and then look at recent work on automatic description generation for graphics. Automatic image description generation is more similar to our scenario, as pictures or key frames of videos are first preprocessed, and then their content is rendered into natural language descriptions. The graphics case generally avoids complex image processing, assuming that the data used to draw the graphics are already at hand, hence places more emphasis on how to verbally convey the information inherent in the graphics, especially on information that is easy to visualize but usually omitted.

As we discussed in the previous chapter, in order to solve this cross-disciplinary task, we need to deal with two main problems, namely automatic image annotation and description generation, that, although closely related, have been previously studied in isolation. When looking at previous efforts in automatic image annotation, we address the problem in terms of the training paradigm employed and their capability dealing with real-world data. Current image annotation approaches fall under two broad categories: discriminative and generative models. The former usually achieve better performance however are heavily reliant on the quality of training data which in turn influences their extendibility. In contrast, generative models can deal with low quality data more easily as well as changes in training set or even vocabulary.

Recall that in this thesis we will focus on news data, where news images with associated captions and documents co-occur naturally. This type of data will be used as it is without any additional manual annotation and will allow us to treat caption generation as a summarization model. We thus examine current advances of text sum-

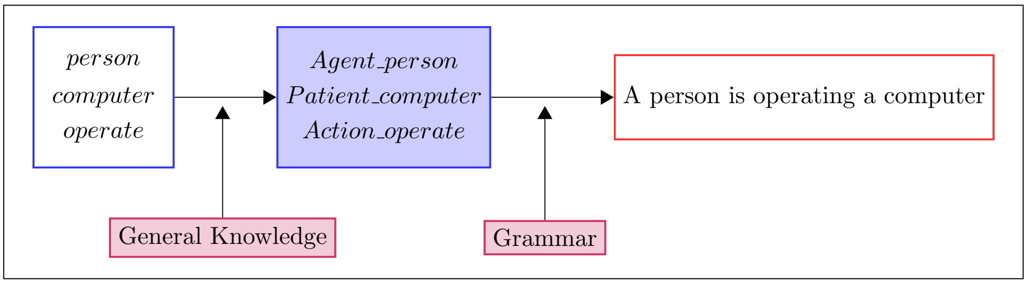


Figure 2.1: This figure shows an example of single sentence generation, where the input is a set of keywords, a knowledge base is used to interpret the specified roles for these keywords, and then a grammar helps create the sentence.

marization, survey existing summarization approaches, both extractive and abstractive, and especially keep an eye on whether extra knowledge bases are employed. Extractive approaches dominate the field of automatic text summarization. The main reason is that without good linguistic analysis, it is possible to output good enough summaries both in terms of their content and grammaticality simply by deciding which sentences present the key ideas of the document. Abstractive summarization, on the other hand, is a more challenging task as sentences need not only be extracted but also rewritten. However, it has the potential of creating more human-like summaries that are more succinct and coherent.

We first describe the task of automatic image description generation, and briefly review previous approaches and related applications. Next, we proceed to review automatic image annotation and text summarization.

## 2.1 Problem Formulation

Natural language generation (NLG) is the task of producing natural language output according to certain input (Jurafsky and Martin, 2000). The input depends on the specific requirements of various applications. For instance, in single sentence generation, it could be a set of concepts with specified relations, or just a set of isolated keywords. And the output is expected to satisfy the input requirements, and also to be grammatical and semantically coherent. These two modules are often referred to as content selection and surface realization, Content selection usually requires a knowledge base to assist in better interpreting the input concepts. In Figure 2.1, we show an example

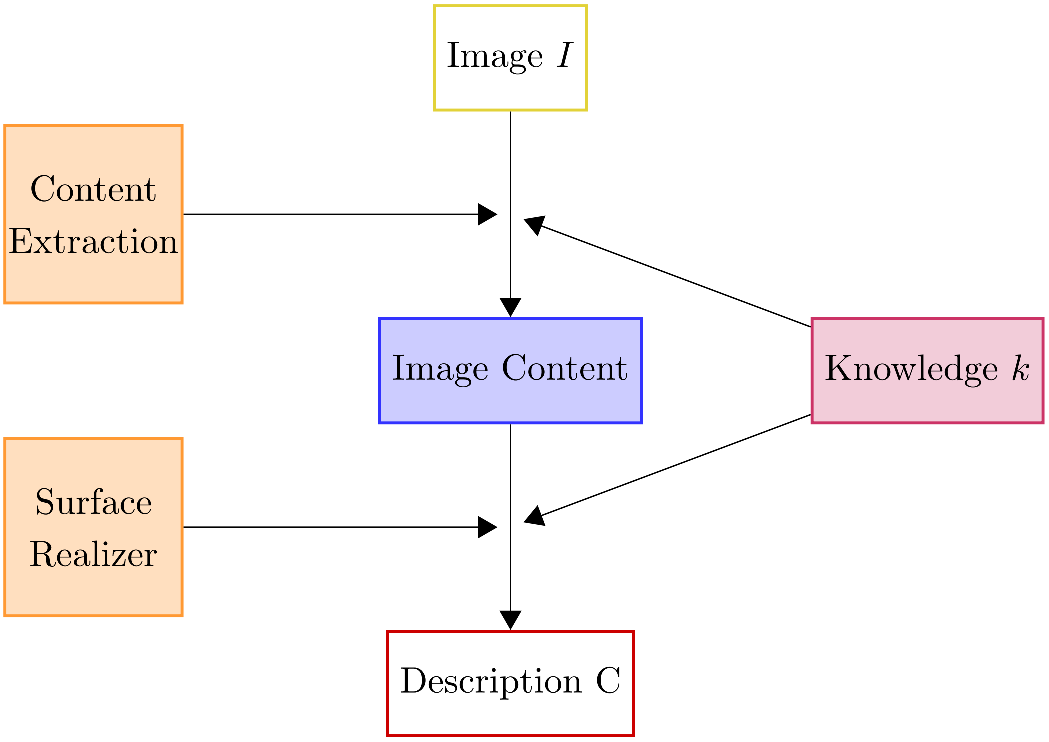


Figure 2.2: This figure illustrates a general pipeline for the task of automatic image description generation

of single sentence generation[1]. Here, the input is *person, computer, operate*, and a knowledge base is then consulted to determine the relations between these words (e.g., *person* is an agent, *computer* is the patient, etc.). Accordingly, a grammar manages them into a grammatical sentence.

However, in the task of automatic image description generation (as shown in Figure 2.2), the input is an image, and therefore, content selection involves interpreting the image and representing its content. Here we assume that a set of keywords is a good enough representation of the image's content. Subsequently, a surface realizer takes these keywords as input to generate a description.

More formally, we define the task of automatic image description generation as below:

**Definition 1.** *Given an image $I$, and a related knowledge database $\kappa$, create a natural language description $C$ which captures the main content of the image under $\kappa$.*

Following the typical natural language generation paradigm, the task involves, first

---

[1] In a multi-sentence case, the first step will evolve to discourse planner aiming to provide a better sentence structure, discourse structure, and so on.

analyzing and representing the image content and then rendering it in natural language. And the knowledge base $\kappa$ must contain two types of information, information about how the images (or image regions) corresponds to words and information about how these words can be combined to create a human-readable sentence.

## 2.2 Automatic Image Description Generation

As a relatively new task, automatic image description generation has not yet received as much attention as automatic image annotation or sentence generation. Following specific applications, two different streams of work have addressed this problem within computer vision and natural language processing, respectively.

In computer vision, there is an increasing demand for describing images or video frames more linguistically, e.g., with description sentences rather than isolated keywords lists. More emphasis has been placed on extracting content from images or video key frames, e.g, by recognizing objects or even interpreting human actions. Generally, the content is first extracted and represented as a keyword list or concept entries in a dictionary, and next a natural language generation module arranges this content into human-readable sentences, often using sentence-templates or a functional grammar-based surface realizer.

Much work in NLP looks at the problem of generating explanatory captions or descriptions for graphics. It is not surprising that these methods avoid the hard image processing problem during content extraction, and focus on expressing the gist of the images by assuming that the data used to produce the graphs are already available.

**Automatically Generating Descriptions for Images** A handful of approaches have been proposed in the literature that automatically generate descriptions for images by examining their content. To begin with, the image is represented by image features, which are then replaced by an abstract representation, essentially a set of description words, according to a visual-to-textual representation dictionary (Héde et al., 2004; Kojima et al., 2002, 2008; Yao et al., 2009). The features used to represent the image content mainly include color information (Héde et al., 2004; Yao et al., 2009; Kojima et al., 2002), textual features (Héde et al., 2004; Yao et al., 2009), detected edges (Kojima et al., 2002, 2008), and so on. For certain applications, some objects are detected and recognized with prior knowledge to supply higher level features (Abella et al., 1995; Kojima et al., 2002; Yao et al., 2009). For instance, in some video surveillance

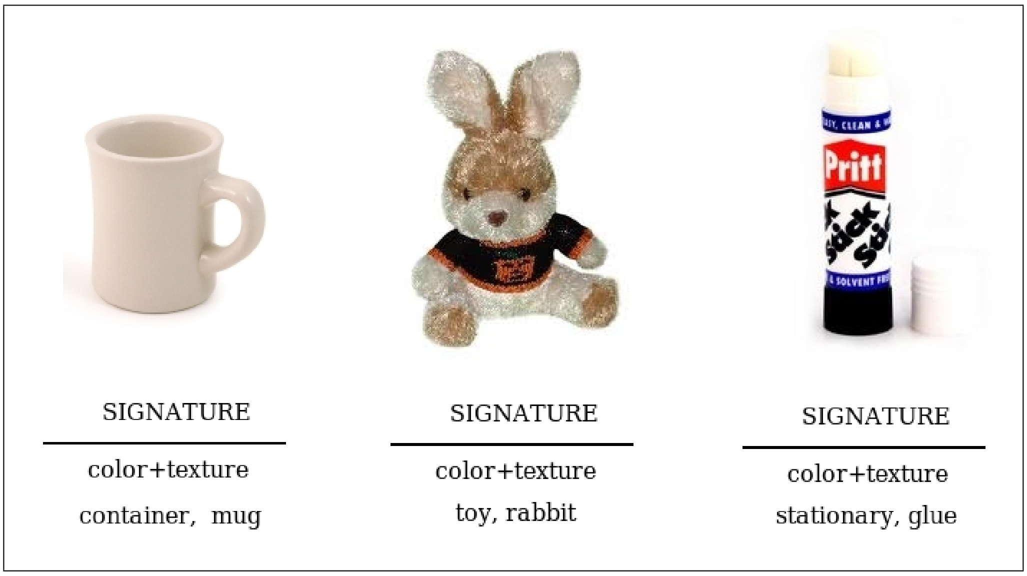


Figure 2.3: Example of an object dictionary (Héde et al., 2004). This manually created dictionary contains images of different objects and their signatures, essentially encoding the correspondence between visual features to keywords.

applications, human heads are recognized and then used to extract human walking trajectories (Kojima et al., 2002). The dictionaries mapping visual features to concepts or keywords are usually constructed by humans (Abella et al., 1995; Kojima et al., 2002; Héde et al., 2004; Yao et al., 2009). The abstract interpretation extracted from the image is in turn used as input for a surface realizer to produce a verbal description (Abella et al., 1995; Kojima et al., 2002; Héde et al., 2004; Yao et al., 2009). A common theme across different models mentioned here is domain specificity, the use of hand-labeled data, and reliance on background ontological information.

For example, Héde et al. (2004) attempt to generate descriptions for images of objects shot in a uniform background. Their work highlights the importance of a content representation with semantic relations in image database related applications, e.g., a phrase "*an orange ball*" explicitly indicates the modified relationship between *orange* and *ball* while isolated words "*orange, ball*" might describe two objects, an *orange* and a *ball*, rather than one (*an orange ball*). Their system relies on a manually created dictionary of objects, each entry is indexed by an image signature (e.g., raw image features, such as color and texture, and two keywords, the object's name and category). Figure 2.3 illustrates an example of such an object dictionary. The model first segments images into regions, retrieves corresponding signatures from the database by comparing the region features with entries in the dictionary, and produces a description sentence using the retrieved signature keywords and selected sentence templates. Researchers from the medical science adopt a similar procedure to generate text to describe the referential positions of renal stones (Abella et al., 1995).

In the applications of video surveillance (e.g., in office scenes), Kojima et al. (2002) first recognize human poses and heads with their moving trajectories from video frames, interpret these numerical values into abstract actions (e.g., *enter*,*exit* and *operate*), and then create scene descriptions using predefined grammars. The interpretation from numerical visual features to human action concepts is based on a manually created concept dictionary. They further improve the method by recognizing more objects, detecting contact points between human bodies and other objects, constructing a more complex concept hierarchy which organizes human bodies, contacts with objects into a sequence of motions in a coarse-to-fine manner (Kojima et al., 2008).

More recently, Yao et al. (2009) present a general framework for generating textual descriptions of images or videos through a pipeline consisting of an image parser, a visual knowledge representation, the semantic web and a text generation engine. Firstly, images are hierarchically decomposed into their constituent visual patterns, which are subsequently converted into structured representations with specified semantic relations including categorical, spatial and functional relations, using an image parser and a visual knowledge database. The text generation engine render this structured representation into a natural language description with the help of the semantic web. The image parser is guided by the visual knowledge representation which supplies an non-ambiguous way to organize the parsing results (a large number of visual patterns and semantic relations among these patterns) into an And-Or Graph. However, both the parser and visual semantic representation are built based on a large-scale ground truth image database which is manually annotated by a full-time team as well as a tool-development team. The text generation step is relatively simpler: a multi-sentence description is generated using a document planner and a surface realizer, where the sentence templates or grammars are predefined according to the specific applications at hand (e.g., the video surveillance case).

The approaches discussed above generate grammatical natural language sentences for images or videos by analyzing the image content. However, note that all of them rely on large amounts of manually created resources. This includes the annotation of the image database for the training purpose, the construction of a visual-textual correspondence dictionary or ontology, and the engineering of application-specific sentence templates or grammars for generation. And most of this data can not be reused cross domains or applications and thus manual effort must be invested for a new one, which is obviously costly and time consuming.

**Automatically Generating Descriptions for Information Graphics** Information graphics such as pie charts, plots and bars are commonplace in many documents. However, information conveyed in the graphics is not always present in the document, or, only a small portion of it is mentioned. An explanatory caption is thus often needed to complement the graphics or bridge the graphics and document. Within the natural language processing community, most previous efforts have focused precisely on generating captions for complex graphical presentations (Mittal et al., 1998, 1995) or on using the captions accompanying information graphics to confirm their intended message, e.g., the author's goal to convey ostensible increase or decrease of a quantity of interest (Corio and Lapalme, 1999; Fasciano and Lapalme, 2000; Elzer et al., 2005). Here, the emphasis is more on how to clearly describe the data in the graph and select proper sentence templates rather than actually to analyze the content of picture. It is thus assumed that the data used to create the graphics is in structured format and already at hand. This entails that there is no image processing involved as the aim is not so much to describe the picture but to create explanatory captions that help users to recover the information conveyed in the graphics but sometimes omitted in the documents.

For instance, Mittal et al. (1995, 1998) tackle the problem of generating "explanatory captions" for users to fully understand information graphics. Their system first analyzes the data used to produce the graphics and focuses on deciding what to say in the captions (e.g., a perceptually complex part, implicit relations among data objects or necessary information but omitted the main thread of the document ) and how to describe them with the help of a sentence realizer and a multi-sentence planner.

Corio and Lapalme (1999) propose a graphic caption generation system which first distinguishes the writer's intentions, e.g., subjective or objective, comparative or descriptive, and then accordingly generate accompanying captions or short text to complement the information available from the graphics. They manually investigate a corpus of 411 journals about Statistics, and then obtain rules for what to say and how to say in the captions of graphs. For example, *descriptive* captions are expected to indicate the general trend of how a value evolved, while *domination* messages are expected to identify the extreme value points. These rules are then applied to instruct a generation module to choose appropriate lexicon and sentence-templates.

Other work has focused on the presentation of graphics content for people with special needs. A prototype system, *iGraph* proposed by Ferres et al. (2006), is designed to help people with visual impairment to better access graphical information. Their system outputs short descriptions for a graph and provides an interface to support natural language querying given a graph. In order to better understand the needs of people with visual impairment, they conduct a series of surveys, to collect information on the users' expectations, and their common requests when given a graph. For example, they find that the most frequently used description words are *X, Y Axes* and *line up/down* while the top request is about the purpose, type and title of the graph (main and axes). The prototype of the system is then improved according to these findings.

In addition, Carberry et al. (2004) address the importance of graphics description generation in document summarization. They argue that these graphical representations play an important role not only in conveying information that is not directly described in the text, but also in deciding what is necessary for the document's summary. Different from other work, Carberry et al. (2004) utilize a computer vision module to analyze the graphs and capture their components as well as the relations among these components. They further argue that graphics descriptions, together with writers' intentions, strongly influence the content selection of the summary, e.g., whether other topics of the document will appear in the summary.

Compared to generating descriptions for images or videos discussed before, it is not difficult to observe that the graphics description generation approaches mainly tackle the problem in terms of how to well present information available in the graphics while paying little attention to extracting contents from these graphics. Therefore, their focuses are on the level of planning the descriptions, where, in order to make readers better understand the text, a model should first make choices on what to say and how to say, in other words, in which structure the selected contents will be presented for the purpose of better communication. The remaining job, rendering the selected contents into human-readable sentences, is taken by a surface realizer, which again relies on manually created sentence-templates or grammars.

## 2.3 Automatic Image Annotation

Over the past decades, image annotation and other related areas, e.g., object recognition have received more and more attention within computer vision and information retrieval. For example, the rapid growth of image collections on the internet indicates the increasing demand for searching and browsing. In practice, given a query, search engines retrieve relevant pictures by analyzing the image caption (if it exists), surrounding text, metadata of the image (e.g., filename and shooting conditions) or user click behaviors. However, since they do not examine the actual content of the images, search engines cannot be used to retrieve unannotated images. The ability of automatically annotating images with keywords would be of significant practical importance for many image related applications. For our purpose, automatic image annotation can be seen as an approximation to full image understanding and can be used as a means to automatically obtain keywords that broadly describe image content.

Automatic image annotation bears some similarity to object recognition, i.e., the task of trying to identify high-level meaningful concepts given a set of low-level visual features of the image. Here, words are assumed to represent the concepts and images are sometimes segmented into regions and represented with various features. Then the problem can be formally defined as:

**Definition 2.** *Given an image I with visual features* $V_I = \{v_1, v_2, ..., v_N\}$ *and a concept set* $W = \{w_1, w_2, ..., w_M\}$*, where M is the number of concepts, the image annotation task is to find the subset* $W_I$ *(*$W_I \subseteq W$*), which can appropriately describe the image I.*

Previous image annotation work can be broadly classified into two streams (Yao et al., 2009). The first class of methods focuses on isolated labels, e.g., *animal, plant, human, sheep, tree, car*, etc. The second class addresses the semantic relations between these concepts, which are usually organized in a hierarchy, e.g., *sheep is a sub-category of animal, human can drive a car, etc* (Fan et al., 2007). Although the latter approaches work with richer information which may be crucial for many image related applications, they usually rely on the training data that should be heavily annotated by humans.

In terms of the learning paradigm employed, existing approaches are mostly either discriminative or generative[2]. The former directly model the conditional probabilities of keywords given image features, $P(w|v_1, v_2, ...)$, while the latter model the joint probability of keywords and image features, $P(w, v_1, v_2, ...)$. We will review related approaches following this distinction, whilst paying attention to the knowledge bases used during modeling.

---

[2] Another stream of work is cast in a semi-supervised learning setting, where each unlabeled sample is assumed to originate from one of the known classes (or concepts) which can be effectively learned from existing annotated training data, mainly through a classifier-based approach (Fan et al., 2005; Schroff et al., 2007).

### 2.3.1 Discriminative Models

Most discriminative image annotation methods originate from the earlier image classification, or scene classification research (Vailaya et al., 1999, 2001). The task is formulated as an $M$-class classification problem. Specifically, all elements of the concept set $W$ are considered as different classes ($M$ in total), and a binary classifier for each concept is trained one by one on the training set. Some methods adopt a ”one vs all” model (Vailaya et al., 1999; Maron and Ratan, 1998; Qi and Han, 2007; Chai and Hung, 2008; Gupta et al., 2008). For each semantic concept, the training data is re-labeled, positive or negative (1 or 0), according to the concept, and a binary classifier is trained on re-labeled data. For each image in the testing set, all $M$ classifiers are used to examine the presence or absence of corresponding concepts.

However, this “one vs all” model has a potential shortcoming as it needs strongly labeled training data, which means all objects appearing in the image must be annotated explicitly, since images with missing labels will negatively affect the accuracy of the classifiers that are trained for these missing labels (Carneiro and Vasconcelos, 2005). Unfortunately, strongly labeled databases, either in large scale or with large vocabulary size, are very expensive and time-consuming to construct, and it is nearly impossible in real-world applications. In addition, under this framework, the number of classifiers is decided by the size of the concept set, and when adding a new concept, all existing classifiers have to be re-considered. If the training set adds new data, then all classifiers have to be re-trained. Some of these shortcomings have been addressed by Carneiro and Vasconcelos (2005), who propose an improved multiple classifiers based method and obtain good results. Here, each concept still defines a single class but multiple instance learning is used to estimate the class density in order to make each class directly compete with the others during annotation. Fan et al. (2005) overcome the data acquisition problem by using a mixture of labeled and unlabeled samples and by introducing a semi-supervised framework which enables multi-level concept modeling and hierarchical classifier training. They use the EM algorithm to obtain the base concept-level classifiers. A hierarchical mixture model is then used to combine these classifiers for higher level concept learning. Jeon and Manmatha (2004) train a maximum entropy model and achieve competitive results with the state of the art (Lavrenko et al., 2003; Feng et al., 2004).

In sum, discriminative image annotation approaches are difficult to port across different databases with different concept sets (Ulusoy and Bishop, 2005). They do not scale or generalize well since these classifier-based models usually require predefined concept sets of fixed size and are sensitive to the changes of training data and concept sets which usually result in manually re-preparing large mounts of new training data. Although these approaches may achieve better results all things being equal, they highly depend on strongly labeled training data, and correspondingly their accuracy is closely related to the quality of the training data (Holub, 2007).

### 2.3.2 Generative Models

Many generative models that have been successfully applied in speech recognition, machine translation and information retrieval, have been ported to image annotation (Mori et al., 1999; Duygulu et al., 2002; Barnard and Forsyth, 2001; Wang and Li, 2002; Barnard et al., 2002; Lavrenko et al., 2003; Blei and Jordan, 2003; Feng et al., 2004; Pan et al., 2004; Jin et al., 2004; Li and Wang, 2006, 2003). The key idea here is to model the joint probability of images and annotated keywords based on the training data. Generally speaking, these approaches first introduce a set of latent variables, and the joint probability is built to describe the relationship between the image features and keywords with the help of these latent variables. In other words, the joint probability can be also considered as a measurement of how much the images and words can mutually describe the latent variables. Formally, for an image $I$, with visual features $V_I$, and annotated words $W_I$, we can define the latent variable conditional joint probability $P_s(V_I, W_I|s)$, where $s$ is one of the latent variables. For the entire training set, we simply sum over all latent variables as:

$$P(V_I, W_I) = \sum_{s \in S} P_s(V_I, W_I|s) P(s), \tag{2.1}$$

where $S$ is the total set of the latent variables and $P(s)$ is the prior probability of the latent variable $s$. The equation is common in most generative models, and different instantiations can be derived from it depending on the assumptions of specific applications. Next we will present several typical models under this framework.

#### 2.3.2.1 Image Annotation as Statistical Machine Translation

Inspired by statistical machine translation, Duygulu et al. (2002) formulate image annotation as the process of learning lexicons from a bitext. Briefly, in statistical machine translation, we are given a parallel corpus of French and English, and our task is to learn how to turn English sentences into French, assuming that the aligned bitext potentially works as a codebook supplying the underlying lexical correspondence between the two languages. Analogously, for the image annotation task, we have an image-words bitext, consisting of images (segmented into regions) and their annotated keywords from the training set, which enable us to translate the regions into text.

Duygulu et al. (2002) segment images into regions and cluster the latter using K-means into 500 classes which they call blobs. They assume that blobs correspond to objects in images. Next, they learn the correspondence between the blobs and words, using the IBM machine translation model 2 (Brown et al. 1993), to capture the probability of translating images into words:

$$P(W|B) = \prod_{n=1}^{D} P(W_n|B_n) = \prod_{n=1}^{Image} \prod_{j=1}^{Word_n} \sum_{i=1}^{Blob_n} P(Alig(w_{nj}, b_{ni}))P(w_{nj}|b_{ni}) \tag{2.2}$$

where $P(w|b)$ indicates the probability of obtaining word $w$ given blob $b$; and $Alig(w_{nj}, b_{ni})$ gives the alignment probability that $b_{ni}$ translates to $w_{nj}$. Using the Expectation Maximization (EM) algorithm, they learn the translation probabilities from blobs to words, which makes it easier to compute the probabilities of keywords given a test image.

Despite this intuitive formulation, their model's performance is not as good as expected for several reasons. First of all, the image regions, segmented by Normalized Cuts (Shi and Malik 2000), are not always meaningful. In machine translation, words are meaningful units, while automatically segmented image regions are not. Automatic image segmentation is still an unsolved problem and it is nearly impossible to find an algorithm to successfully deal with all types of images (more details will be provided in Section 2.3.4). Furthermore, the K-means clustering process may lead to unreliable clusters. This is due to the large variance that objects exhibit from different viewing angles, scales, or other transformations. An object might be clustered into different groups when it was shot under different conditions. Another important issue is the quality of the training data which should be ideally strongly labeled, since the modeling procedure operates on the region level. Regarding to corpus size, in statistical machine translation training data usually consists of millions of sentences, with dozens of words in each sentence, but in Duygulu et al. (2002), the size of the training data was 4500 images, with less than 5 annotated words and 10 blobs per image. This dataset is substantially smaller and thus not enough to reliably capture the correspondence between regions and keywords in IBM model 2.

#### 2.3.2.2 Image Annotation Based on the Continuous Relevance Model

The task of image annotation could be considered as the process of modeling the relevance of a set of keywords given a document (the document here is the image) based on the latent variables, i.e., pairs of image-keywords. Generative approaches for modeling relevance have been actively investigated in information retrieval (Lavrenko, 2004; Lavrenko and Croft, 2001; Metzler et al., 2004), and applied to the image annotation task (Lavrenko et al., 2003; Feng et al., 2004; Jeon et al., 2003). These methods originate from the relevance-based language model (Lavrenko and Croft, 2001), and try to learn the joint distribution $P(V,W)$ of words $W$ and image regions $V$. The key assumption here is that the process of generating images is conditionally independent from the process of generating words. Each annotated image in the training set is treated as a latent variable. Recall formula (2.1), which can be rewritten as:

$$P(V_I,W_I) = \sum_{s\in S} P(V_I|s)P(W_I|s)P(s), \tag{2.3}$$

where $S$ are all annotated images in the training set. The conditional probabilities $P(V_I|s)$ are estimated using a Gaussian kernel distribution and $P(W_I|s)$ is estimated using the mulinomial distribution (see Lavrenko et al. 2003) or the multiple Bernoulli distribution (see Feng et al. 2004). We discuss the continuous relevance model in more details in Chapter 4.

The parameters of these distributions need to be estimated from a labeled image dataset. In the model proposed by Feng et al. (2004), the word generative distribution $P(W_I|s)$ is estimated using a multiple Bernoulli model instead of a multinomial one, which places more emphasis on the presence of a word rather than its prominence. For example, given an image of two persons, the multiple Bernoulli model will focus on whether the word "*people*" has appeared in the annotation, while the multinomial model will care about how many times "*people*" has appeared in the annotation. In practice, the presence of a word is more useful than its prominence since it can make the estimated probability distribution concentrate on the concepts or objects and partially avoid the negative effect of weak labeling.

Jin et al. (2004) further improve on the multiple Bernoulli-based model by adding a bigram language model:

$$P(V_I,W_I) = \sum_{s\in S} P(V_I|s)P(W_{I1},W_{I2}|s)P(s). \tag{2.4}$$

This language model helps address the correlation between annotated keywords, and the improved model prefers words that frequently co-occur with the annotation keywords that are already selected. Similar improvements include introducing beam-search during the iterations over the whole vocabulary and then discarding impossible candidate words according to already selected keywords (Moran, 2009).

The continuous-space relevance model is relatively simple in structure but effective because of its reasonable assumptions and parameter estimation. In addition, it is not sensitive to the strong/weak labeling issue, as the probabilistic nature of the definition for relevance as well as Bayesian estimation for parameters can smooth the negative effect of weak labeling or other data quality issues. Moreover, it can naturally handle multiple labels for multiple objects in one cluttered area, e.g., *a desk with a laptop on its surface and a chair in front of it.* However, to some extent, the simplifying assumptions made above seem too strict and eliminate the potential of the latent variables since the model places more emphasis on the aspect of modeling words.

#### 2.3.2.3 Image Annotation based on Topic Models

Researchers in the image annotation community have also been inspired by modeling text document. Barnard and Forsyth (2001) propose a generative hierarchical model, related to the hierarchical mixture model proposed by Hofmann (1998). In this model, data is generated by a fixed hierarchy of nodes with leaves corresponding to clusters each of which has some probability to generate words or image regions. Blei and Jordan (2003) extend Latent Dirichlet Allocation (LDA, Blei et al. 2003) to the image annotation task and propose CorrLDA to model the relations between words and images. The model assumes that image regions ($v_n$) are generated from a multivariate Gaussian distribution $P(v_n|z_n)$ conditioned on factors ($z_n$) of a multinomial distribution ($z_n \sim Mult(\theta)$) which is derived from a Dirichlet distribution ($\theta \sim Dir(\theta|\alpha)$). And each keyword is drawn conditionally on the factor ($z$) that has just generated a randomly selected image region. More formally:

$$\begin{aligned} P(V_I, W_I, \theta) =& P(\theta|\alpha) \\ & \times \Big( \prod_{n=1}^{RegionN_I} P(z_n|\theta) P(v_n|z_n) \Big) \\ & \times \Big( \prod_{m=1}^{WordN_I} P(y_m|RegionN_I) P(w_m|y_m, z) \Big) \end{aligned} \tag{2.5}$$

where $RegionN_I$ and $WordN_I$ denote the number of regions and words in image $I$, $\alpha$ and $\theta$ are the priors of the Dirichlet and Multinomial distributions, respectively. This is a mixture model thus allowing to explore the relations between mixture components.

Similarly, standard latent semantic analysis (LSA) and its probabilistic variant (PLSA), are applied to address the multimodal relations from a multimodal space consisting of image features combined with the corresponding keywords (Pan et al., 2004; Monay and Gatica-Perez, 2003, 2007; Fergus et al., 2005; Sivic et al., 2005). In addition, Fei-Fei and Perona (2005) propose a variant of LDA for learning natural scene categories.

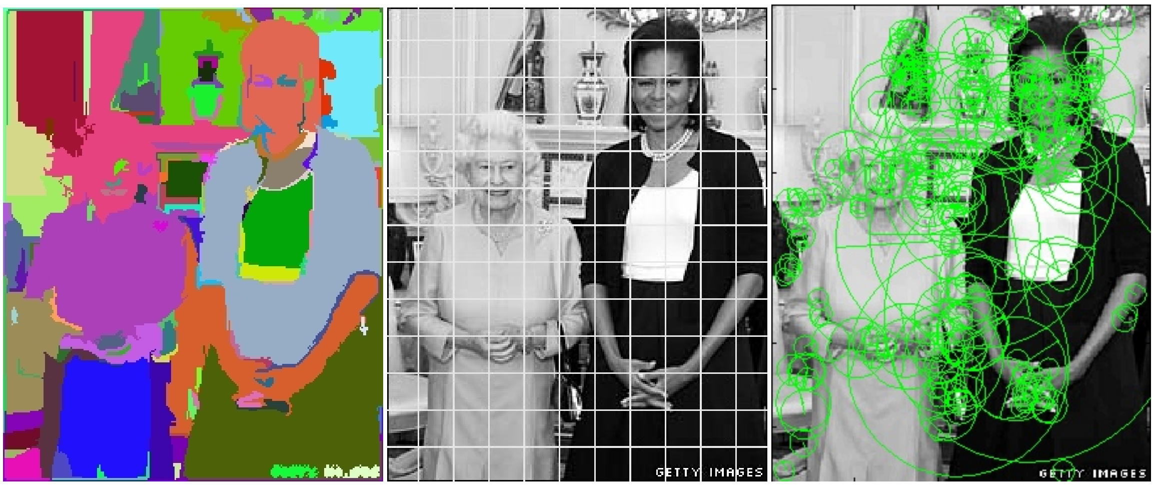
### 2.3.3 Other Statistical Methods

Other well-known statistical methods have also been successfully applied to the image annotation task. Mori et al. (1999) simply count the most frequent co-occurring words for each image region. This very naive model works well on an encyclopedia collection database (Mori et al., 1999). Pan et al. (2004) estimate the correspondence between image regions and words through the correlation between two weighted co-occurrence frequency matrices. Wang and Li (2002) develop an image annotation system using the 2-D Multi-resolution Hidden Markov Model (MHMM), which can explore the statistical dependence among image regions across multiple resolutions. Active learning has also been employed to reduce the size of the training database (Jin et al., 2004). Other work has adopted multiple instance learning (MIL) in the visual feature space to learn the visual distributions for different concepts, and formulate the problem in an M-class classification setting. The use of MIL greatly reduces the negative effect of weak labeling (Maron and Ratan, 1998; Carneiro and Vasconcelos, 2005).

### 2.3.4 Challenges in Automatic Image Annotation

Despite numerous efforts, automatic image annotation remains a challenging task. This is due to three inter-related factors: (a) the quality of image segmentation and preprocessing, (b) ambiguity inherent in natural language and issues of synonymy, and (c) the availability and quality of training data.

**Image Features** As mentioned earlier, images are typically segmented into regions before modeling in automatic image annotation task. It is generally assumed that these regions have some underlying high-level meanings represented by a set of low-level features, such as color, size, edges, texture, relative positions and so on. There are mainly two ways of obtaining image regions, either using a generic image segmen-



(a) (b) (c)

Figure 2.4: Image partitioned into regions of varying granularity using (a) the normalized cut image segmentation algorithm, (b) uniform grid segmentation, and (c) the SIFT point detector.

tation algorithm to segment the image into several regions (see Figure 2.4(a)) or averagely dividing the image into a grid, with certain number of rectangles (see Figure 2.4(b)).

It is well known in computer vision and image processing communities that image segmentation is a challenging problem (Shi and Malik, 2000). Until now, there is no generic image segmentation algorithm that can deal well with all kinds of images in all conditions. Normalized cuts (Shi and Malik, 2000) are widely used to segment images as they produce acceptable output for a wide range of occasions and non-surprisingly are popular in image annotation research (Barnard and Forsyth, 2001; Duygulu et al., 2002; Barnard et al., 2002; Lavrenko et al., 2003; Blei and Jordan, 2003; Jeon et al., 2003; Pan et al., 2004; Jin et al., 2004). The Normalized Cuts algorithm (Shi and Malik, 2000) treats image segmentation as a graph partitioning problem. A novel criterion which combines intra-group similarity and inter-group dissimilarity is used to optimize the partitions. In the literature, normalized cuts is reported to achieve better segmentation results in many applications compared to other methods (Datta et al., 2008) .

However, surprisingly, Feng et al. (2004) show that averagely cut rectangular regions yield better results than those obtained by Normalized Cuts. In order to avoid unnecessary errors induced by image segmentation algorithms, some approaches (Feng et al., 2004; Li and Wang, 2003; Wang and Li, 2002; Jeon and Manmatha, 2004; Lavrenko et al., 2004; Mori et al., 1999; Carneiro and Vasconcelos, 2005) directly divide the image evenly into rectangular regions. Specifically, Wang and Li (2002) adopt the 2-D Multi-resolution Hidden Markov Model, which has been successfully applied in image segmentation research and can explore the statistical dependence among image rectangles across multiple resolutions, thereby avoiding the reliance on image segmentations. The size of the image region usually depends on the database.

Once region segmentation has taken place, the next step is to extract appropriate features for these image regions. Feature extraction algorithms play a very important role in image processing, and results for related applications are greatly affected by the way that visual features are extracted. Generally speaking, most previous content-based image retrieval systems extract both global features such as a color histogram, and local features including object shape, size, texture, position and so on (Datta et al., 2008). Most automatic image annotation algorithms adopt similar strategy. Proposed by Mori et al. (1999), an RGB histogram and a multi-resolution local energy histogram (computed by the local density after a Sobel filtering) are used to represent the image regions. In more recent work (Barnard and Forsyth, 2001; Duygulu et al., 2002; Barnard et al., 2002; Lavrenko et al., 2003; Blei and Jordan, 2003; Jeon et al., 2003; Lavrenko et al., 2004; Pan et al., 2004; Jin et al., 2004), frequently used visual features computed for each image region include region size, position, convexity, moment, average RGB value, average Lab value, standard deviation of RGB and Lab value, and oriented energy of different directions using Gabor filtering. These wide range of features contain as many aspects of images as possible, and is found in many approaches. In other work (Feng et al., 2004; Jeon and Manmatha, 2004; Carneiro and Vasconcelos, 2005; Wang and Li, 2002; Li and Wang, 2003, 2006) rectangular blocks are used as a proxy for the image regions, and only the RGB, Lab features, and texture energy (responses of Gabor filtering, DCT filtering or wavelet transformation) are considered. The most important features for representing images are color and texture information, including LUV, Lab color space and the output of wavelet transformation and Gabor filtering (Datta et al., 2005, 2008).

However, it is important to note that such low-level visual features are not enough to handle the great variance of objects appearing under different conditions. For instance, color and texture features alone are not discerning and robust enough to represent an object when its appearance has scale changes or certain transformations, e.g, when it is shot from different angles and distances, or in different illuminating conditions.

This indicates the demand for a more discriminative and robust representation. Rather than extracting features from everywhere of the image, the new idea is to first detect points of interest which are supposed to be discriminative and, to some extent, invariant to scale, position, slight illumination changes or certain transformations (Lowe, 2004; Mikolajczyk and Schmid, 2003; Herbert Bay and Gool, 2008) and then extract features from the local regions around these points. The most popular implementation of this idea is the Scale Invariant Feature Transform (SIFT) algorithm (Lowe, 1999, 2004). The algorithm first samples an image with the Difference-of-Gaussians point detector at different scales and locations (see Figure 2.4(c)). Importantly, this detector is not sensitive to small changes of affine transformation, scale, rotation and illumination. Each detected region is represented with the SIFT descriptor which is a histogram of edge directions at different locations and scales (more details will be given in Chapter 5). SIFT features have been shown to be superior to other descriptors (Mikolajczyk and Schmid, 2003) and are considered the state of the art in object recognition (Bosch, 2007).

Another related issue concerns the fact that the visual features do not naturally occupy a discrete space as words do. To render image regions more word-like, some existing algorithms (Mori et al., 1999; Duygulu et al., 2002; Barnard et al., 2002; Jeon et al., 2003; Jeon and Manmatha, 2004; Pan et al., 2004; Jin et al., 2004) cluster their features into a certain number of blobs (e.g., using K-means). Ideally, each cluster should represent a meaningful unit, however, in practice nonsensical blobs are often created. An alternative is to directly build an image annotation model on the extracted continuous visual features and use a Gaussian kernel distribution to model the image regions which improves the overall performance (Lavrenko et al., 2003; Wang and Li, 2002; Blei and Jordan, 2003; Feng et al., 2004; Lavrenko et al., 2004; Li and Wang, 2003; Carneiro and Vasconcelos, 2005). In sum, if the model requires image regions to correspond to visual terms (in an analogy to words), then it might be a good idea to cluster the region features into a discrete space but with proper feature representations (i.e., color features are better to characterize natural image regions), otherwise it is preferable to work with their original continuous space.

**Vocabulary Construction** When large scale image annotation data is created manually, it is difficult to guarantee that annotators will consistently use the same word to describe a given visual object. It is often the case that humans use synonyms or semantically similar words, e.g., *boy/teenager*, *automobile/car*, *ox/bull*, or *train/locomotive*, which increase the vocabulary and lead to data sparseness. One solution is to cluster semantically related words into categories (Duygulu et al., 2002; Wang and Li, 2002; Li and Wang, 2003, 2006; Barnard et al., 2002). Another solution is to construct a hierarchical structure to describe the relations among the words using WordNet (Miller, 1995) or other structured lexicon databases (Fan et al., 2007).

Beyond synonyms, it is important to decide whether all words should be modeled. This is rather difficult under the machine translation based framework mentioned in Section 2.3.2.1, since for some underlying abstract concepts (e.g., *meeting*, *run* or *peace*) it is very hard to find corresponding regions in the images. However, this is feasible under models with loose constraints for the mapping between visual and textual modalities (Barnard et al., 2002; Lavrenko et al., 2003; Feng et al., 2004) which can capture the correlations among regions or words.

So far, much effort has concentrated on modeling object names (e.g., *car*, *tiger*). It is of interest and importance for future investigation to extend the annotation task to proper names, including locations and person names as well as abstract nouns and even events. Person names are particularly challenging (Deschacht and Moens, 2007). Because if one solely relies on face recognition, it is very hard to accurately identify the face corresponding to the person using small, low resolution images.

**Data Acquisition and Annotation** Similarly to other image processing and computer vision problems, the diversity of real world images entails that the training data should be large and varied so as to train the image annotation models reliably.

However, the fact that the current training data are manually annotated entails that it is not economic to have large amounts of high quality training data. This is a labor intensive and thus very expensive task that must be repeated for every new database.[3] Furthermore, most existing algorithms require a strongly labeled training database with all semantic concepts appearing in the image correctly labeled. However, it is almost impossible to construct a large real-world database with such high standards. The most popular database used in image annotation research is the Corel data set. It contains hundreds of CDs, each with 100 images representing one distinct topic such as "*man*",

---

[3] Although there is an increasing trend to collect annotations using online worker communities, such as Amazon Mechanical Turk (Amazon, 2009), the quality of such data still remains an unsolved issue, especially in computer vision where workers may not select the best label when facing thousands of candidate words, or different workers may give totally unrelevant keywords for the same objects due to their different background knowledge (Russell et al., 2008).

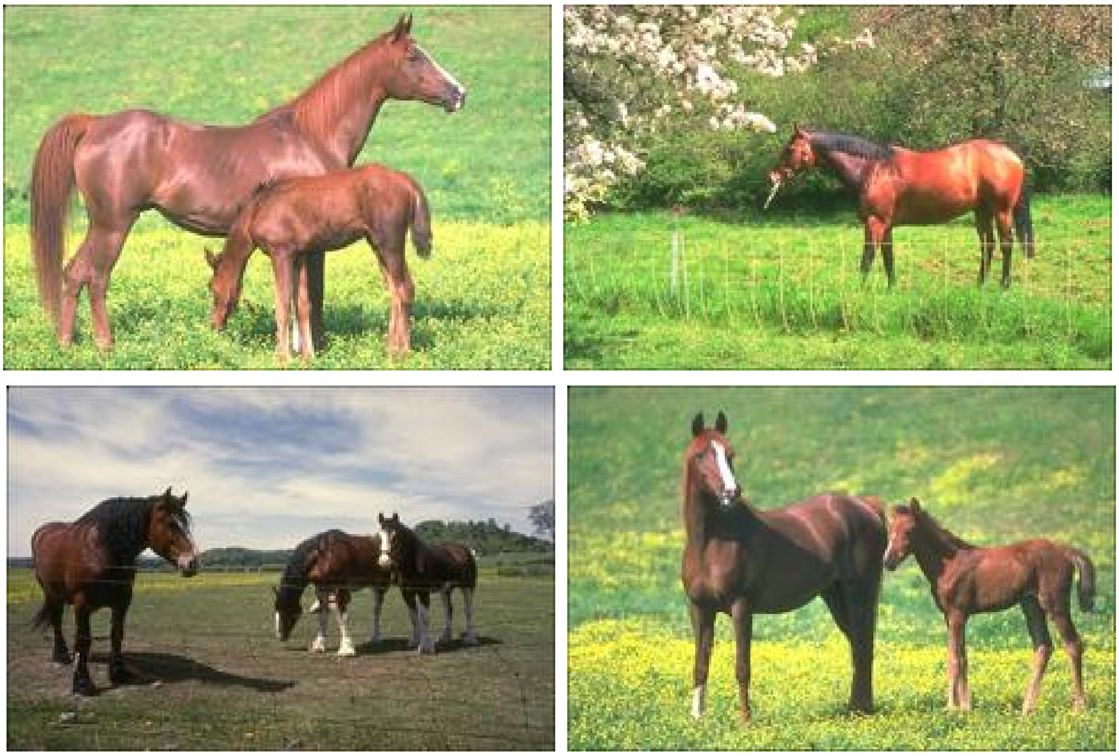


Figure 2.5: Examples from Corel imageset, exemplifying the concept *horse* with keyword descriptions *horse, grass, sky*

"*sunset*", "*horse*". Generally, images in the Corel data set depict less than five objects (Wang and Li, 2002) with 1-5 keywords per image (Feng et al., 2004). Corel contains clusters of many closely related images which in turn share keyword descriptions (see examples in Figure 2.5), thus allowing models to learn image-keyword associations reliably (Tang and Lewis, 2007). Although useful for comparing image annotation models, the Corel dataset is considered relatively simpler than what would be representative of real-world applications. There is another data set, NIST's Video TREC, which contains 12 30-minutes video sections of CNN or ABC news from which 5200 key frames have been extracted, and partially annotated in a hierarchical way by the participants of TREC (Feng et al., 2004). The TREC database is more difficult than Corel because it contains more complex scenes and more objects (Feng et al., 2004). But both databases require manual annotation and both of them are currently weakly labeled.

## 2.4 Text Summarization

Recall that extracting image content is the first step towards generating an image description. Next, we must render this content into natural language. This task mainly involves natural language generation, i.e., producing natural language outputs from non-linguistic inputs. Generally, a background knowledge base is required to structure the image content and specify relations among these contents whereas a surface realizer (either using templates or grammars), is employed to produce the description accordingly.

In this thesis, we adopt a knowledge-lean approach for this task, which means that we will not utilize manually created rules, grammars or sentence templates. Recall that we focus on news image caption generation, where a news image is available together with its associated document and the task is to automatically generate a caption for the news image. Without access to the associated articles, we would be exposed to a traditional NLG problem (Yao et al., 2009). To some extent, the accompanying text documents allow us to perform the caption generation task in a fashion akin to text summarization and realize the caption generation task without human involvement. Specifically, during generation, we have the extracted image content and the associated text document at hand, and our task is to create a summary from the associated document given the image content. Analogously, traditional summarization system is expected to condense a source text, containing one or more documents, into a text shorter in size that still conveys the main information in the original, without, however, taking any visual information into account.

Compared to general NLG systems, summarization models rely less on human labor in many aspects. Firstly, the source text provides a fertile field for lexical choices. Word morphology and necessary function words are often manually imported into an NLG system as rules (Knight and Hatzivassiloglou, 1995; Langkilde and Knight, 1998). With respect to grammaticality, most traditional NLG systems either adopt sentence-templates or rely on predefined grammars to create human-readable sentences (Jurafsky and Martin, 2000). In summarization, the source text naturally supplies grammatical sentences or phrases that can be used to produce grammatical summaries. Additionally, sentence templates or similar paradigms are not suitable for general-purpose generation applications due to their inability of generating diverse, expressive and flexible sentences in previously unseen domains.

A successful summarization system should be able to automatically decide what to say in a limited space, while keeping sentences grammatical, and making sure coherence is preserved when multi-sentence output needed (Mani, 2001). We will look at the current developments in the field of text summarization and review both extractive and abstractive models. Extractive approaches create summaries by reusing pieces of the source document, e.g., sentences or paragraphs, while the abstractive approaches tend to use paraphrases and re-organize the main topics of the source document in a more succinct manner.

**Extractive Summarization**[4] Much work to date in summarization places more emphasis on what to say rather, than how to say it. In other words, automatic text summarization is currently dominated by extractive approaches, where a summary is created simply by identifying and subsequently concatenating the sentences from the original text. A key problem in text summarization is how to represent the content of documents and accordingly find appropriate criterions to select the most informative sentences in order to build the summary.

Existing document content representations include simple unigram frequencies (Nenkova and Vanderwende, 2005; Vanderwende et al., 2007; Haghighi and Vanderwende, 2009), syntactic features (such as clause information or dependencies) (Barzilay and McKeown, 2005), and topic representations (Daumé III and Marcu, 2006; Haghighi and Vanderwende, 2009). For non-probabilistic representations, distance measures (e.g., Euclidean distance, or cosine distance) are often used as selection criterions while divergence-based measures are used for probabilistic ones (e.g., topic representations) (Haghighi and Vanderwende, 2009).

SumBasic, proposed by Nenkova and Vanderwende (2005), is a simple but effective extractive model for multi-document summarization. The algorithm is based on the observation that content words used frequently in a document set are likely to appear in human-written summaries. Therefore a sentence is scored according to how many frequently appeared content words it has, and the summary is built up by these highest scored sentences, together with measures to penalize repeated content. This algorithm utilizes the unigram frequencies over the document set, but ignores the fact that a word appearing many times in only one document could have totally different status in a summary compared to a word appearing the same number of times evenly across the whole document set.

---

[4] It is outside the scope of this thesis to review all current work in summarization. Interested readers may refer to Mani (2001) and Spärck Jones (1999); Spärck Jones (2007) for an overview on text summarization. Here, we concentrate on approaches related to our own work.

Daumé III and Marcu (2006), also Haghighi and Vanderwende (2009), propose models based on the idea that the document's content can be represented by the distribution of topics corresponding to coarse-gained categories such as education, sports, and so on. These topic distributions are modeled though topic models, e.g., variants of Latent Dirichlet Allocation (LDA, Blei et al. 2003), and sentences whose topic distributions are most *similar* to the whole document set are selected as the summary. These models work considerably well on the multi-document summarization task of the Document Understanding Conference (DUC 2006).

Without a great deal of linguistic analysis, extractive models are able to output grammatical summaries for a wide range of documents, independently of style, text type, and subject matter. Unfortunately, these selected sentences are usually verbose and contain redundant or irrelevant information. A manual investigation conducted by Jing and McKeown (2000) shows that the output of extractive models is very different from human written summaries. They manually examined 30 articles with corresponding human-written summaries and found that most of the latter could be decomposed into pieces of original article text, which, accordingly, motivated them to propose a cut-and-paste style summarization model. Instead of extracting sentences from documents, humans often construct a brand-new sentence according to their background knowledge or simply extracting useful constituents (words or phrases) across the whole documents and arranging them into a grammatical and coherent summary.

**Abstractive Summarization** In contrast to extractive approaches which have been widely used in the past decades, abstractive methods have received less attention. This is for a good reason as substantially more effort must be invested in creating fluent and grammatical summaries over and above identifying the right content. Abstractive models, in most cases, first identify the key content of documents in the form of constituents, e.g., words or phrases, which are then organized into a grammatical sentence. Identifying content units bears some similarities with extractive models, where various keyword extraction or topic detection approaches are adopted to select candidate words or phrases, which must then be reordered. Candidate constituents may be also deleted or paraphrased in order to generate informative and concise summaries.

To handle to the grammaticality issue, some approaches utilize statistical language models to recover local word order (Witbrock and Mittal, 1999; Banko et al., 2000; Jin and Hauptmann, 2001a, 2002), while others depend on human heuristics and syntactic information (Zajic et al., 2002; Zhou and Hovy, 2003; Dorr et al., 2003; Bonnie et al., 2004). The former may create more varied sentences as there are less constraints involved, while the latter typically produce more grammatical realizations, yet require manually created rules (these are much smaller in size than the knowledge bases used in a original NLG system).

Abstractive models usually produce more condensed sentences with controllable length. The cut-and-paste style gives abstractive models considerable freedom in lexical choice. And this also makes them portable across different summarization scenarios. However, it is commonly agreed that generating grammatical sentences from scratch with limited human involvement remains a difficult problem, especially if one relies solely on a language model with limited history length. One compromise could be to lower the expectations for the output, e.g., the sentences will not be entirely grammatical but readable enough to express the gist of the story in the original article. Headline generation is such a task that generates a very short title-like summary for a given document (Witbrock and Mittal, 1999; Banko et al., 2000; Jin and Hauptmann, 2001a,b, 2002; Zajic et al., 2002; Zhou and Hovy, 2003; Dorr et al., 2003; Bonnie et al., 2004). For example, news headlines are usually less than 10 words long and highlight the most important part of the news stories. Creating very short single-document summaries was one of the summarization tasks in the Document Understanding Conference[5](DUC 2004).

Banko et al. (2000) propose a bag-of-words model for the task of headline generation. Their model selects content words according to the probability of a word appearing in the headline given that the same word appears in the corresponding document and is independent from other words in the headline. A bigram language model is adopted to score and rank possible realizations. The length of the headlines is also considered probabilistically to favor outputs of reasonable length (around 5 words). The model has a clean structure and utilizes a language model to generate acceptable word orders.

However, Jin and Hauptmann (2002) claim that using a language model solely for word order purposes will favor some *common* uninformative words to appear in the headlines and even negatively influence the inclusion of other more important words in the headlines. They propose to eliminate this bias by subtracting the *commonness* of headline words from the language model score, where *commonness* is modeled according to how often a word appears in the headlines of the whole corpus. Their human experiment shows that the headlines generated by the model taking *commonness* into

---

[5] `http://duc.nist.gov/duc2004/`

account are more readable than the ones produced by the model without considerations about *commonness*.

Another possible way of dealing with word order is through heuristics. Zhou and Hovy (2003) first identify keywords according to frequency, position, and probabilistic features (the ones used by Jin and Hauptmann (2001a)), then expand these keywords with their near bigram neighbors. A *window* that contains the largest number of selected words is considered as the headline and a set of hand-written rules are used to post process the headline. Other approaches include sentence compression which helps make the summary more concentrate on the gist hence avoid redundancies (Knight and Marcu, 2002; Clarke, 2008), and sentence fusion which provides a flexible framework for both single- and multi- document summarization by taking syntactic information into account (Barzilay and McKeown, 2005; Filippova, 2009).

## 2.5 Summary

In this chapter, we reviewed related work on image and graphics description generation. Despite their differences, both applications rely on the background knowledge bases containing correspondences that help interpret visual information into textual concepts, and fine-gained sentence templates or grammars. The development of such knowledge bases usually requires significant human involvement. For example, in order to obtain an abstract representation from various raw image features, one has to retrieve them from a visual-textual correspondence dictionary which is typically manually created beforehand. Analogously, during the generation step, various sentence templates or grammars in the desired domain need to be created by human annotators.

Obviously, the heavy reliance on hand crafted knowledge greatly limits further development and applications of current image description generation models. In this thesis, we will look at the task in a knowledge-lean way, and utilize resources where images and their captions co-occur naturally. Specifically, we will concentrate on news articles, their images and captions. Such data is admittedly noisy but cost-free.

We also presented an overview of previous approaches for automatic image annotation and text summarization with emphasis to those most relevant to our task.

In the rest of the document, we will first introduce the news data we employ throughout, and then address the problem of automatically learning the visual-textual relations from this noisy database. Finally, we formulate our generation task in a text summarization framework which is less knowledge intensive compared to traditional

NLG models.

# Chapter 3

# Data Description and Problem Formulation

In the previous chapter, we reviewed work related to image caption generation. A common theme across existing related methods is the reliance on manually created knowledge bases. Specifically, they often employ a hand-crafted dictionary of visual-textual correspondences in order to extract the concepts present in the image, and rely on predefined sentence-templates or grammars to render the extracted image content into nature language. Manually building knowledge bases for both content extraction and surface realization is time-consuming, costly and has to be repeated for new domains. Therefore, it is of great importance to look at knowledge-lean ways to perform the task, which do not rely on the availability of large scale multimodal correspondence dictionaries or hand-written sentence-templates and grammars. In our work, we learn this information automatically from data.

The availability of appropriate image datasets is therefore crucial for our task. An ideal dataset should (1) be representative of real-world data, (2) relatively easy to collect as we hope to rely on minimal or no human involvement, (3) include images with annotations that will potentially supply visual-textual correspondences, (4) contain auxiliary information that could allow us to mine related linguistic information in order to help us create human readable descriptions, and (5) contain gold standard captions for evaluating the output of our system.

Existing image datasets used in computer vision or image retrieval related areas are many, however, not readily suitable to our task. Most of them are built for traditional image annotation (such as the Corel dataset), image segmentation, object recognition (Martin et al., 2001; Fei-Fei et al., 2004; Griffin et al., 2007; Schroff et al., 2007;

Russell et al., 2008), and other more sophisticated tasks (Russell et al., 2008; Yao et al., 2009; Barnard et al., 2008; Deng et al., 2009). These datasets consist mainly of images and annotation keywords. The former are usually pictures with one or two prominent objects in the center of a relatively simple background, whereas the latter are object names ranging from 20 to 300 in total. Object contours are sometimes outlined in the datasets used for image segmentation and object recognition. Some datasets even provide semantic relationships among segmented image regions (Yao et al., 2009). However, it should be noted that nearly all of these datasets are either post-processed or entirely built by human annotators. Existing datasets can be used to learn the correspondences between images and words for the limited number of objects for which annotation keywords are available. However, they do not contain gold standard caption annotations.

In this chapter, we first survey existing datasets in computer vision and related areas, and different paradigms for collecting data. We then motivate why we harvest our database from a news website, introduce our BBC news image dataset and discuss its properties. Accordingly, we recast our task in view of the BBC news image dataset.

## 3.1 Image Datasets

The availability of high quality image datasets has been a central issue in many computer vision and image retrieval applications, such as image classification, annotation, segmentation, object recognition, and so on. Generally speaking, each entry in an image database will consist of an image and its labels which refer to object names or other object related information, such as the contours, or locations of the objects and, possibly, how they relate to each other in the form of image regions. Image databases containing more complex scenes, objects under different appearances, or even more detailed annotations would prompt further research in object recognition and related areas. Recently, a handful of image databases have emerged and are being widely used as benchmark datasets in the object recognition and image annotation communities.

**Datasets with Light Annotation** Current image datasets used for image classification or vanilla image annotation tasks usually require light annotations for each image. In other words, an image is often labeled with keywords only, referring to the scene it captures (e.g., events or locations), or the names of objects it depicts. These datasets can be obtained by automatic methods aided with human post-processing (e.g., clas-

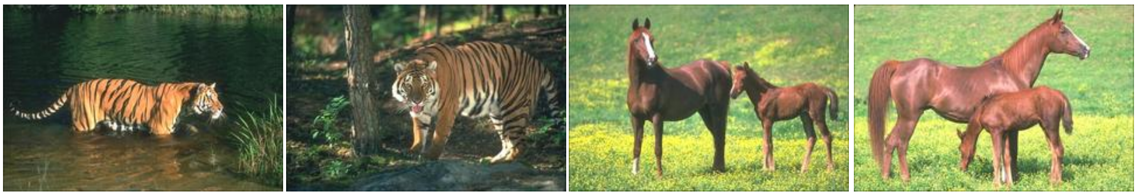


Figure 3.1: Images from the Corel database; the first two are in the theme *tiger*, which is assigned with keywords: *cat, tiger, grass, forest*, while the last two images are in the theme *horse*, with keywords *horse, grass, sky*.

sifiers combined with human post-cleaning). For example, in the databases used for image classification tasks, only one word (e.g., indoor, outdoor, urban, countryside, or natural), is given to each image (Vailaya et al., 1999, 2001). The vocabulary is small (usually less than 20 words ) and the labels are usually more general terms rather than specific object names.

The databases used for vanilla image annotation contain more keywords referring to object names or other abstract concepts. The Corel database has been extensively used as a testbed for the development and evaluation of image annotation models. It is a collection of stock photographs with a simple labeling scheme. Figure 3.1 shows examples from the Corel dataset: the first two images are from theme *tiger*, and the last two are from theme *horse*. Note that the images have a relatively simple background and a prominent animal in the foreground. Unfortunately, this database is not representative of real-world image collections. It has a limited number of themes with many closely related images which in turn share keyword descriptions. It is therefore relatively easy to learn the associations between images and keywords and do well on annotation and retrieval tasks (Tang and Lewis, 2007; Westerveld and de Vries, 2003).

Given the relatively simple labeling protocol, it is also feasible to utilize existing image search engines combined with either manual cleaning or automatic approaches (which, to some extent, still need manually created training data). If necessary, this process can be performed iteratively for better results. For each word in a set of predefined object names, Schroff et al. (2007) obtain a candidate pool of images with their original webpages from existing image search engines by using the object name as a query. And a classifier trained on manually labeled textual data (such as the surrounding text, in-out web links and metadata, such as filenames, image tags, etc) is used to filter out symbolic and abstract images (e.g., comics, maps and plots). Afterwards, more content related textual features are applied to re-rank the filtered candidate pool.

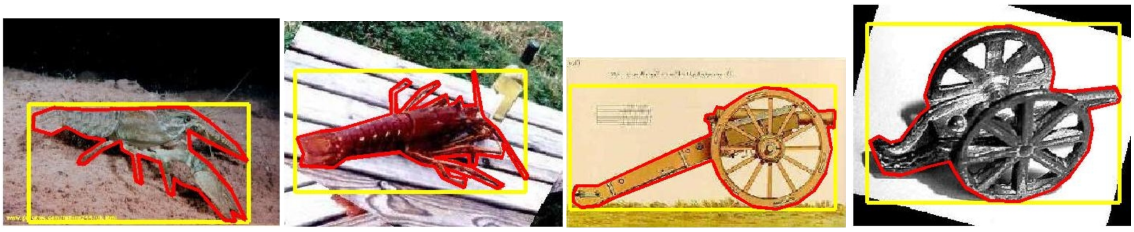


Figure 3.2: Images from the Caltech 101 database; the first two are labeled with the word *crayfish*, while the last two images are labeled with *canon*. The yellow rectangle outlines the bounding box for each object, and red line gives the contour of the object.

High ranking images are further used as positive samples, together with randomly selected negative ones, to train a visual feature based classifier in order to further clean the pool of the given query. This semi-automatic approach reduces the amount of human involvement but produces relatively coarse annotations, usually one word per image.

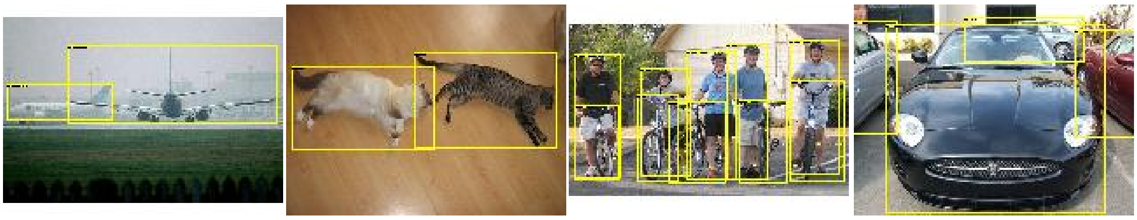
**Datasets with Detailed Annotation** There also exist image datasets that contain more detailed information about the depicted objects. For instance, in the Caltech 101 dataset[1] (Fei-Fei et al., 2004) (see Figure 3.2 for example), objects are annotated with keywords and in addition, the main object is outlined with a bounding box. Its contour is also provided and all this information is entered manually. As shown in Figure 3.3, the PASCAL Visual Object Classes (VOC) Challenge 2007 dataset[2] (Everingham et al., 2007) and ImageNet[3] (Deng et al., 2009) have their objects manually labeled by bounding boxes. Datasets used for object detection, recognition and image segmentation often require such detailed annotations since these applications need not only the object names, but also the accurate contours and locations of the objects (Fei-Fei et al., 2004; Griffin et al., 2007; Deng et al., 2009; Russell et al., 2008; Martin et al., 2001; Barnard et al., 2008; Yao et al., 2009; Everingham et al., 2007). Due to the detailed annotation requirements (determining bounding boxes or contours for objects), most of these databases either adopt an automatic approach combined with human post-processing, or completely rely on human annotations, and most of their vocab-

---

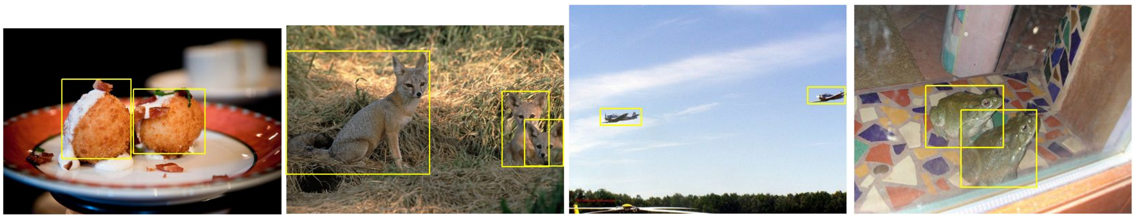
[1] The Caltech 101 dataset can be found in `http://www.vision.caltech.edu/Image_Datasets/Caltech101/`, and the Caltech 256 dataset in `http://www.vision.caltech.edu/Image_Datasets/Caltech256/`

[2] `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/`

[3] `http://www.image-net.org/`



airplane cat bike people car



croquette fox airplane frog

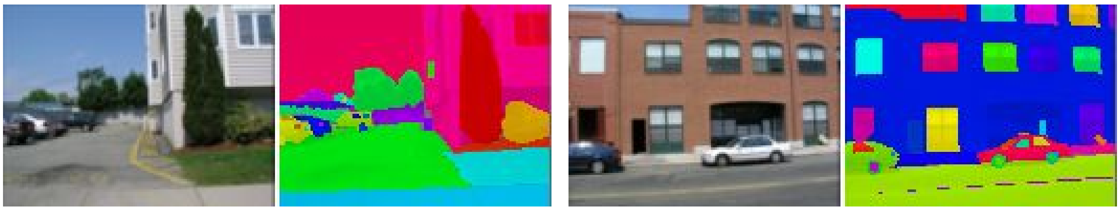
Figure 3.3: The top row shows images with bounding boxes from the PASCAL VOC 2007 dataset. The bottom row shows images from ImageNet. Images in both datasets are annotated with object names and bounding boxes. The PASCAL VOC2007 dataset tends to include more cluttered images but contains only 20 object categories while ImageNet has a vocabulary of several thousand words.

ularies are, therefore, only a few hundred words[4] (Fei-Fei et al., 2004; Griffin et al., 2007; Martin et al., 2001; Everingham et al., 2007). Also note that some of datasets for object recognition only annotate one object for each image since they only focus on the most prominent object (foreground) while ignoring the background (Fei-Fei et al., 2004; Griffin et al., 2007; Schroff et al., 2007).

Fei-Fei et al. (2004) and Griffin et al. (2007) utilize existing image search engines available on the internet to collect related image candidates according to specific queries from a predefined vocabulary, then manually filter the search results and further provide more detailed annotations (e.g., bounding box or contours of objects). The former create a database for 101 objects with around 9,000 images while the latter contains over 20,000 images for 256 objects. In a similar framework, Deng et al. (2009) collect a much larger dataset, ImageNet, with 5,000 synsets over 3,000,000 images. The manual cleaning procedure here is achieved through Amazon Mechanical Turk (Amazon, 2009). This data collection project is ongoing and their ultimate goal is to created a visual version for WordNet (Miller, 1995), with the visual synsets corresponding to WordNet synsets.

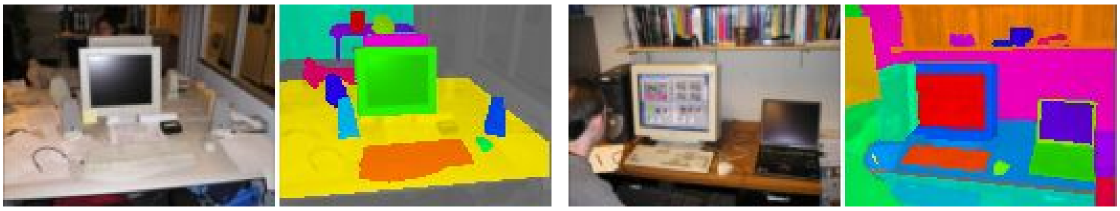
---

[4] ImageNet has 5,000 synsets now, and is still growing.



car sky grass tree building fence

car sky building road window gate



screen mouse speaker keyboard desk

laptop desk screen mouse keyboard man

Figure 3.4: Images with annotations (abridged version) from the LabelMe dataset. Each image is accompanied with a segmentation mask and corresponding keywords labeled by internet users. LabelMe allows users to annotate an image with any words they think appropriate, so there are repetitions of keywords in most images (e.g., *grass*, *grassland*).

Interestingly, a new web based collaborative paradigm has been used to successfully build several large scale image databases (Russell et al., 2008; von Ahn and Dabbish, 2004). The ESP game[5] proposed by von Ahn and Dabbish (2004) draws internet users into a game scenario, essentially an object recognition task. Data is collected by randomly pairing two internet users and encouraging them to “guess each other’s mind” for a question about the content of a given picture. This effort has collected more than million images with annotation keywords. Russell et al. (2008)[6], in a similar fashion, collected a large scale image database (LabelMe) with detailed object segmentation information from the web. They made an annotation tool and their images publicly available and asked internet users to label the objects with keywords and outline their contours (examples of LabelMe are illustrated in Figure 3.4). The LabelMe dataset contains over 20,000 images or video frames with over 3,000 descriptions (around 200 object categories) in total.

The LHI dataset[7] (Yao et al., 2009) hosts over a million images and video frames with hundreds of objects. These are deep segmented, and parsed in a top-down fashion.

---

[5] `http://espgame.org/gwap/`

[6] `http://labelme.csail.mit.edu/`

[7] `http://www.imageparsing.com/`

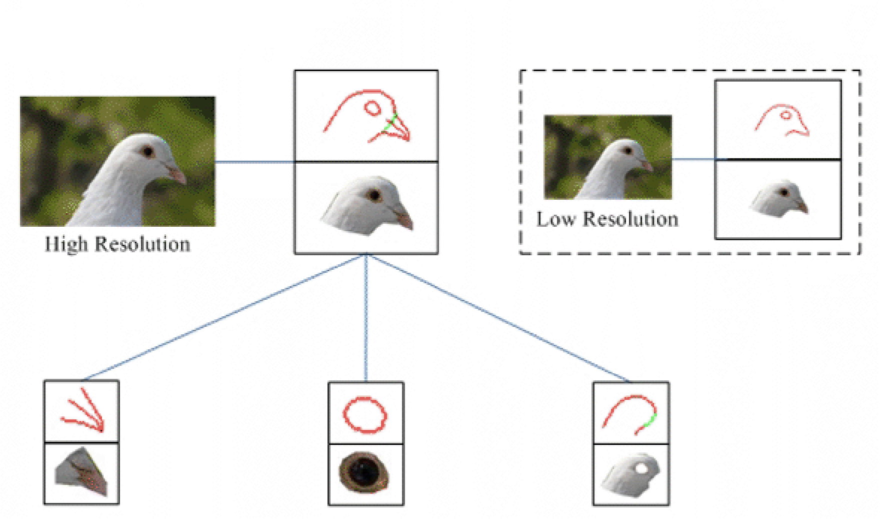


Figure 3.5: An image with deep segmentation, deep parsing and semantic relations from LHI image dataset. This top-down hierarchy shows that a pigeon head consists of eye, beak, and skull.

Furthermore, regions are associated with semantic relations taken from a background knowledge base (see Figure 3.5 for example). The segmentation, parsing and association data were manually created by a full-time 23-annotator team spanning 2 years. The LHI database contains the most detailed annotations to date ranging from simple object localization to naming, segmentation and top-down hierarchical relations.

The image databases used for object recognition and related applications are usually annotated in several respects with information about the objects, their locations, and often contain bounding boxes or contours of the objects. Furthermore, in most cases, all depicted objects are labeled irrespectively of size and whether they appear in the foreground or background (especially in datasets for image segmentation). This type of dataset unavoidably require substantial human involvement and does not scale across different types of images and tasks.

## 3.2 Datasets Created from News Resources

Previous work on automatic image description generation deals with images or video frames shot in specific conditions, with a limited vocabulary (e.g., object names) and fine-gained knowledge bases (e.g., Kojima et al. (2002) construct a knowledge base that maps the moving trajectory of the human head into abstract actions, such as *enter*, *exit*, and *sit down*). In contrast, in this thesis, we are aiming to explore knowledge-lean ways for extracting the image content and rendering it appropriately. Thus, visual-textual correspondences as well as grammar constraints will be learned from data. We will also focus on real-word images and employ a larger vocabulary. Compared to the Corel and Caltech datasets discussed above, real-world images are more challenging as they contain more objects embedded often in cluttered scenes.

Since we are going to generate captions for images, and not just isolated keywords, we expect our vocabulary to contain not only object names, but also other words, such as abstract concepts, verbs, adjectives, (e.g., *meeting, run, game*) and function words. Another unavoidable issue is that, to some extent, we need gold standard captions for tuning our models and evaluating their output. However, existing image datasets used in computer vision or image retrieval are not readily suitable for the caption generation task.

First of all, a major concern is the absence of rich linguistic information. Detailed semantic relations, in the form of a knowledge hierarchy are specified among image regions in the LHI dataset, however, these alone can not be used to create human-readable descriptions without the help of additional sentence-templates or grammars that are usually created manually.

Secondly, although the annotation keywords (usually referring to object names) together with the images in existing datasets can be used to learn the visual-textual correspondences, most of their vocabularies are relatively small (around 300) and focus solely on object names or categories. For example, Caltech 101 and Caltech 256 only have 101 and 256 object names, respectively. These datasets contain several types of annotations, but gold standard captions are notably missing. Unfortunately, many existing annotations such as object positions and bounding boxes or contours are not relevant to our task. Note that the LHI database contains information that is potentially useful for text generation (e.g., spatial relationships among object regions and image parsing information). However, its domain specificity (e.g., its focus on video surveillance) makes it difficult to use for our caption generation task.

Last but not least, our aim is to generate image captions in a learning-from-data fashion, and thus make use of as little manually created knowledge as possible. All existing databases are either manually filtered or post-processed intensively, and therefore, hard and expensive to adapt to our requirements (e.g., by hiring annotators to write captions for a large dataset).

**News Image Resources** Although existing image databases are not exactly suitable for our task, their smart use of web resources encourages us to explore the large amount of data available on the internet.

Given all the issues discussed above, we aim to relieve the data acquisition bottleneck associated with image related applications by taking advantage of publicly available resources where images and their textual descriptions co-occur naturally. News articles associated with images and their captions spring readily to mind (e.g., BBC News, Yahoo! News, and CNN News). Many on-line news providers supply pictures with news articles, some even classify news into broad topic categories (e.g., business, world, sports, entertainment, technology, etc). News images often display several objects and complex scenes, and are usually associated with captions describing their contents. Captions are slightly different from image descriptions; they can be denotative (describe the objects the image depicts) or connotative (describe sociological, political, or economic attitudes reflected in the image). However, they are image specific and employ a rich vocabulary, which is in marked contrast to the previous databases. For example, in Corel, Caltech 101 and Caltech 256, images contain one or two salient objects and a limited vocabulary (typically around 300 words or less).

So, rather than laboriously annotating images with their keywords, we could simply treat the caption words as the image annotation keywords. These annotations are admittedly noisy and far from ideal, but the co-occurrence of the images and caption words suggest the possibility of modeling the correlations between the visual and textual modalities with less supervision. Importantly, the news images and captions are not standalone, they come with news articles whose content is shared with the images, and which can also contribute to modeling the multimodal correlations. Besides, the rich linguistic information present in the news documents allow us to gather syntactic and lexical knowledge necessary for generating human-readable descriptions.

Figure 3.6 shows two example webpages from BBC News website. The news image on the right shows a girl sitting in front of a computer, with a caption reading *“Children were found to be far more internet-wise than parents“*, and the article talks about the technological gap between kids and their parents. The caption here is not a set of isolated annotation keywords; the words *children* and *parents* could be considered as literal annotations (although parents do not appear in the image), while *internet-wise* and *found* are connotative since there are no specific regions in the picture that *depict* these words. The left example shows a policeman hitting the protestors, with the caption *”Students accused the police of brutality.“*, where *students* and *police* are literal while *accused* and *brutality* are connotative.

## 3.3 BBC News Dataset

In what follows, we present a new database consisting of news articles, images, and their captions which we collected from the BBC News website[8] from 2006-07-11 to 2006-10-19. Specifically, we downloaded 3,361 news articles from this site and created a database where each news article is accompanied with an image and its caption (as shown in Figure 3.6). The dataset covers a wide range of topics including national and international politics, technology, sports, education, etc.

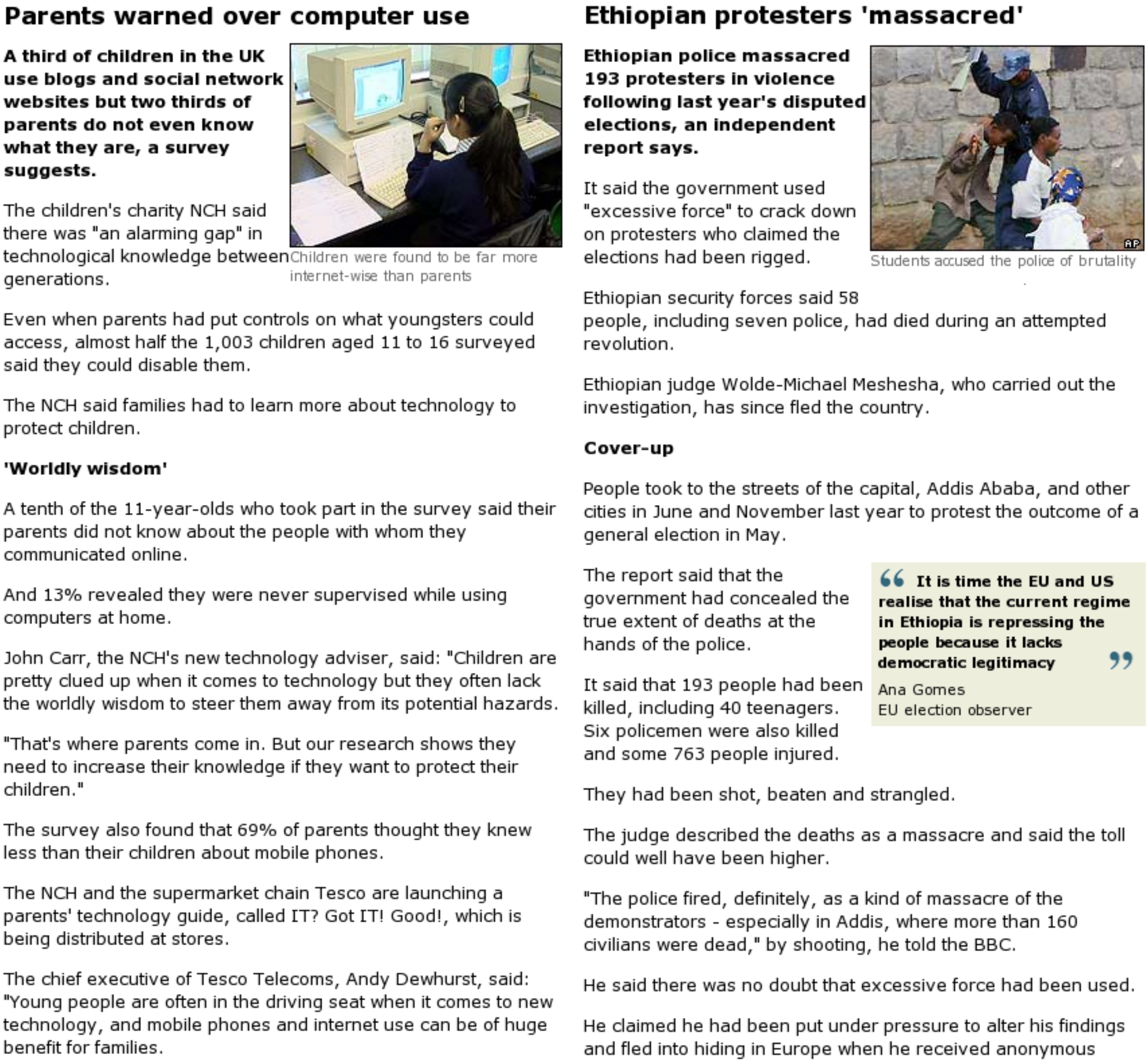
News articles normally use color images which are around 200 pixels wide and 150 pixels high. The average caption length is 9.5 words, the average sentence length is 20.5 words, and the average document length 421.5 words. The caption vocabulary is 6,180 words and the document vocabulary is 26,795. The vocabulary shared between captions and documents is 5,921 words.

In contrast to existing image databases, our dataset contains more challenging images. The latter come with captions that can be considered as noisy annotations while traditional image databases are manually cleaned or post-processed resulting in explicit annotation keywords. Our captions are image specific and focus on the news event that is shared between the news image and document, whilst traditional datasets tend to annotate objects with details for isolated regions, e.g, the positions and contours of objects, which are not relevant to our task. The vocabulary employed in our dataset includes not only object names but also abstract concepts and beyond nouns all other parts of speech. Furthermore, it is much larger in size compared to traditional databases which only focus on few object names, around 300. Importantly, our dataset has a unique component, the news document, that is not available in existing datasets.

In our dataset, the news image, its caption and accompanying document co-occur naturally in order to tell readers a news story. This interplay among different modalities will offer readers complementary views on the same or similar events. In our case, the news article contains detailed information covering the five W's (who, what, when, where, why), and often provides the necessary backdrop against which to interpret the image and its caption. News images are usually selected by journalists based on the document with the aim of highlighting important aspects of the story. Along with news titles, the first sentences and section heads, image captions are the most commonly read

---

[8] `http://news.bbc.co.uk/`



## Parents warned over computer use

**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.**

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

The children's charity NCH said there was "an alarming gap" in technological knowledge between generations.

Even when parents had put controls on what youngsters could access, almost half the 1,003 children aged 11 to 16 surveyed said they could disable them.

The NCH said families had to learn more about technology to protect children.

**'Worldly wisdom'**

A tenth of the 11-year-olds who took part in the survey said their parents did not know about the people with whom they communicated online.

And 13% revealed they were never supervised while using computers at home.

John Carr, the NCH's new technology adviser, said: "Children are pretty clued up when it comes to technology but they often lack the worldly wisdom to steer them away from its potential hazards.

"That's where parents come in. But our research shows they need to increase their knowledge if they want to protect their children."

The survey also found that 69% of parents thought they knew less than their children about mobile phones.

The NCH and the supermarket chain Tesco are launching a parents' technology guide, called IT? Got IT! Good!, which is being distributed at stores.

The chief executive of Tesco Telecoms, Andy Dewhurst, said: "Young people are often in the driving seat when it comes to new technology, and mobile phones and internet use can be of huge benefit for families.

## Ethiopian protesters 'massacred'

**Ethiopian police massacred 193 protesters in violence following last year's disputed elections, an independent report says.**

Students accused the police of brutality

It said the government used "excessive force" to crack down on protesters who claimed the elections had been rigged.

Ethiopian security forces said 58 people, including seven police, had died during an attempted revolution.

Ethiopian judge Wolde-Michael Meshesha, who carried out the investigation, has since fled the country.

**Cover-up**

People took to the streets of the capital, Addis Ababa, and other cities in June and November last year to protest the outcome of a general election in May.

The report said that the government had concealed the true extent of deaths at the hands of the police.

> **It is time the EU and US realise that the current regime in Ethiopia is repressing the people because it lacks democratic legitimacy**
>
> Ana Gomes
> EU election observer

It said that 193 people had been killed, including 40 teenagers. Six policemen were also killed and some 763 people injured.

They had been shot, beaten and strangled.

The judge described the deaths as a massacre and said the toll could well have been higher.

"The police fired, definitely, as a kind of massacre of the demonstrators - especially in Addis, where more than 160 civilians were dead," by shooting, he told the BBC.

He said there was no doubt that excessive force had been used.

He claimed he had been put under pressure to alter his findings and fled into hiding in Europe when he received anonymous

Figure 3.6: Two example web pages from our BBC news dataset.

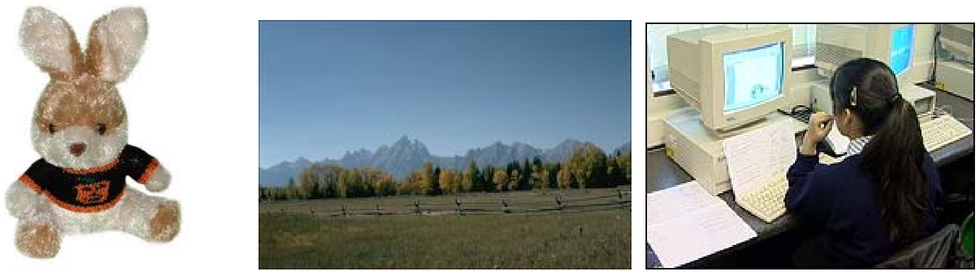


Figure 3.7: Images from a simple object dictionary (Héde et al., 2004), the Corel dataset, and BBC news dataset, from left to right, respectively.

parts in a news article. A good caption must be informative and easy to read, clearly identify the subject of the picture, establish the picture's relevance to the news article, provide background for the picture, and ultimately draw the reader into the article.

Most previous image datasets used in image description generation tasks contain pictures shot in uniform or fixed background, with one or two prominent objects in the foreground (see the left picture in Figure 3.7). News images deviate from this setting. They are usually cluttered depicting more complex scenes, contain more and less prominent objects, and are often rendered in low resolution. For instance, cluttered indoor scenes are common in news images (see the right picture in Figure 3.7) in contrast to the blue sky and green grassland frequently appearing in the Corel dataset (see the middle picture in Figure 3.7). A concern here is that detecting and recognizing all objects from images under such noisy conditions is still beyond the capability of current computer vision research. However, we do not claim that this image dataset is universally suitable for all computer vision related applications. Importantly, our aim is to employ this dataset to observe the correlation between the visual and textual modalities without explicitly performing object recognition.

In our setting, we consider the image captions as annotations for the news images. This is admittedly noisy since captions are not written in order to exhaustively list every object in the image but to describe the main event or related aspects thereof in the news document. To assess the level of noise in the dataset, we randomly selected 240 image-caption pairs and manually examined whether the content words (e.g., nouns, verbs, and adjectives) present in the captions could describe the image. We found out that the captions express the picture's content 90% of the time. Furthermore, approximately 88% of the nouns in subject or object position directly denote salient objects in the pictures.

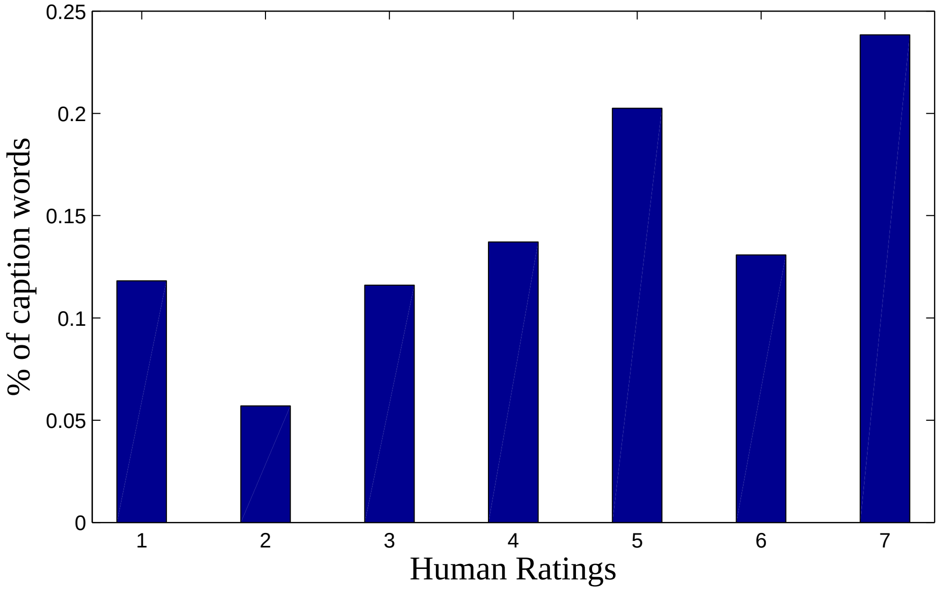


Figure 3.8: Histogram of the ratings for content words in the original image captions given by humans.

We also conducted a human study to assess the quality of the caption words being treated as annotation keywords. 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 describe the news image given the document. We used a $[1-7]$ rating scale, and encouraged participants to give high ratings for words which were closely related to both image and document. We randomly selected 30 document-image pairs and included the content words from their original human-authored captions. We collected ratings from 26 unpaid volunteers, all self reported native English speakers. We used WebExp (Keller et al., 2009) to conduct the experiment over the internet. Figure 3.8 shows a histogram of the relative number of caption content words using a 7-point rating scale. More than 71.5% of the content words in the original captions were given a rating of 4 or higher, which suggests that most caption words are perceived to be descriptive of the image content. We thus conclude that the captions contain useful information about the picture and can be used for our purposes. We will further validate this experimentally in next chapter.

Although text documents are absent in traditional image datasets, they form a crucial part in our database. The importance of including documents is twofold: firstly, news documents contain the necessary background knowledge which the image depicts or supplements. Secondly, the availability of news articles allows us to benefit from the rich linguistic information naturally embedded in the text, and approximate the natural language generation with methods akin to text summarization.

## 3.4 Assumptions and Problem Formulation

Since we are using a non-standard database, namely captioned images embedded in documents, it is important to clarify how this impacts our task. We thus make the following assumptions:

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 the right example of Figure 3.6, the caption mentions the *police* shown in the picture but also the brutality of the police, and the students' response to the police's brutality.
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[9].
3. Since our images are implicitly rather than explicitly labeled, we do not assume that we can extract *all* objects present in the image. Instead, we hope to be able to model event-related information such as "what happened", "who did it" and "where" with the help of news document. Our annotations are therefore more semantic in nature than traditionally assumed.

**Problem Formulation** Given our dataset and the assumptions above, we are now ready to recast the image caption generation task as follows:

**Definition 3.** *Given a news image I, and its associated news document D, create a natural language caption C that captures the main content of the image given D.*

Our training data thus consists of document-image-caption tuples like the one used in Figure 3.6. During testing, we are given a new document and an associated image for which we must generate a caption. This task includes two subtasks, namely extracting the image's content and rendering it into human-readable form. We consider the

---

[9] As we mentioned in previous section, in our pilot experiment, about 88% of salient objects have been specified in the image captions

original image captions as a gold standard. Specifically, we are not going to propose a state-of-the-art image annotation model which enumerates *every* object , nor do we create a literal sentence that lists all objects in the image.

## 3.5 Summary

In this chapter, we discussed the data resources available to our task. We examined dominant image databases currently used in computer vision and outlined their shortcomings with regard to the caption generation task. We proposed to overcome the data acquisition bottleneck associated with our task by taking advantage of publicly available resources where images and their textual descriptions co-occur naturally. These resources are many and easy to collect from the internet. Examples include news articles and their images, Wikipedia entries, and so on. Our image dataset is collected from the BBC News website, where news articles are typically accompanied with captioned images. Although our data is noisy in nature, covering a wide range of topics, with approximate annotations and low quality images, we argue that it can be used as a testbed for automatic image caption generation, under certain conditions (e.g., that we do not aim to label all objects depicted in the image). In next chapter, we discuss how we extract semantic information from a news image with the help of its associated news document.

# Chapter 4

# Extended Continuous Relevance Image Annotation Model

In this thesis we will explore the feasibility of automatically generating captions for news images. This task comprises of two components, namely extracting the image content and rendering it into natural language form. We propose to build an automatic image annotation model as an approximation of the image's content.

Recall that our dataset is captioned news images and their associated documents. In this settings, we define image annotation as below:

**Definition 4.** *Given a news image I with visual features* $V_i = \{v_1, v_2, \ldots, v_N\}$ *, its associated news document D, and a set of keywords* $W = \{w_1, w_2, \ldots, w_M\}$*, the task is to find automatically the keyword subset* $W_I \subset W$*, which can appropriately describe the image I aided by document D.*

Note that this definition is different from traditional image annotation task where the associated documents are absent in their datasets.

In the previous chapter, we have discussed what elements of a dataset are crucial and necessary for our image caption generation task, and found that existing image databases built for computer vision and image retrieval research are not readily suited to our task. We therefore proposed to collect an image dataset from the BBC News website which contains multiple article-picture pairs. Furthermore, images are routinely accompanied by captions. News images in this dataset are challenging as they are usually depicting cluttered scenes, contain more but less prominent objects, and are often rendered in low resolution. Moreover, the caption is not explicitly crafted for this task, but is image specific focusing on the news event that is shared between the image and the document. We further assume that the caption can be considered as coarse-gained annotation of the image, and the accompanying news document describes the content of its corresponding image. Our dataset differs from the traditional image dataset in that images are weakly annotated and accompanied with a news document that can potentially describe the image content in more detail.

With such an image dataset at hand, we will first address the following questions:

1. Is it possible to create an image annotation model from noisy data that has not been explicitly hand labeled to serve as a content extraction component? Despite the initial investigation mentioned in Chapter 3, we need to demonstrate that this image dataset can be used to capture the correlation between visual and textual information.

2. What is the contribution of the associated news document? In other words, is it really necessary to take the document into account during modeling if we assume that the caption is the annotation for the image? Taking the auxiliary news document into consideration, will surely increase the computational complexity for an image annotation model, but this can be justified as long as we demonstrate a substantial increase in model performance.

3. What is the contribution of the image? As we have discussed before, most existing image retrieval systems available on the internet only consider textual resources found adjacent to the images as well as “user click information”. Here, we will try to access if the visual modality is contributing to better model performance. For example, we could simply tag an image with the most topical document words without any image processing.

In this chapter we first review the relevance model for information retrieval and introduce the continuous relevance image annotation model (CRM, Lavrenko et al. 2003) in more detail. Next, we describe how we adapt this model to our dataset by taking the auxiliary news document into account. We utilize the document information as a complementary resource and demonstrate that it can assist in estimating the conditional probability of a keyword given an image, and further to prune the model’s output. We also discuss our data preprocessing, experimental setup, and present our results.

## 4.1 Continuous Relevance Model For Image Annotation

Over the past decades, the task of automatic image annotation have received increasing attention within computer vision and image retrieval. A large number of approaches have been proposed in the literature under many distinct learning paradigms. Despite differences in application and formulation, all these methods essentially attempt to learn the correlation between image features and keywords from examples of annotated images. In Chapter 2, we have briefly reviewed recent work for this task, which can be broadly categorized into discriminative and generative approaches. The latter try to model the joint probability of images and annotation keywords. Generally speaking in computer vision, generative models are good at handling low quality data, e.g., partially labeled data, and can relatively easily deal with changes of dataset and keyword vocabulary, compared to discriminative approaches (Ulusoy and Bishop, 2005; Holub, 2007).

We argue that the generative paradigm is more appropriate to our setting for several reasons. Firstly, we aim to perform the image caption generation task in a knowledge-lean way. We therefore expect every component in our model to be *learned from* data with minimal human involvement. This would impose more difficulties for classification-based discriminative models as one would need to train and maintain a large number of classifiers. Secondly, our dataset is noisy in nature, and compared to the traditional datasets, contains more challenging data, e.g., cluttered scenes and multiple labels, which can be better handled by generative models since they are not as sensitive to the quality of labeling as discriminative models are (Ulusoy and Bishop, 2005) (see Chapter 2 for an overview).

Among already existing generative models for image annotation, the Continuous Relevance Model (CRM, Lavrenko et al. 2003), is a good point of departure. This model captures the joint probability of images and annotated keywords *directly*, without requiring an intermediate clustering stage, and employs a relatively simple structure where expectations are computed over every single point in the training set and therefore parameters can be estimated without the EM algorithm. Indeed, it is one of the state-of-the-art generative image annotation models. And as we will show below, the model can be easily extended to incorporate information outside the image and its keywords.

The CRM model originates from the Relevance-based Language Model (Lavrenko and Croft, 2001). In information retrieval, a central problem is to estimate the probabilities of words given the relevant documents $P(w|RelevantDocuments)$, where *Relevant Documents* are the documents relevant to the given query $q$, and are usually difficult to explicitly and automatically specify. The relevance-based language model proposes to estimate $P(w|RelevantDocuments)$ by using query words and plain training set alone without explicitly identifying which documents are relevant to query $q$. Formally, this conditional probability can be derived as:

$$\begin{aligned} P(w|RelevantDocuments) &\approx P(w|q_1,q_2,..,q_J) && (4.1)\\ &= \frac{P(w,q_1,q_2,..,q_J)}{P(q_1,q_2,..,q_J)} && (4.2)\\ &\propto P(w,q_1,q_2,..,q_J) && (4.3)\end{aligned}$$

where $q_1,q_2,..,q_J$ is a query containing $J$ words. We can then make a strict assumption that both query words $q_1,q_2,..,q_J$ and word $w$ are sampled identically and independently from a universe distribution $\mathcal{S}$ which is assumed to be the source of all text. In the other words, we can describe the generative process as :

1. from source $\mathcal{S}$, select a sample $s$ with probability $P(s)$,
2. from sample $s$, select a word $w$ with probability $P(w|s)$,
3. for $j = 1,2,...,J$, from the same sample $s$,
    - select a word $q_j$ with probability $P(q_j|s)$ .

In practice, the source distribution $\mathcal{S}$ can be limited to the training set and sample $s$ is therefore each entry in the training set.

Then, we can write down the joint probability $P(w,q_1,q_2,..,q_J)$ as

$$\begin{aligned} P(w,q_1,q_2,..,q_J) &= \sum_{s\in\mathcal{S}} P(s)P(w,q_1,q_2,..,q_J|s) && (4.4)\\ &= \sum_{s\in\mathcal{S}} p(s)p(w|s)P(q_1,q_2,..,q_J|s) && (4.5)\\ &= \sum_{s\in\mathcal{S}} p(s)p(w|s)\prod_{j=1:J} P(q_j|s), && (4.6)\end{aligned}$$

and accordingly write the query prior by marginalizing over $w$ as:

$$\begin{aligned} P(q_1,q_2,..,q_J) &= \sum_{w\in Vocabulary} P(w,q_1,q_2,..,q_J) && (4.7)\\ &= \sum_{w\in Vocabulary}\sum_{s\in\mathcal{S}} p(s)p(w|s)\prod_{j=1:J} P(q_j|s). && (4.8)\end{aligned}$$

The resulting approximation of $P(w|RelevantDocuments)$ can be further used to compute conditional document probability ($P(W_d|RelevantDocuments)$, and rank the candidate documents accordingly.

Lavrenko et al. (2003) adapt the idea of estimating the conditional probability $P(w|RelevantDocuments)$ through query words to the scenario of image annotation. They estimate the joint distribution $P(W,V)$ of keywords $W$ and image regions $V$, by assuming that an image and its annotation keywords are closely related and express the same meaning with respect to the image content. Let $\mathcal{S}$ denote an image training set, each entry of which contains an image $I$ represented by a set of regions $V_I$ and annotation keywords $W_I$. It is assumed that the process of generating keywords is conditionally independent from the process of generating image regions. The independence assumption is further made within the generation of annotation keywords, and image regions. Then, the generative process can be described as:

1. from training set $\mathcal{S}$, select an entry $s$ with probability $P(s)$,
2. from sample $s$, select annotation keywords $W_I$ with probability $P(W_I|s)$,
3. from sample $s$, select image regions $V_I$ with probability $P(V_I|s)$.

where $P(W_I|s)$ is characterized by a multinomial distribution $P(*|s)$ and $P(V_I|s)$ by Gaussian kernel distributions $P_G(*|s)$. Then, the joint probability of image regions and annotation keywords can be written as:

$$P(V_I, W_I) = \sum_{s \in \mathcal{S}} P(V_I|s)P(W_I|s)P(s). \tag{4.9}$$

Equation 4.9 can be used to rank keywords given a test image. The $n$-best of which are selected as the annotations for the test image[1]. The model is evaluated on the Corel dataset, which contains 5,000 images and 371 words. The CRM model employs a relatively simple structure but performs surprisingly more efficiently and better compared to other latent variable-based models (Blei and Jordan, 2003).

## 4.2 Extended Continuous Relevance Model for Image Annotation

In this section, we describe how we adapt the CRM model to our BBC news dataset by utilizing the associated news document information.

---

[1] They used 5 in their original experiment

In CRM model, the expectations of annotation keywords $W_I$ and image regions $V_I$ are computed over every entry in the training set, as shown in Equation (4.9). For each entry $s$, there are three factors, $P(V_I|s)$, $P(V_I|s)$, and $P(s)$. The latter is the prior probability of entry $s$, which is assumed to be drawn from a uniform distribution:

$$P(s) = \frac{1}{N_{\mathcal{S}}} \tag{4.10}$$

where $N_{\mathcal{S}}$ is number of the image-annotation pairs in the training database $\mathcal{S}$.

When estimating $P(V_I|s)$, the probability of image regions given the current entry, Lavrenko et al. (2003) reasonably assume Gaussian kernel distributions for the generation of image regions:

$$\begin{aligned} P(V_I|s) &= \prod_{r=1}^{N_{V_I}} P_g(v_r|s) \\ &= \prod_{r=1}^{N_{V_I}} \frac{1}{n_{s_v}} \sum_{i=1}^{n_{s_v}} \frac{\exp\{(v_r - v_i)^T \Sigma^{-1} (v_r - v_i)\}}{\sqrt{2^k \pi^k |\Sigma|}} \end{aligned} \tag{4.11}$$

where $N_{V_I}$ is the number of regions in image $I$, $v_r$ the feature vector for region $r$ in image $I$, $n_{s_v}$ the number of regions in the image of latent variable $s$, $v_i$ the feature vector for region $i$ in $s$'s image, $k$ the dimension of the image feature vectors and $\Sigma$ the feature covariance matrix. According to Equation (4.11), a Gaussian kernel is fit to every feature vector $v_i$ corresponding to region $i$ in the image of current state $s$. Each kernel here is determined by the feature covariance matrix $\Sigma$, and for simplicity, $\Sigma$ is assumed to be a diagonal matrix: $\Sigma = \beta I$, where $I$ is the identity matrix; and $\beta$ is a scalar modulating the bandwidth of the kernel whose value is optimized on the development set.

Lavrenko et al. (2003) estimate the word probabilities $P(W_I|s)$ using a multinomial distribution. This is a reasonable assumption in natural language processing where word frequency (we only consider content words here) tends to reflect the importance of the word in the document, e.g., words that appear multiple times in a document tend to receive higher probabilities in the multinomial framework. However, this is not always true in image annotation case. Images are usually annotated according to whether the objects are present in the pictures (irrespectively of whether they appear frequently ). So, it is rare that an image is annotated multiple times with the same keyword in most existing datasets. The difference of annotation lengths also matters in the multinomial framework. In our dataset, the annotations have varying lengths, and rarely keywords are repeated. As we are more interested in modeling the *presence* or *absence* of words in the annotation, we follow Feng et al. (2004) in using the multiple-Bernoulli distribution to generate words in the CRM setting.

In Chapter 3, we made the assumption that the associated news document describes the content of the image (in addition to the caption), We propose to bring the document information into the model through a linear combination of document and caption words. Then, the probability of sampling a set of keywords $W$ given current entry $s$ from an underlying multiple Bernoulli distribution that has generated the entry $s$ can be written as:

$$P(W|s) = \prod_{w \in W} P(w|s) \prod_{w \notin W} (1 - P(w|s)) \tag{4.12}$$

where $P(w|s)$ denotes the probability of the $w$ component of the multiple Bernoulli distribution.

Now, when estimating $P(w|s)$ we include the news document information as follows:

$$P_{est}(w|s) = \alpha P_{est}(w|s_c) + (1-\alpha) P_{est}(w|s_d) \tag{4.13}$$

where $\alpha$ is a smoothing parameter tuned on the development set, $s_c$ is the annotation keyword in current entry $s$ and $s_d$ its corresponding document.

Equation (4.13) smooths the influence of the annotation words and allows to offset the negative effect of the noise inherent in our dataset. Since our images are implicitly annotated, there is no guarantee that the annotations are all appropriate. By taking into account $P_{est}(w|s_d)$, it is possible to annotate an image with a word that appears in the document but is not included in the caption.

We follow Feng et al. (2004) and use a Bayesian framework for estimating $P_{est}(w|s_c)$. Specifically, we assume a beta prior (conjugate to the Bernoulli distribution) for each word:

$$P_{est}(w|s_c) = \frac{\mu\, b_{w,s_c} + N_w}{\mu + |S|} \tag{4.14}$$

where $\mu$ is a smoothing parameter estimated on the development set, $b_{w,s_c}$ is a Boolean variable denoting whether $w$ appears in the annotation $s_c$, and $N_w$ is the number of latent variables that contain $w$ in their annotations.

We estimate $P_{est}(w|s_d)$ using maximum likelihood (Ponte and Croft, 1998):

$$P_{est}(w|s_d) = \frac{num_{w,s_d}}{num_{s_d}} \tag{4.15}$$

where $num_{w,s_d}$ denotes the number of times word $w$ appears in the accompanying document of entry $s$ and $num_{s_d}$ the number of all tokens in this document. Note that we purposely leave $P_{est}$ unsmoothed, since it is used as a means of balancing the probabilities of words appearing in annotations. So, if a word does not appear in the document, the possibility of selecting it will not be greater than $\alpha$ (see Equation (4.13)).

Unfortunately, including the document information in the estimation of $P_{est}(w|s)$ increases the vocabulary which in turn increases computation time. Given a test image-document pair, we must evaluate $P(w, V_I)$ for every $w$ in our vocabulary which is the union of the caption and document words. We reduce the search space, by scoring each document word with its $tf * idf$ weight (Salton and McGill, 1983) and adding the $n$-best candidates to our caption vocabulary. In this way, the vocabulary is not fixed in advance for all images but changes dynamically depending on the document at hand.

It is easy to see that the output of our model is a ranked keyword list. Following by common practice (Duygulu et al., 2002; Jeon et al., 2003), we take the $k$-best words to be the annotations for a test image $I$ where $k$ is a small number and the same for all images.

So far we have taken account of the auxiliary document rather naively, by considering only its vocabulary in the estimation of $P(W|s)$. Crucially, documents are written with one or more topics in mind. The image (and its annotations) are likely to represent these topics, so ideally our model should prefer words that are strong topic indicators of the document. A simple way to implement this idea is by pruning our ranked keyword list according to a topic model estimated from the news document collection.

Specifically, we use Latent Dirichlet Allocation (LDA) as our topic model (Blei et al., 2003). LDA represents documents as a mixture of topics and has been previously used to perform document classification (Blei et al., 2003) and ad-hoc information retrieval (Wei and Croft, 2006) with good results. Given a collection of documents and the desired number of topics, the LDA model estimates the probability of topics per document and the probability of words per topic. We will provide more detail about LDA model in Section 5.1.2 of Chapter 5.

Specifically, we use the LDA model to infer the $m$-best topics in the accompanying document. We then select from the output of our model those words that are most likely according to these topics. To give a concrete example (as shown in Figure 4.1), let us assume that for a given image our model has produced a ranked list of annotation keywords, $w_1$, $w_2$, $w_3$, $w_4$, $w_5$. However, according to the LDA model neither $w_2$ nor $w_5$ are likely topic indicators. We therefore remove $w_2$ and $w_5$ and substitute them with words further down the ranked list that are topical (e.g., $w_6$ and $w_7$). An advantage of

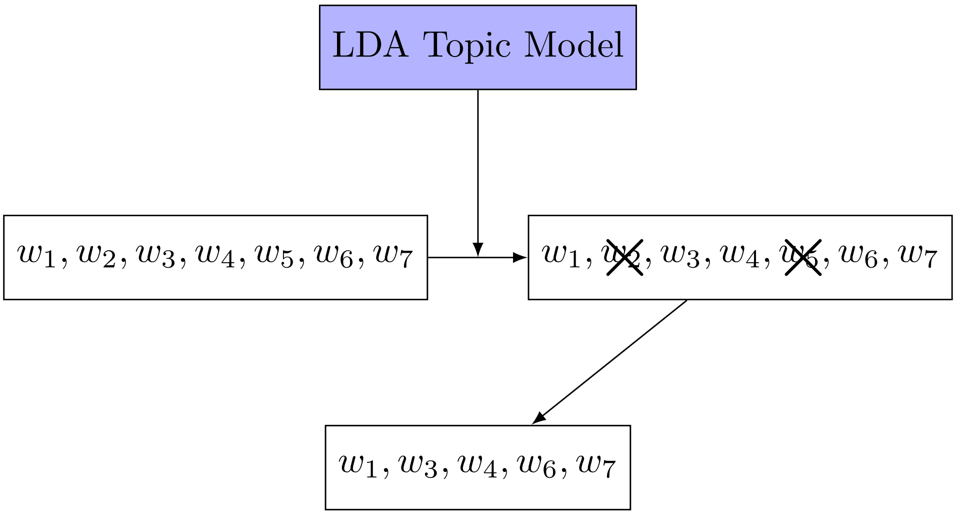

LDA Topic Model
$w_1, w_2, w_3, w_4, w_5, w_6, w_7$
$w_1, \cancel{w_2}, w_3, w_4, \cancel{w_5}, w_6, w_7$
$w_1, w_3, w_4, w_6, w_7$

Figure 4.1: An example of annotations pruned by LDA topic model. $w_1, w_2, \ldots$ are ranked annotation keywords according to the probability $P(w, V_I)$. However, $w_2$ and $w_5$ are identified as non-topical indicators by the topic model and will be discarded from the final annotation output.

using LDA is that at test time we can perform inference on the test document without retraining the topic model.

## 4.3 Experimental Setup

In this section we discuss our experimental design for assessing the performance of the model presented above. We give details on our training procedure and parameter estimation, describe the preprocessing of our data, and present the baseline methods used for comparison with our approach.

**Data** Our model was trained and tested on the BBC news dataset introduced in Chapter 3. The documents and captions were part-of-speech tagged and lemmatized with Tree Tagger[2](Schmid, 1994). Words other than nouns, verbs, and adjectives (see Table 4.1) were discarded. We assume that a word that appears at least 5 times in the training set is learnable and rare words were removed to avoid unreliable estimation. After filtering, the average caption length is 5.35 words and the average document length is 133.85 words. Our captions have a vocabulary of 2,167 words and documents 6,253,

---

[2] The Tree Tagger achieved an accuracy of 96.36% on the PennTree bank (Marcus et al. 1994)

| Words | Selected POS Tags |
|---|---|
| Nouns | NN, NNS, NP, NPS |
| Verbs | VVD, VVG, VVN, VVP, VVZ |
| Adjectives | JJ, JJR, JJS |

Table 4.1: Part of speech tags used to select content words from image captions and documents.

| Color |
|---|
| average of RGB components, standard deviation<br>average of LUV components, standard deviation<br>average of LAB components, standard deviation |
| Texture |
| output of DCT transformation<br>output of Gabor filtering (4 directions, 3 scales) |
| Shape |
| oriented edge (4 directions)<br>ratio of edge to non-edge |

Table 4.2: Set of image features used in our experiments.

and the overlap between them is 2,056 words. In total, our vocabulary consists of 8,309 words.

Images are typically segmented into regions prior to modeling ( e.g., using normalized cuts (Barnard and Forsyth, 2001; Duygulu et al., 2002; Lavrenko et al., 2003; Blei and Jordan, 2003; Jeon et al., 2003), or averagely cut rectangles (Feng et al., 2004; Li and Wang, 2003; Wang and Li, 2002; Jeon and Manmatha, 2004)). We impose a fixed-size rectangular grid on each image rather than attempting segmentation using a general purpose algorithm such as normalized cuts (Shi and Malik, 2000). Using a grid avoids unnecessary errors from image segmentation algorithms, and reduces computation time (Feng et al., 2004). Taking the small size and low resolution of the news images into account, we averagely divided each image into $6 \times 5$ rectangles and extracted features for each region. We used 46 features (in terms of color, texture, and shape), which are summarized in Table 4.2.

**Model Parameters** The model presented above has a few parameters that must be selected empirically on the development set. These include the parameters for relevance model ($\beta$,$mu$, and $alpha$), the vocabulary size, which is dependent on the $n$ words with the highest $tf * idf$ scores in each document, and the number of topics for the LDA-based pruning component. The relevance model parameters are interrelated, we hence exhaustively examined all combinations of the three parameters. We empirically tuned the bandwidth parameter $\beta$ from 0.1 to 100, the smoothing parameter $\mu$ from 0.1 to 300, and parameter $\alpha$ from 0.05 to 1.0; the optimal parameter setting is $\beta = 0.075$, $\mu = 0.1$, and $\alpha = 0.9$. We obtained best performance with $n$ set to 100 (no cutoff was applied in cases where the associated document was less than 100 words). We trained an LDA model with 20 topics on our document collection using David Blei's implementation.[3] We used this model to prune the output of our annotation model according to the three most likely topics in each document.

**Baselines** We compared our model against the following baselines.

1. $tf * idf$**:** our first baseline ranks the document's content words (i.e., nouns, verbs, and adjectives) according to their $tf * idf$ weights and selects the top $k$ to be the final annotations.

2. **DocTitle:** the second baseline simply annotates the image with the document's title. Again we only use content words (the average title length in the training set is 4.0 words). Titles are an intuitive baseline as they often succinctly summarize the contents of the news documents.

3. **CRM:** we compare our model with the original CRM model (Lavrenko et al., 2003). It is trained solely on image-caption pairs, uses a vocabulary of 2,167 words and the same features as our extended model.

**Evaluation** During test, we are given an unannotated image $I$ with its associated document and are asked to automatically produce suitable annotations for $I$. Given a set of image regions $V_I$, we use Equation (4.9) to derive the joint distribution $P(w, V_I)$. We consider the $n$-best words as the annotations for image $I$. We present results using the top 10, 15, and 20 annotation keywords. We assess our model's performance using precision/recall and F1. In our task, precision is the percentage of correctly annotated

---

[3] Available from `http://www.cs.princeton.edu/~blei/lda-c/index.html`.

| Model | Top 10 | | | Top 15 | | | Top 20 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Pre. | Rec. | F1 | Pre. | Rec. | F1 | Pre. | Rec. | F1 |
| $tf * idf$ | 4.37 | 7.09 | 5.41 | 3.57 | 8.12 | 4.86 | 2.65 | 8.89 | 4.00 |
| DocTitle | 9.22 | 7.03 | 7.20 | 9.22 | 7.03 | 7.20 | 9.22 | 7.03 | 7.20 |
| CRM | 9.05 | 16.01 | 11.81 | 7.73 | 17.87 | 10.71 | 6.55 | 19.38 | 9.79 |
| ExtModel | 14.72 | 27.95 | 19.82 | 11.62 | 32.99 | 17.18 | 9.72 | 36.77 | 15.39 |

Table 4.3: Automatic image annotation results on the BBC News database.

keywords over all annotations that the system suggested:

$$precision = \frac{correctly\ annotated\ words}{system\ output}. \tag{4.16}$$

Recall, is the percentage of correctly annotated words over the number of genuine annotations in the test data:

$$recall = \frac{correctly\ annotated\ words}{genuine\ annotations}. \tag{4.17}$$

F1 is the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}. \tag{4.18}$$

These measures are averaged over test set.

## 4.4 Results

As we discussed previously, our experiments were driven by three questions: (1) can we build an annotation model from the BBC news dataset that has not been explicitly hand labeled like traditional image datasets? (2) can the model benefit from the auxiliary document? As mentioned earlier, considering the associated document increases the model's computational complexity, which can be justified as long as we demonstrate a substantial increase in performance. (3) what is the contribution of the image? Here, we are trying to assess if the image features matter.

Our automatic evaluation results are summarized in Table 4.3. We compare the annotation performance of the model described in this chapter (ExtModel) with the original continuous relevance model (CRM, Lavrenko et al. 2003) without considering documents, and two other simpler models which do not take the image into account

($tf * idf$ and DocTitle). First, note that the original CRM model performs best when the annotation output is restricted to 10 words with an F1 of 11.81% (recall is 9.05 and precision 16.01). F1 is marginally worse with 15 output words and decreases by 2% with 20. This model does not take any document-based information into account, it is trained solely on image-caption pairs. On the Corel dataset the same model obtains a precision of 19.0% and a recall of 16.0% with a vocabulary of 260 words. Although these results are not strictly comparable with ours due to the different nature of the dataset (in addition, we output 10 annotation words, whereas Lavrenko et al. (2003) output 5), they give some indication of the decrease in performance incurred when using a more challenging dataset. Unlike Corel, our images have greater variety, less-overlapping content and employ a larger vocabulary (2,167 vs. 260 words).

When the document is taken into account (see ExtModel in Table 4.3), F1 improves by 8.01% (recall is 14.72% and precision 27.95%). Increasing the size of the output annotations to 15 or 20 yields better recall, at the expense of precision. Eliminating the topic model pruning step from the extended model decreases F1 by 0.62%. Incidentally, LDA can be also used to prune the output of CRM model. LDA also increases the performance of this model by 0.41%. We also evaluated the ExtModel on different partitions of the BBC dataset in order ensure that the model perfoms consistently. We obtained a precision of $14.24 \pm 0.63$, a recall of $28.00 \pm 1 : 56$ and F1 score of $18.96 \pm 0.87$ for the top 10 annotation words.

Finally, considering the document alone, without the image yields inferior performance. This is true for the $tf * idf$ model and the model based on the document titles. Interestingly, the latter yields precision similar to the CRM model. This is probably due to the fact that the document's title is in a sense similar to a caption. It often contains words that describe the document's gist and expectedly some of these words will be also appropriate for the image. In fact, in our dataset, the title words are often a subset of those found in the captions.

Examples of the annotations generated by our model are shown in Figure 4.2. We also include the annotations produced by the CRM model and the two baselines. As it can be seen that our model annotates the image with words that are not always included in the caption. Some of these are synonyms of the caption words (e.g., *child* and *intelligent* in left image of Figure 4.2), whereas others express additional information (e.g., *mother*, *woman*). Also note that images of complex scenes remain challenging (see the center image in Figure 4.2). Such images are better analyzed at a higher resolution and probably require more training examples.

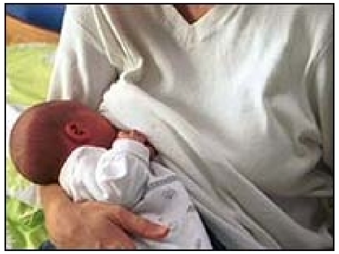
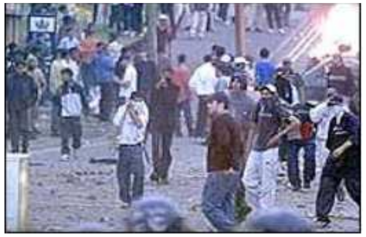
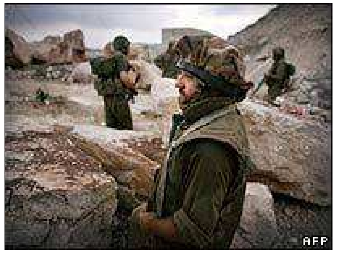
| |  |  |  |
|---|---|---|---|
| *tf* ∗ *idf* | **breastfeed**, medical, intelligent, health, child | culturalism, faith, Muslim, separateness, ethnic | **ceasefire**, Lebanese, disarm, cabinet, Haaretz |
| DocTitle | Breast milk does not boost IQ | UK must tackle ethnic tensions | Mid-East hope as ceasefire begins |
| Lavrenko03 | woman, **baby**, hospital, new, day, lead, good, England, look, family | bomb, city, want, day, fight, child, attack, face, help, government | war, carry, city, security, **Israeli**, attack, minister, force, government, leader |
| ExtModel | **breastfeed**, intelligent, **baby**, mother, **tend**, child, study, woman, sibling, advantage | aim, Kelly, faith, culturalism, community, Ms, tension, commission, multi, tackle, school | **Lebanon**, **Israeli**, Lebanese, aeroplane, **troop**, Hezbollah, Israel, force, **ceasefire**, grey |
| Caption | Breastfed babies tend to be brighter | Segregation problems were blamed for 2001's Bradford riots | Thousands of Israeli troops are in Lebanon as the ceasefire begins |

Figure 4.2: Examples of annotations generated by our model (ExtModel), the continuous relevance model (CRM), and the two baselines based on $tf * idf$ and the document title (DocTitle). Words in bold face indicate exact matches, underlined words are semantically compatible. The original captions are in the last row.

## 4.5 Summary

In this chapter, we detail our efforts to adapt the continuous relevance model (CRM), a state-of-the-art generative image annotation model, to our BBC news dataset which contains images, their captions, and associated document. Specifically, we extended the CRM model by taking into account the latter. We smoothed the conditional probabilities of keywords given a news image, and further pruned the model's output by assuming that image content words should be strong topic indicators of the associated news document.

The main purpose of our experiments was to validate whether our non-standard BBC news dataset is appropriate for capturing the correlation between visual and textual information. Our experimental results provide evidence that it is possible to create an image annotation model from this noisy data that has not been explicitly hand labeled. We also show that the image annotation model benefits substantially from the additional news document, beyond the caption or image. The model performs better than either the original CRM model, which does not take the news documents into account, or models that are solely text-based without any regard for the image. However, note that our analysis of the accompanying document was rather shallow, limited to word frequency and a post-pruning step, which suggests that more sophisticated modeling can be performed to mine our dataset.

# Chapter 5

# Image Annotation Based on Topic Models

In the previous chapter, we modified a state-of-the-art image annotation model, namely the continuous relevance model (CRM) and deployed it on our BBC news dataset in order to extract keywords given news images and their associated news documents. Our extensions of the CRM model are twofold and concern how to exploit the information present in the document. First, when estimating the conditional probability of a keyword given a captioned image, we also consider its likelihood in the collateral news document. Secondly, we further prune the model's output by assuming that image content words should be strong topic indicators of the associated document. Thus, the image content can be expressed by a ranked keyword list. Our experimental results provide evidence that it is possible to create an image annotation model from noisy data (like our BBC news dataset) that has not been explicitly hand labeled and show that the availability of the news document (compared to traditional image databases) helps improve model performance.

However, the associated news document is not an integral part of the model properly. Rather it is utilized in complementary fashion as a post-filtering tool. Moreover, in this setting, there are few opportunities to make visual information interact with the textual information present in the news text document. However, when discussing our assumptions in Chapter 3, we argued that our dataset is noisy in nature but the captioned news images and news documents are naturally structured together and closely related in content. Because the news document contains more detail about the image content, it should be mined further and more cleverly in a uniform way during modeling.

In this chapter, we propose a probabilistic image annotation model that learns to automatically label news images from the noisy data. We exploits the multimodal redundancy inherent in our BBC news dataset by assuming that news images with captions and their associated articles have been generated by a shared set of latent variables or topics. Specifically, we represent the images using scale-invariant feature transform (SIFT) descriptors (Lowe, 2004) in order to account for similar objects appearing at slightly different scales and transformations. The resulting descriptors are utilized as *visual terms* for each image. We further assume that documents and images can be described by a common multimodal vocabulary consisting of textual words and visual words. Due to polysemy and synonymy many terms (both textual and visual) will refer to the same underlying concept or topic. We infer these underlying topics using Latent Dirichlet Allocation (LDA, Blei and Jordan 2003), a probabilistic model of text generation, based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. Different from previous work, we do not make priority assumptions on either modality, and let the textual and visual words drive the definition of the topic space simultaneously. Our annotation model takes these topic distributions into account while finding the most likely keywords for an image and its associated document. We also illustrate how the model can be straightforwardly modified to perform automatic text illustration and report performance on this task as well.

In the remainder of this chapter, we will first review two topic model based image annotation approaches that are related to our model. Then, we describe our image annotation framework and the experiments we conducted on the BBC news dataset in more detail. Based on this framework, we also show how our model can be extended to perform a text illustration[1] task and present our evaluation results.

## 5.1 Topic Modelling for Image Annotation

We have briefly introduced topic models applied to the image annotation task in Section 2.3.2.3 Chapter 2. These models span from the standard latent semantic analysis (LSA, Zhao and Grosky 2003; Pan et al. 2004) and its probabilistic variant (PLSA, Monay and Gatica-Perez 2003, 2007; Fergus et al. 2005; Sivic et al. 2005; Bosch 2007), latent dirichlet allocation and its variants (LDA, Blei and Jordan 2003; Barnard et al. 2002; Fei-Fei and Perona 2005; Wang et al. 2009; Li et al. 2009; Ahmed et al.

---

[1] We use the terms *story picturing* and *text illustration* interchangeably throughout this chapter

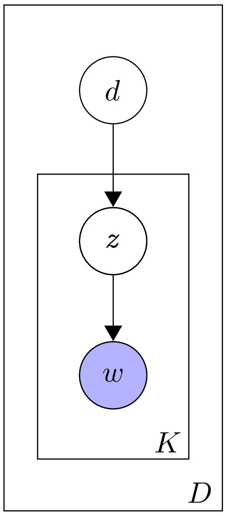


Figure 5.1: Graphical model representation of PLSA. $K$ indicates total number of topics and $D$ the document collection.

2009).

Generally speaking, the common core of these models is to define a latent topic space to facilitate the synergy between visual and textual modalities. A topic is usually defined as a distribution over multimodal data, and then an image and its annotations are assumed to be generated by mixture of different topics. Hence, the central task during training is how to construct the latent topic representation from annotated images, and during testing, it is important to infer the topic proportions for the test images. Before we present our model, we will first review two methods which are based on PLSA and LDA, respectively.

**PLSA based Image Annotation Models** Monay and Gatica-Perez (2007) use PLSA (Hofmann, 2001) to perform the image annotation task. PLSA is a generative model[2] which performs mixture decomposition on co-occurrence data through the EM algorithm.

Given a text document collection $D$, the process of generating it can be described as follows and shown graphically in Figure 5.1:

1. Select a document $d$ with probability $P(d)$
2. Then choose a topic $z$ with probability $P(z|d)$

---

[2] Compared to LDA, PLSA is not a fully generative model (Griffiths et al., 2007)

3. Generate a word $w$ with probability $P(w|z)$

Then, the joint probability of word $w$ and document $d$ can be written as :

$$P(w,d) = P(d)P(w|d) = P(d)\sum_{z} P(w|z)P(z|d). \tag{5.1}$$

Generally speaking, a PLSA model (parameterized mainly by $P(z|d)$ and $P(w|z)$) can be solved through the EM algorithm by maximizing the likelihood on the observed dataset. Given a corpus $D$, the log likelihood of the complete data will be:

$$LL = \sum_{d\in D}\sum_{w}\{n(d,w)[\log P(d) + \log\sum_{z_{1:K}} P(z|d)P(w|z)]\} \tag{5.2}$$

where $n(d,w)$ indicates the number of times word $w$ appears in the document $d$.

In the Expectation step of the EM algorithm, the conditional probability of topic $z$ given the observed data $(d,w)$ can be updated according to Bayes' theorem :

$$P(z_{new}|d,w) = \frac{P(w|z)P(z|d)}{\sum_{z_{1:K}} P(w|z)P(z|d)} \tag{5.3}$$

For the Maximization step, maximizing the expected log likelihood of $D$ yields new iterative steps as:

$$\begin{aligned} P_{new}(w|z) &= \frac{\sum_{d\in D} n(d,w)P(z|d,w)}{\sum_{w}\sum_{d\in D} n(d,w)P(z|d,w)} & (5.4)\\ p_{new}(z|d) &= \frac{\sum_{w} n(d,w)P(z|d,w)}{\sum_{w} n(d,w)} & (5.5)\end{aligned}$$

where $\sum_{w} n(d,w)$ indicates the length of document $d$. The two steps are repeated until the log likelihood of $D$ converges or other early stopping condones are satisfied[3].

The obtained parameter $P(z|d)$ can be considered as the topic proportions for document $d$, and $P(w|z)$, the distribution over words for topic $z$, can be deemed as the definition (interpretation) for topic $z$. Given a PLSA model trained on a corpus, it is possible to infer the topic proportions $P(z|d_{new})$ for an unseen document $d_{new}$ by performing a *fold-in* step in which topic definitions $P(w|z)$ are fixed and all other procedures in EM remain the same. The topic proportion for new document $P(z|d_{new})$ will be obtained when EM finishes.

Monay and Gatica-Perez (2007) propose several image annotation models based on the PLSA model. They first render images into word-like visual terms by clustering algorithm, and derive three annotation models, PLSA-Words, PLSA-Mixed and

---

[3] More details about the model fitting with the EM algorithm and further improvements can be found in Hofmann (2001)

PLSA-Features, from the original PLSA structure. The main difference among the three models is how to obtain the topic structures $P(z|d)$ from the training data and accordingly define the latent topic space (topic definition $P(w|z)$). PLSA-Words uses annotation words alone to obtain the topic proportions $P(z|d)$ as well as the textual part of the topic representation $P(w_t|z)$, and then folds the image features into the model in order to obtain the visual aspect $P(w_v|z)$. By contrast, PLSA-Features relies solely on the images to determine the topic proportions and further folds-in the annotation keywords into the model. The last model, PLSA-Mixed, uses both images and annotation words to infer the topic space, and follows normal PLSA procedures to estimate $P(w|z)$. Monay and Gatica-Perez (2007) report image annotation results on the Corel dataset, and show that a solely keyword-derived topic representation (PLSA-Words) works marginally better than the one derived only from image features (PLSA-Features), and when both images and keywords are used, the model (PLSA-Mixed) performance decreases significantly.

**Correspondence LDA** Correspondence LDA (CorrLDA), proposed by Blei and Jordan (2003), has been successfully employed for modeling annotated images in the Corel domain. CorrLDA also uses the notion of *topic* when modeling the generation of images and their keywords. In this model, the visual modality drives the definition of the latent space to which the textual modality is linked.

In terms of the generative process, given an image-annotation pair, CorrLDA first generates image regions from a Gaussian LDA model. The latter is essentially a Gaussian multinomial mixture model where a multinomial distribution is used to draw the mixture components and this distribution is characterized by a Dirichlet prior; then, for each annotation keyword, it first uniformly selects a region from the image, and next, generates a word according to the topic which has been used to generate the selected image region.

Formally, an image-annotation pair is represented by $N$ image regions ($r_{1:N}$) and $M$ annotation words ($w_{1:M}$), $\alpha$ is a Dirichlet prior, $z_{1:K}$ indicate latent topics, $y_{1:M}$ are index variables ranging from 1 to $N$, $Norm(\mu_{1:K}, \sigma_{1:K})$ is a $K$-dimension multivariate Gaussian distribution, and $\beta$ is the matrix of word probabilities given topic. Thus, the above generative process can be summarized as follows and shown graphically in Figure 5.2:

1. Choose $\theta|\alpha \sim Dir(\alpha)$

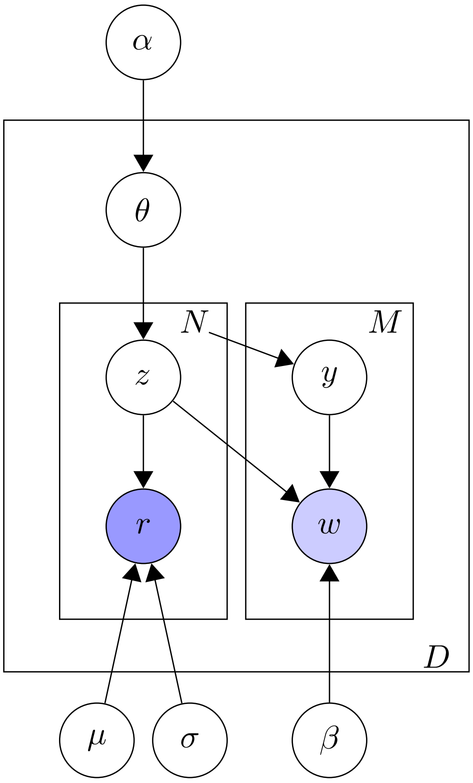


Figure 5.2: Graphical model representation of CorrLDA. $N$ and $M$ indicate the number of image regions $\{r\}$ and keywords $\{w\}$ in each pair, respectively, and $D$ the collection of captioned images.

2. For $n \in 1,...,N$ :
   (a) Choose topic $z_n|\theta \sim Mult(\theta)$
   (b) Choose region description $r_n|\{z_n, \mu_{1:K}, \sigma_{1:K}\} \sim Normal(\mu_{z_n}, \sigma_{z_n})$
3. For $m \in 1,...,M$ :
   (a) Choose region index $y_m|N \sim Unif(1,...,N)$
   (b) Choose word $w_m|\{y_m, z_{1:K}, \beta_{1:K}\} \sim Mult(\beta_{z_{y_m}})$

Hence, this generative process can derive the following joint distribution of observed images with annotation keywords and latent variables:

$$\begin{aligned}&P(r_{1:N}, w_{1:M}, \theta, Z, Y|\alpha, \beta_{1:K}, \mu_{1:K}, \sigma_{1:K})\\ =&P(\theta|\alpha)\prod_{n=1}^{N} P(z_n|\theta)P(r_n|z_n, \mu_{1:K}, \sigma_{1:K})\prod_{m=1}^{M} P(y_m|N)P(w_m|y_m, z_{1:K}, \beta_{1:K}).\end{aligned} \tag{5.6}$$

The parameters to be estimated are $\alpha$, $\beta$, $\mu$ and $\sigma$, the latent variables are $z_{1:K}$ (also $\theta$) and $y_{1:M}$, while direct inference from Equation (5.6) is intractable. Blei and Jordan (2003) propose a variational inference strategy to estimate the model parameters. Alternatively, a sampling approach can be used to simulate the generative process and collect parameters after reaching the target distribution or enough iterations.

Here is an example of using Gibbs sampling to solve the model. During the process of sampling, we need to compute two types of conditional probabilities in order to assign topic assignments for image regions and annotation keywords, respectively. For each image-annotation pair, the conditional probability of sampling the $n$th image region can be written as:

$$\begin{aligned} & p(z_n|\theta, z_{-n}, y_{1:M}, r_{1:N}, w_{1:M}) \\ & \propto p(z_n|z_{-n}, \theta) p(r_n|z_n, \mu_{1:K}, \sigma_{1:K}) \prod_{m=1}^{M} p(w_m|y_m, z_{1:N}, \beta) \end{aligned} \tag{5.7}$$

where $z_{-n}$ are all the topic assignments excluding the $n$th region, and the conditional probability of sampling the $m$th keyword will be:

$$\begin{aligned} & p(y_m|\theta, z_{1:N}, y_{-m}, r_{1:N}, w_{1:M}) \\ & \propto p(y_m|z_{1:N}) p(w_m|y_m, z_{1:N}, \beta) \prod_{n=1}^{N} p(z_n|\theta) p(r_n|z_n, \mu_{1:K}, \sigma_{1:K}) \end{aligned} \tag{5.8}$$

where $y_{-m}$ are the current topic assignments excluding the $m_{th}$ keyword. Other parameters will be updated accordingly after each pass over the whole corpus.

During testing, a sampling process can be re-run with trained model parameters solely on the test images; after enough iterations, one can collect subsequent samples with an appropriate lag and use these samples to compute the topic proportions.

CorrLDA addresses the relation between image regions and annotation words during modeling by drawing word topic assignments from the topics which have generated the image regions. This indeed brings certain dependency between region generation and word generation into the model. Intuitively, one can think of this process as generating an image first, and then asking human annotators to label keywords for it.

**Discussion** Both the models discussed above utilize topic modeling to capture the correspondence between visual and textual modalities. The derived topic representations provide a way to set up relations between images and annotation keywords. Both models perform the annotation task on the Corel dataset. As we discussed in Chapter 3, this dataset is, unfortunately, not representative of real-world data, hence not an ideal testbed for the task. The dataset contains many related images which in turn share keywords. This simplified annotation scheme further makes the derived topic representations not reliable or robust enough.

An issue with PLSA based models is the imbalance of the two modalities (i.e., visual vs. textual). In the Corel dataset, the average annotation length is less than 5 words while Monay and Gatica-Perez (2007) extract averagely 240 visual terms per image. This means that on average, visual terms are nearly 50 times more than textual words per image and contain not only more details from salient foreground objects but also details from background scenes. In contrast, images in Corel are only annotated with one or two foreground object names. Therefore, a PLSA model that prioritizes the textual aspect works best whereas PLSA-Mixed works worst since it places equal importance on both modalities.

In CorrLDA, images are segmented into regions which are modeled by multiple Gaussian distributions and further drive the construction of the latent space to which the textual modality is linked. There is a strong assumption here: annotation keywords strictly correspond to topics which have been used to generate the regions of the current image. This assumption is based on the observation in Corel that only salient objects in the images are labeled and the prerequisite that image regions are accurately segmented. But this will impose difficulties with more complex real-world data, where there is no one-to-one correspondence between annotation keywords and salient objects, or where annotation keywords are abstract or expressed by a combination of several objects. Take *wind* as an example, it is impossible to find one or more image regions that correspond to *wind*; similarly the keyword *badminton* may correspond to multiple objects in a picture, such as players, shuttlecocks, rackets or even badminton nets. The assumption also puts more emphasis on accurately modeling the image regions, which is still beyond the capabilities of generic image segmentation algorithms.

Despite of the differences in modeling structure, these topics model-based approaches are also restricted by the dataset itself and the preprocessing of data. PLSA-based models suffer problems from the imbalanced multimodal data where visual terms are far more than textual words in the Corel dataset. CorrLDA, on the other hand, relies on accurate image segmentation algorithms and well-prepared training data. Neither of the two is readily suited to our BBC news dataset which is admittedly noisy and is not explicitly annotated by human. However, from existing approaches, we find that topic models are able to capture the interplay between visual and textual

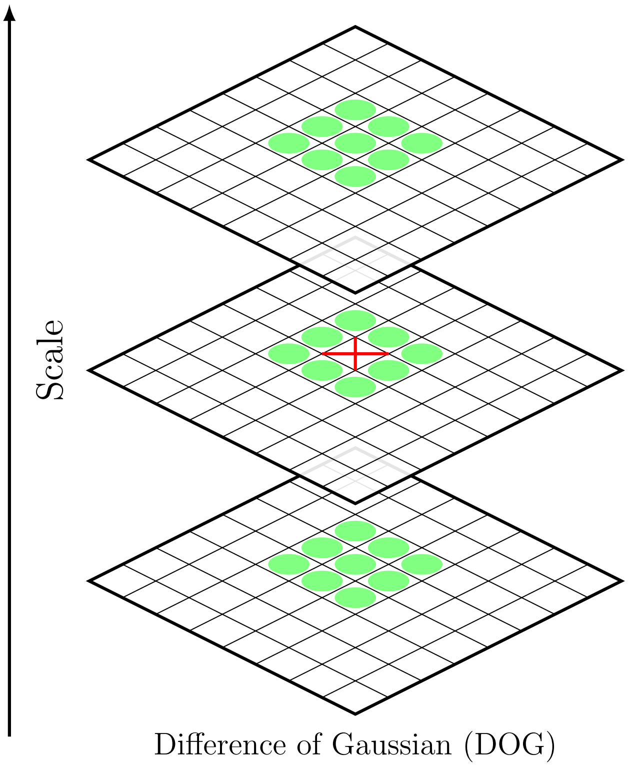


Figure 5.3: In difference-of-Gaussian images, extreme points (marked with *"cross"*) are located by comparing all neighbors in the scale space.

information in a generative manner. This is a good news for us since our purpose of building an image annotation model here is to provide image content for the generation module. Topic models are also naturally structured and can easily integrate multimodal data in a generative framework. As they are inherently probabilistic, topic models can further provide a more meaningful way, compared to a raneked list of keywords, to describe the content of an image, which are also convenient for our ultimate image caption generation task.

In what follows, we will describe a topic model for our data and accordingly perform the image annotation task. We also show how this model can be straightforwardly modified to perform automatic text illustration.

### 5.1.1 Image and Text 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 image into regions, using either a fixed-grid layout or an image segmentation algorithm. Regions have been described by a standard set of features, including color, texture, and shape, and subsequently treated as continuous vectors (e.g., Barnard et al. 2002; Blei and Jordan 2003) or in quantized form (e.g., Duygulu et al. 2002; Monay and Gatica-Perez 2007). Through this process, the low-level image features are made to resemble word-like units.

Here, we go one step further and represent each image by a bag of visual words, thereby converting visual features from a continuous space onto a discrete space. In order to do this we use the Scale Invariant Feature Transform (SIFT) algorithm (Lowe, 1999, 2004). The general idea behind the algorithm is to first sample an image with the difference-of-Gaussians point detector at different scales and locations (shown in Figure 5.3). Each detected region is represented with the SIFT descriptor which is a histogram of directions at different locations in the detected region (shown in Figure 5.4). 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 (Mikolajczyk and Schmid, 2003) and are considered state-of-the-art in object recognition (Bosch, 2007).

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 stands for 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, $[w_{v_1}, w_{v_2}, ..., w_{v_L}]$, where $w_{v_i} = n$ only if $I$ has $n$ regions labeled with $v_i$.

Since visual and textual modalities have now the same status, 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 here 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.

### 5.1.2 Latent Dirichlet Allocation

For ease of exposition, we first describe the basics of Latent Dirichlet Allocation (LDA, Blei et al. 2003), a probabilistic model of text generation and then move on to discuss our image annotation and text illustration models both of which make use of probabil-

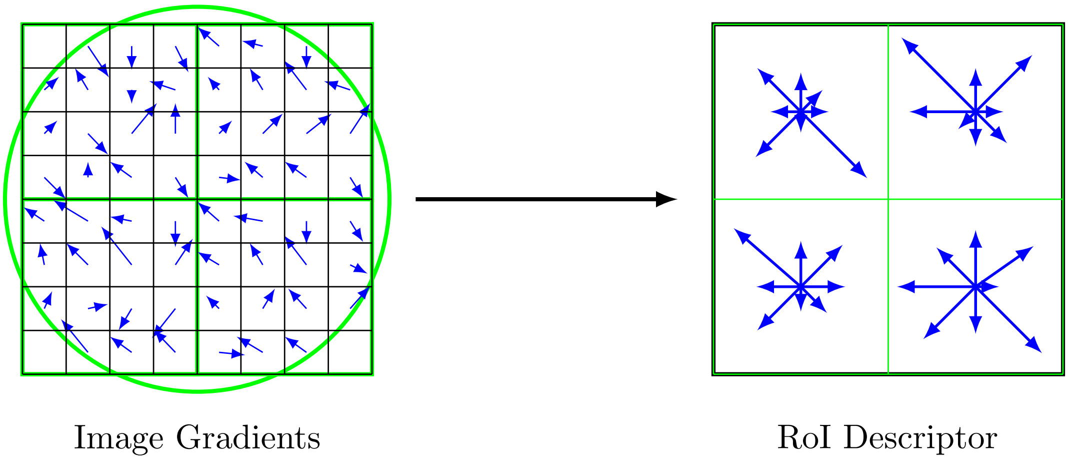


Figure 5.4: Once a keypoint is detected, its region of interest (RoI) will be located around the keypoint according to its position and scale (the circle on the left). The gradient magnitude and direction are computed for each sample point in that region (shown on the left). These orientations are then used to created a histogram. Here, the histogram comprises $2 \times 2$ sub-histograms, each with 8 bins (directions) from a $8 \times 8$ region. So, the dimension of the shown descriptor is $2 \times 2 \times 8 = 32$ (in our experiments, we use descriptors of $4 \times 4 \times 8 = 128$ dimensions extracted from a $16 \times 16$ region).

ities estimated by LDA.

LDA can be represented as a three level hierarchical Bayesian model shown graphically in Figure 5.5. Given a corpus consisting $M$ documents, each document is modeled using a mixture over $K$ topics (assumed to follow a multinomial distribution $\theta$ 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. (2003) define the generative process for a document $d$ as follows:

1. Choose $\theta|\alpha \sim Dir(\alpha)$
2. For $n \in 1, 2, ..., N$ :
   (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 $\theta$ is drawn from a Dirichlet prior with parameters $\alpha$ whose role is to create a smoothed topic distribution. Once $\alpha$ and $\beta$ are sampled, then each document is generated according to the topic proportions $z_{1:K}$ and word probabilities over topics $\beta$. The probability of a document $d$ in a corpus can be obtained as:

$$P(d|\alpha,\beta) = \int_{\theta} P(\theta|\alpha) \left( \prod_{n=1}^{N} \sum_{z_k} P(z_k|\theta) P(w_n|z_k,\beta) \right) d\theta. \tag{5.9}$$

The central computational problem in LDA is to compute, $P(\theta, z_{1:k}|d, \alpha, \beta)$, the posterior distribution of the hidden variables given a document. Although directly computing this distribution is intractable, in general, a variety of approximate inference algorithms have been proposed in literature including variational inference (Blei et al., 2003) and several forms of Markov chain Monte Carlo (Jordan, 1999).

Our model follows the convexity-based variational inference procedure described in Blei et al. (2003). This inference algorithm involves two steps, (a) introducing variational parameters in order to find the tightest lower bound for the target posterior distribution, and (b) obtaining the tight lower bound through minimizing the Kullback-Leibler (KL) divergence between the introduced variational distribution and the true posterior distribution.

Figure 5.6 shows an example of the graphical model used to approximate the posterior distribution in the original LDA model (we refer interested readers to Blei et al.

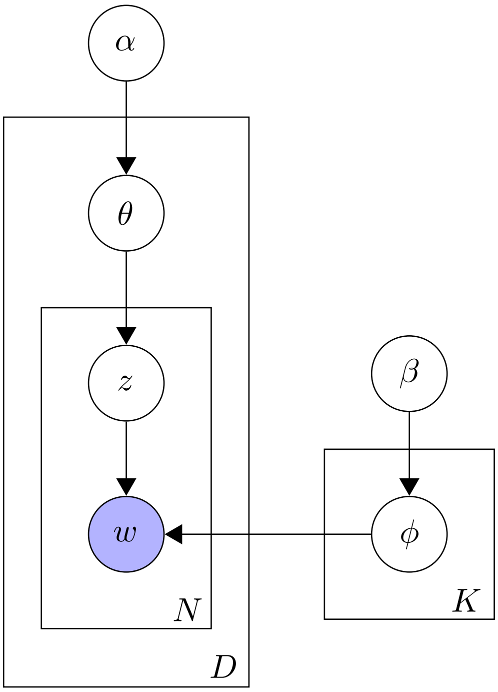


Figure 5.5: The LDA topic model model; shaded nodes represent observed variables, unshaded nodes indicate latent variables. Arrows indicate conditional dependencies between variables, whereas plates (the rectangles in the figure) refer to repetitions of sampling steps. The variables in the lower right corner refer to the number of samples.

(2003) and Blei (2004) for more details). Compared to Figure 5.5, the new graphical model drops node w and the edges from $\theta$ to z and z to w, while introducing two free variational parameters $\gamma$ and $\phi$. This will give us a tractable family of distributions on the latent variables, which can be further characterized by the the variational distribution (shown in Figure 5.6):

$$q(\theta, z_{1:K}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n), \tag{5.10}$$

where the Dirichlet parameter $\gamma$ and multinomial parameters $\phi_{1:N}$ are free variational parameters.

In order to find the tightest lower bound for the posterior distribution of interest, Blei et al. (2003) first minimize the KL divergence between $q(\theta, z_{1:K}|\gamma, \phi)$ and the true posterior as :

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma, \phi)} D(q(\theta, z_{1:K}|\gamma, \phi)||P(\theta, z_{1:k}|d, \alpha, \beta)). \tag{5.11}$$

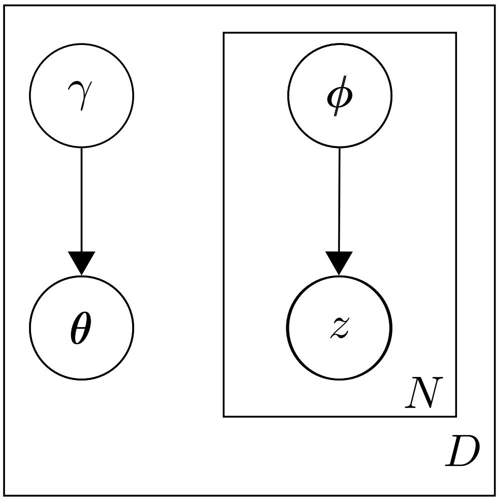


Figure 5.6: Graphical representation of the model used to approximate the posteriors in LDA. $\gamma$ and $\phi$ are free variational parameters.

This optimization process will determine the optimal $\gamma^*$ and $\phi^*$ for the tightest lower bound under the current model parameters $\alpha$ and $\beta$.

Given a training corpus $D$, parameters $\alpha$ and $\beta$ can be estimated by maximizing the log likelihood of the observed data $D$:

$$\begin{aligned}(\alpha^*, \beta^*) &= \arg\max_{\alpha,\beta} \log P(D|\alpha,\beta) &(5.12)\\ &= \arg\max_{\alpha,\beta} \sum_{d\in D} \log P(d|\alpha,\beta). &(5.13)\end{aligned}$$

As directly computing $P(d|\alpha,\beta)$ is intractable, Blei et al. (2003) propose to maximize, regarding to $\alpha$ and $\beta$, the tight lower bound of $P(d|\alpha,\beta)$ which is parameterized by $(\gamma^*, \phi^*)$ that are determined during the variational inference described above.

Blei et al. (2003) use a variational EM to perform this procedure: (E-step) for each document $d$, they perform variational inference to find the optimal variational parameters $(\gamma^*, \phi^*)_d$, and then (M-step) maximize the log likelihood on the data $D$ with respect to $\alpha$ and $\beta$, which is approximated by maximizing the tightest lower bound (characterized by $\{(\gamma^*, \phi^*)_d, d \in D\}$) for the log likelihood of $D$. When this EM procedure converges, we obtain the optimal variational parameters and model parameters[4].

Now, with an LDA model trained on a document collection at hand, we can obtain two sets of parameters, $P(w|z_{1:K})$, the word probabilities given topics and $P(z_{1:K}|d)$,

---

[4] More details about variational inference and parameter estimation can be found in Blei et al. (2003) and Blei (2004)

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

$$p(z|d_{new}) \approx \frac{\gamma}{\sum_{j=1}^{K} \gamma_j} \tag{5.14}$$

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:

$$p(w|d_{new}) \approx \sum_{k=1}^{K} P(w|z_k) \frac{\gamma_k}{\sum_{j=1}^{K} \gamma_j} \tag{5.15}$$

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

### 5.1.3 LDA based 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 $I$:

$$W_I^* = \arg\max_W P(W|I) \tag{5.16}$$

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

$$W_I^* = \arg\max_W \prod_{w \in W} P(w|I) \tag{5.17}$$

Recall that each entry in our dataset is an image-caption-document tuple $(I,C,D)$ under the assumption that the accompanying news document describes the content of the image. 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 $D$:

$$W_I^* = \arg\max_{W_t} P(W_t|I,D) \tag{5.18}$$

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:

$$W_I^* = \arg\max_{W_t} P(W_t|I,D) = \arg\max_{W_t} \prod_{w_t \in W_t} P(w_t|I,D). \tag{5.19}$$

Since $I$ and $D$ 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 $P(w_t|I,D)$:

$$P(w_t|I,D) \approx P(w_t|d_{Mix}) \tag{5.20}$$

Substituting Equation (5.20) into (5.19) yields:

$$W_I^* = \arg\max_{W_t} P(W_t|I,D) \approx \arg\max_{W_t} \prod_{w_t \in W_t} P(w_t|d_{Mix}). \tag{5.21}$$

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 Equation (5.14). We then substitute Equation (5.15) into Equation (5.21):[5]

$$\begin{aligned} W_I^* &\approx \arg\max_{W_t} \prod_{w_t \in W_t} P(w_t|d_{Mix}) \\ &= \arg\max_{W_t} \prod_{w_t \in W_t} \sum_{k=1}^{K} P(w_t|z_k)P(z_k|d_{Mix}) \\ &\approx \arg\max_{W_t} \prod_{w_t \in W_t} \sum_{k=1}^{K} P(w_t|z_k)\frac{\gamma_k}{\sum_{j=1}^{K}\gamma_j} \end{aligned} \tag{5.22}$$

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 $d_D$, 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 a single modality:

$$\begin{aligned} P^*(z_{1:K}|d_{Mix}) &\approx q_1 P(z_{1:K}|d_{Mix}) \\ &+ q_2 P(z_{1:K}|d_D) \\ &+ q_3 P(z_{1:K}|d_I) \end{aligned} \tag{5.23}$$

---

[5] During training, the model has access to all three elements $(I,C,D)$, so the mixed document $d_{Mix}$ is 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 $D$, but not $C$.

where $P(z_{1:K}|d_D)$ and $P(z_{1:K}|d_I)$ are inferred on $d_D$ 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 $(1 - q_1 - q_2)$.

In sum, calculating $P(W_t|I,D)$ boils down to estimating the probabilities $P(w_t|d_{Mix})$ according to Equations (5.22) and (5.23) which in turn 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 $I$.

Importantly, the presented model differs from the extended CRM model presented in Chapter 4. It 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 formulation will be advantageous when generating captions for an image. As we shall see in Chapter 6, our generation model is also probabilistic and this will allow for easy integration of the two components. The image annotation model will essentially indicate which words express the image content. In this chapter, however, we evaluate the image annotation model on its own, before discussing image caption generation.

### 5.1.4 Experiments

In this section we discuss our experimental design for assessing the performance of the model presented above. We give details on our training procedure and parameter estimation, describe our features, and present the baseline methods used for comparison with our models.

**Data** We evaluated the image annotation task on our BBC news dataset. Following the pre-processing described in Chapter 4, we performed part-of-speech tagging and lemmatization on documents and captions, and excluded from the vocabulary words other than nouns, verbs, and adjectives. Low frequency words (appearing less than five times) were also discarded.

We preprocessed the images as follows. We first extracted SIFT keypoints with descriptors from each image, 150 on average. As mentioned in Section 5.1.1, we used K-means to quantize these features into a discrete set of visual terms. We varied K experimentally (from 100 to 2000).

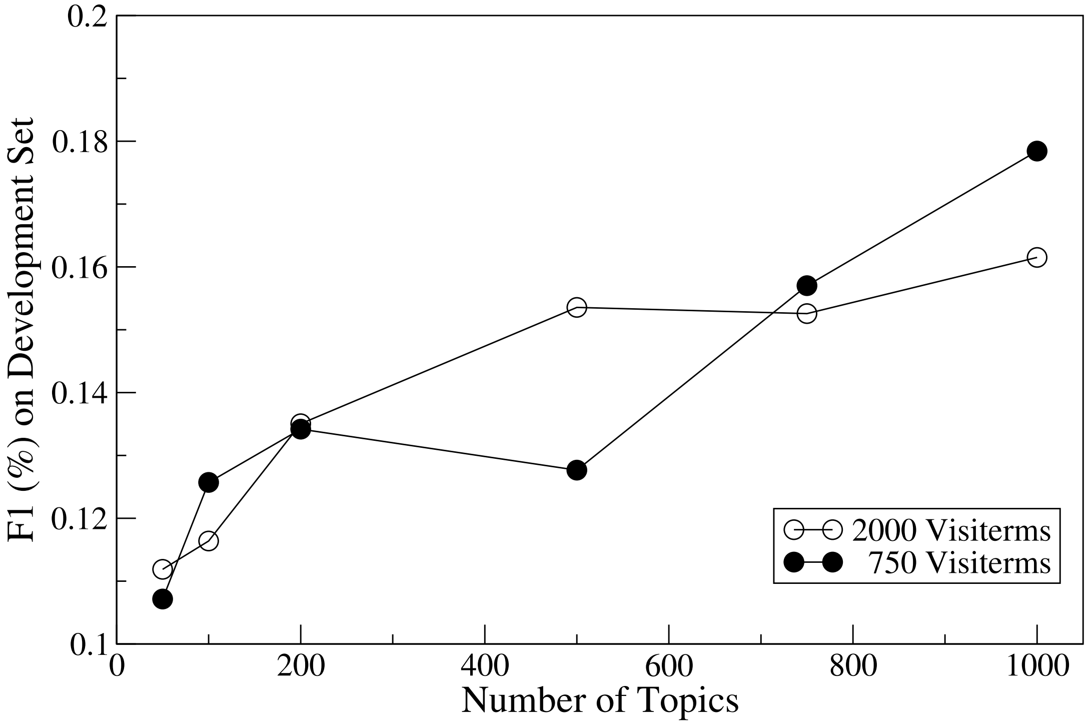


Figure 5.7: Image annotation performance of MixLDA model on the development set under different topic numbers using 2000 and 750 visual terms.

**Model Parameters** We trained the LDA topic model on the multimodal document collection $\{d_{Mix}\}$. We varied the number of topics from 15 to 1,000. In our experiments, hyperparameter $\alpha$ was set to 0.1 and the word-topic probability table $\beta$ 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 Equation (5.24)) on the development set. The best values were $q_1 = 0.84$, $q_2 = 0.12$, and $q_3 = 0.04$.

Evidently, the number of visual terms and topics are interrelated. We exhaustively examined all combinations of the number of visual terms (ranging from 100 to 2000) and topics (ranging from 50 to 1000) on the development set. We obtained best results on image annotation with 1,000 topics and 750 visual terms. Figure 5.7 shows how performance on the image annotation task varies with different topic numbers. Here, we use 750 and 2,000 visual terms and the results are reported on the development set.

**Evaluation** In this chapter, we evaluate the image annotation task on its own. As described in Chapter 4, we use precision/recall and F1 measures and compare against top 10 words.

**Baselines** For the image annotation experiments, we compared our model (MixLDA) against the following baselines:

1. **TxtLDA:** we trained a vanilla LDA model on the document collection without taking the images into account. This model estimates the predictive probability of textual word $w_t$ given text document $D$:

   $$P(w_t|d_D) = \sum_{k=1}^{K} P(w_t|z_k)P(z_k|d_D). \tag{5.24}$$

   This model assumes that the most likely words in the given test document are the annotations for the accompanying image.

2. **ContRel:** our second baseline is the continuous relevance model introduced in Lavrenko et al. (2003). This model is trained solely on image-caption pairs and uses the same settings described in Chapter 4.

3. **ExtRel:** the third baseline is the extended CRM model presented in Chapter 4.

4. **CorrLDA:** our fourth baseline is CorrLDA, which is similar to our model in terms of structure. CorrLDA was originally built on the Corel dataset which consists of image and keyword pairs. However, our BBC news dataset contains image-caption-document tuples. We therefore created two implementations: the first one is the CorrLDA model trained on news images and their captions and tested on news images alone; while the second version is the model that has access to news images-caption-document tuples during training, and is tested on image-document pairs. We optimized the parameters for both models on the development set. We used Gibbs sampling to perform inference on the data using HBC toolkit[6]. We varied the number of topics (from 10 to 200) on the development set. We obtained best results with 50 topics.

5. **PLSA-Mixed**, **PLSA-Words** and **PLSA-Features:** finally, we compared our model with three PLSA-based annotation models : PLSA-Mixed, PLSA-Words and PLSA-Features (Monay and Gatica-Perez, 2007)[7]. PLSA-Mixed directly trains a PLSA model on the mixed documents (consisting of a concatenation of textual words and visual terms). PLSA-Words and PLSA-Features are asymmetric versions of PLSA. In PLSA-Words, textual words define the topic structure

---

[6] See `http://www.umiacs.umd.edu/~hal/HBC/` for more details

[7] We used an implementation of PLSA available at `http://people.csail.mit.edu/fergus/iccv2005/bagwords.html`. See Sivic et al. (2005) and Fergus et al. (2005) for more details.

| Model | Top 10 | | |
|---|---|---|---|
| | Precision | Recall | F1 |
| CorrLDA(C) | 5.33 | 11.80 | 7.36 |
| PLSA-Features(C) | 8.8 | 18.5 | 12.0 |
| PLSA-Words(C) | 8.99 | 20.1 | 12.6 |
| PLSA-Mixed(C) | 8.37 | 15.9 | 11.1 |
| ContRel | 9.10 | 16.00 | 11.80 |
| CorrLDA(D) | 3.87 | 8.74 | 5.36 |
| PLSA-Features(D) | 10.2 | 21.80 | 13.8 |
| PLSA-Words(D) | 10.26 | 22.60 | 14.04 |
| PLSA-Mixed(D) | 10.30 | 22.60 | 14.16 |
| ExtRel | 14.70 | 27.90 | 19.80 |
| TxtLDA | 7.30 | 16.90 | 10.20 |
| MixLDA | 16.30 | 33.10 | 21.60 |

Table 5.1: Automatic image annotation results on the BBC News database. The bottom half of the table includes models that are evaluated on image-caption-document tuples, while models in the top half are evaluated on image-caption pairs.

which is fixed when visual data is *folded-in* into the model, while in PLSA-Features, visual words dominate the construction of topic space. Similarly to CorrLDA, we first tested a vanilla version of the three models (PLSA-(C)) without access to associated news documents, and then trained these models on image-caption-document tuples and tested them on image-document pairs (PLSA-(D)). We optimized the parameters for the three models on the development set. We varied the number of topics experimentally. Following Monay and Gatica-Perez (2007), we experimented with 25 to 200 topics and obtained best performance with 200. As a sanity check, we also ran PLSA with 500 and 1,000 topics but the results were inferior to the 200 topics. We also experimented with the size of the visual vocabulary. We report results with 2,000 visual terms.

### 5.1.5 Results

Our results on the image annotation task are summarized in Table 5.1. Here, we compare our own model (MixLDA) which is trained on both visual and textual information against: (a) an LDA model trained solely on textual information (TxtLDA), (b) the Continuous Relevance model (ContRel; Lavrenko et al. 2003) which learns annotations from image-caption pairs, (c) the extended version of the continuous relevance model that also takes collateral document information into account (ExtRel; Chapter 4), (d) the CorrLDA model trained on image-caption pairs, and (e) three PLSA based models, one that ascribes equal importance to the visual and textual modalities (PLSA-Mixed(D)), one that prioritizes the textual modality (PLSA-Words(D)) and one that emphasizes the visual modality (PLSA-Features(D)) (all three models were trained with access to accompanying news documents). We performed significance testing on F1 using stratified shuffling (Noreen, 1989), an instance of assumption-free approximative randomization testing.

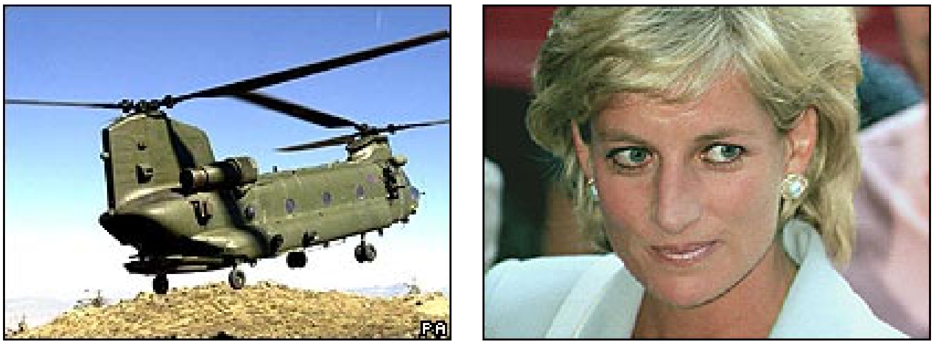
Let us first discuss the performance of TxtLDA and MixLDA. These two models are closely related — they both rely on the probabilities $P(w_t|d)$ to generate the image keywords — save one important difference. MixLDA uses a concatenated representation of words and visual features $d_{Mix}$ while attempting to simultaneously model the visual and textual modalities. In other words, MixLDA assumes that the two modalities have equal importance in defining the latent space, which as the results suggest is a useful assumption. MixLDA outperforms TxtLDA in terms of precision (by 9%), recall (by 16.2%) and improves F1 (by 11.4%); the difference is significant ($p < 0.01$)

Perhaps surprisingly, TxtLDA performs comparably to ContRel, a model that takes both visual and textual information into account. However, bear in mind that ContRel is based only on captions without document information at all, which are sparser and noisier. TxtLDA has access to a larger document collection and therefore knowledge about which words are topical in a given document. The topics here play a similar role to the image, i.e., they highlight important entities or actions mentioned in the text. Even though TxtLDA is not as precise as the continuous relevance model, our results suggest it is possible to obtain an image annotation model without any image processing. This result corroborates the strategy adopted by most commercial search engines to index images using their surrounding text[8].

Before we discuss the performance of PLSA based models and CorrLDA, we examine how the news documents affect the models' performance. As we can see in Table 5.1, access to news documents improves the PLSA models, albeit to a different extent. PLSA-Mixed(D) improves F1 considerably by 3.06%, while PLSA-Words(D) and PLSA-Features(D) are in the same ballpark; F1 increases marginally by 1.44% and

---

[8] However, note that most search engines take users' behavior into account as well.



| TxtLDA | |
|---|---|
| Afghanistan, Taleban, soldier, British, zone, kill, force, Microsoft, **troop**, NATO | police, Burgess, time, letter, **crash**, case, death, operation, investigation, jail |
| PLSA-Mixed (D) | |
| Afghanistan, **troop**, soldier, NATO, force, Taleban, British, kill, **operation**, kill | death, home, police, father, time, Paul, family, bank, know, coroner |
| ExtRel | |
| **helicopter**, Afghanistan, Blair, commander, **troop**, haul, support, NATO, British, minister | **Diana**, inquest, coroner, **crash**, workload, **Paris**, investigation, consuming, Burgess, Stevens |
| MixLDA | |
| Afghanistan, **troop**, Blair, British, NATO, **helicopter**, soldier, support, **operation**, commander | **Diana**, police, case, **crash**, **Princess**, report, **death**, inquest, **Paris**, Burgess |
| Caption | |
| Troops need more Chinook helicopters to carry out operations | Princess Diana died in a car crash in Paris in 1997 |

Figure 5.8: Examples of keywords generated by TxtLDA, PLSA-Mixed(D), ExtRel and MixLDA. Words in bold face indicate exact matches. The original captions are shown in the last row.

1.88%, respectively. While Monay and Gatica-Perez (2007) report large performance differences among the three models on the Corel dataset, we find that when taking accompanying documents into account, these models yield similar results on our dataset. The auxiliary news document helps maintain the balance between the visual and textual information and makes them mutually beneficial, which essentially provides more accurate estimation for the conditional probabilities of words (both visual and textual) given topics during the *fold-in* procedure, especially for the PLSA-Mixed(D) model where the latent topic space is essentially modeled by the two modalities simultaneously. Interestingly, when taking documents into account, CorrLDA(D) performs worse, F1 decreases by 2%. In CorrLDA, text words are generated according to the topics that have already been used to generate the images. This indicates the assumption that the procedure of generating textual modality depends on the one of visual. This is not the case when we considering accompanying documents in our dataset.

Compared to other models, CorrLDA performs significantly ($p < 0.01$) worse than both the PLSA(D) models and TxtLDA. Although CorrLDA delivers good 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 the 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 persons in an image (e.g., the word *badminton* is used to collectively describe an image depicting players, shuttlecocks, and rackets). As can be seen in Table 5.1, all three PLSA-(D) models are superior to TxtLDA and ContRel. This is not surprising, TxtLDA does not exploit any visual information and ContRel does not take advantage of any collateral document information. But neither of the three PLSA-(D) based models performs comparably to MixLDA.

We also used mean average precision (mAP), an evaluation measure common in information retrieval, to compare PLSA-(D) models and MixLDA[9]. The MAP value for MixLDA was 35.01% while for PLSA-Mixed(D) was 26.26%, for PLSA-Words(D) 26.26% and PLSA-Features(D) 26.12%. Intuitively, this means that a hypothetical query-retrieval system would find the relevant images earlier if the image database was annotated with MixLDA.

The extended relevance model improves considerably upon TxtLDA, CorrLDA, and PLSA(D) models but is significantly worse ($p < 0.01$) than MixLDA. On the sur-

---

[9] Average precision is the average of the precision scores at the rank locations of each relevant document. Mean Average Precision is the mean of the Average Precision scores for a group of queries (more details can be found in Monay and Gatica-Perez (2007))

face, MixLDA seems similar to ExtRel, both models take advantage of visual and textual information. ExtRel 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 non-topical keywords from the model's output. MixLDA is conceptually simpler, LDA is the actual model rather than a post-processing step, and exploits 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 ExtRel, MixLDA improves precision by 1.6%, recall by 5.2% and the overall F1 by 1.8%.

Figure 5.8 illustrates examples of annotations generated by TxtLDA, PLSA-Mixed(D), ExtRel and MixLDA for two news images. For comparison we also show the gold standard image captions. Note that TxtLDA and PLSA-Mixed(D) fail to generate any words relating to the objects shown in the image. They find primarily words relating to the topics of the associated articles such as *troops* and *crash*. On the contrary, ExtRel and MixLDA are more successful at identifying the objects shown in the pictures, since they take visual information into account.

## 5.2 Topic Modelling for Text Illustration

In this section, we will show how the proposed image annotation framework can be applied to text illustration (Joshi et al., 2006), a task which has received less attention in the literature, but is routinely performed by news writers who often have to search large image collections in order to find suitable pictures for their text.

Here, the model takes a document as input and suggests images that match the document's content. The task can be formally described as:

**Definition 5.** *Given a text document d, and a pool of candidate images, find an image I from the pool which best describes this document.*

Given the BBC news dataset, a text illustration model has access to the collection of image-caption-document tuples during training. During testing, the model is given a document and must find the image that best illustrates it.

A handful of models have been proposed in the literature for illustrating documents with images automatically. The majority of these are based on visual similarity, instance-based learning (Barnard and Forsyth, 2001) and typically use visual ranking-based schemes (Joshi et al., 2006). Here, we present a relatively simple model, again under the topic model framework.

Given a test document $D$ and a candidate image database $I_{1 \ldots N}$ with captions $C$, we must find the image or images which best describe the document. We can simply compute the predictive probability of each visual term in the vocabulary given $D$ by marginalizing over the document topics $z_{1:K}$:

$$P(w_v|D) = \sum_{z_{1:K}} P(w_v|z_k)P(z_k|d_D) \tag{5.25}$$

where $w_v$ is a visual term and $P(w_v|z_k)$ the probability of $w_v$ given topic $z_k$ is learned from the training data.

Equation (5.25) delivers a ranked list of visual terms according to a given document. We could multiply these probabilities together mirroring Equation (5.21), however this is not reliable. In contrast to textual words, for which we may infer whether they are linguistically meaningful (e.g., by resorting to their part of speech or stopword list), there is no easy way of knowing which visual words are important or content specific. Relying solely on frequency is not reliable either, as frequent visual terms may simply represent features common in nearly all images which are not discriminative enough for identifying objects, while, certain combinations of these frequent terms, sometimes may be meaningful enough to represent an object. To avoid a bias towards frequent but potentially irrelevant visual words, we output a fixed number of visual terms and select the image with the highest overlap as the correct illustration.

### 5.2.1 Experiments

**Data and Model Parameters** We performed the text illustration task on our BBC news dataset., We obtained best model parameters (1,000 topics and 2,000 visual terms) on the development set. For the purposes of simulating a real story picturing engine, we created a larger candidate image pool of 450 image-caption pairs and tested on 300 of them.

**Baselines** For our text illustration experiments, the proposed model was compared with four baselines:

1. **Overlap:** the first baseline is essentially word overlap. We select the image whose caption has the largest number of words in common with the test document.

2. **VectorSpace Model:** our second baseline is a straightforward implementation of the vector space model (Salton and McGill, 1983). Specifically, documents and captions are represented by vectors whose components correspond to term-document co-occurrences. We follow common practice in weighting terms by their tf-idf values. We measure the cosine similarity between the test document and image captions and output the image whose caption is most similar to the document.

3. **TxtLDA:** like the vector space model described above, our third baseline models text illustration as an information retrieval problem and does not take image-related features into account. Importantly, this latter model is probabilistic — the images most relevant to a document are found by maximizing the conditional probability of the candidate captions $C$ given the document $d_D$:

$$\begin{aligned} P(C|d_D) &= \prod_{w_c \in C} P(w_c|d_D) \\ &= \prod_{w_c \in C} \sum_{z=1}^{K} P(w_c|z) P(z|d_D). \end{aligned} \tag{5.26}$$

   where $w_c$ are the caption words, $P(w_c|z)$ the conditional distribution of each $w_c$ given a topic $z$, and $P(z|d_D)$ the conditional distribution of $z$ given $d_D$, the document we wish to illustrate. This approach emphasizes similarity through topic modeling; the document in question has a topic distribution which is likely to have generated the set of words associated with the captions. We report results with 1,000 topics.

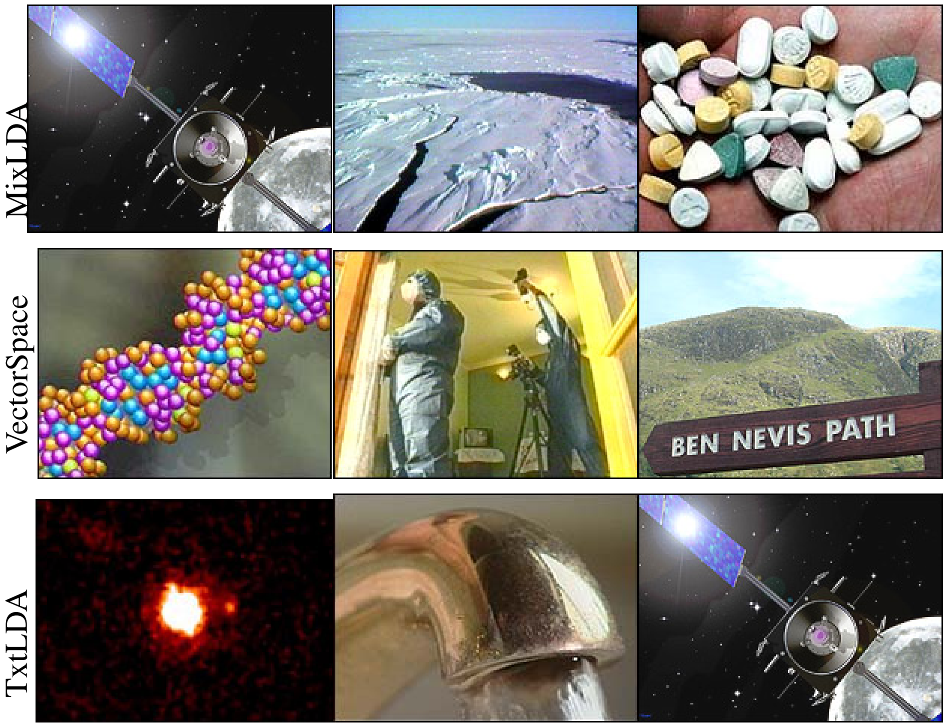
**Evaluation** In the text illustration task, we are given a test document $d$ and a pool of candidate images $I_{1...N}$ with captions $C_{1...N}$. The model is expected to find an image from the candidate pool that matches the test document. We use Equation (5.25) to output a ranked list of $M_I$ visual terms. The image having the highest overlap with the top 30 visual terms is selected as the illustration for the test document. All illustration models were evaluated using top 1 accuracy, which is the percentage of successfully matched image-document pairs in the test set.

### 5.2.2 Results

Table 5.2 presents our results on the automatic text illustration task. Here, we compare our multimodal topic model (MixLDA) against three text-based baselines, namely

| Model | Accuracy |
|---|---|
| TxtLDA | 31.0 |
| Overlap | 31.3 |
| VectorSpace | 38.7 |
| MixLDA | 57.3 |

Table 5.2: Text Illustration results on the BBC News database.



Europe's lunar satellite, the Smart 1 probe, is about to end its mission by crashing onto the Moon's surface. It will be a spectacular end for the robot which has spent the past 16 months testing innovative and miniaturized space technologies. Smart 1 has also produced detailed maps of the Moon's chemical make-up,to help refine theories about its birth. The impact, which will be watched by professional and amateur telescopes, is set for 0543 GMT (0643 BST) on Sunday. The robotic craft should come down on the nearside at mid-southern latitudes, in an area called the Lake of Excellence.

Figure 5.9: Top three images delivered by MixLDA, Overlap, and TxtLDA for illustrating the document (abridged version) in the bottom row.

word overlap (Overlap), a standard vector space model (VectorSpace), and TxtLDA. We examined whether differences in performance are statistically significant using a $\chi^2$ test. As can be seen, MixLDA significantly ($p < 0.01$) outperforms these models by a wide margin (accuracy is 57.3% for MixLDA vs. 31.0% for TxtLDA, 38.7% for the vector space model, and 31.3% for word overlap). These results are encouraging given the simplicity of our model. They also indicate that substantial mileage can be gained by taking into account the visual modality.

Figure 5.9 shows the three best illustrations found by MixLDA, VectorSpace (incidentally, Overlap delivered the same ranking as VectorSpace) and TxtLDA. The images are presented in ranked order, i.e., the first image was given a higher score than the second one, etc. The document discusses Smart 1 Probe, a lunar satellite about to end its mission by crashing onto the moon's surface. MixLDA identifies an image depicting this satellite. The second best picture is also relevant, it resembles the moon's surface. The VectorSpace model does not find any related images, the first one is a DNA image, the second one depicts policemen at a crime scene and the third one Ben Nevis, the highest mountain in the British Isles.

## 5.3 Generalization

In order to investigate the generalization ability of our MixLDA model, we also evaluate its image annotation performance on two further datasets which we downloaded from the CNN news website and Yahoo! news website, repectively. The former is similar in size to the BBC dataset (2881 image-caption-document tuples for training, 240 for testing and 240 for development) while the latter contains about 9000 tuples for training, 2000 for testing and 3000 for training. We adopted a similar parameter tuning procedure as mentioned in Section 5.1.4. On the CNN dataset, our MixLDA model was trained on 1000 topics with 2000 visual words and achieved a precision of 19.2%, a recall of 35.5% and F1 score of 25.5%; on the Yahoo! dataset, the MixLDA was trained on 500 topics with 750 visual words and got a precision of 16.0%, a recall of 38.6% and F1 score of 22.63%, both of which are comparable to our results on the BBC news dataset (see Table 5.1). This indicates that our MixLDA model behaves consistently across different news image datasets.

## 5.4 Summary

In this chapter, we have presented a probabilistic image annotation model. The latter exploits the synergy between the visual and textual modalities inherent in the news image dataset and postulates that visual terms and textual words are generated by common (hidden) topics which are captured probabilistically through Latent Dirichlet Allocation. We can thus represent visual or textual meaning as a distribution over topics and further compute predictive word probabilities given an image by taking these distributions into account. We have evaluated our model on the BBC news dataset and experimentally shown that it is robust to the noise inherent in such data. It improves upon competitive approaches, including the extend continuous relevance model introduced in Chapter 4 and other topic model-based models. We also show that with minor modifications the model can be used to automatically illustrate a document with an appropriate image. Our approach shows promise for multimodal search and image retrieval and other applications which have been traditionally text-based.

In the following chapter, we will present an image caption generation model that employs the keywords extracted by the image annotation model discussed here to represent the content of the image and further render it in natural language.

# Chapter 6

# Image Caption Generation

Automatic image caption generation is of great interest to many image related applications. Examples include image search engines and tools for helping people with visual impairment to access multimedia information in the same way as sighted people. However, relatively little work has focused on the interplay between visual and linguistic information in literature. Existing efforts often follow a two-step natural language generation framework consisting of content selection and surface realization. As we discussed in Chapter 2, previous approaches generate descriptions for images or graphics. The former usually analyze the image content, suggest keywords according to a manually created dictionary containing visual and textual correspondences, and then realize the extracted content into human-readable sentences with the help of predefined sentence-templates or grammars (Abella et al., 1995; Kojima et al., 2002, 2008; Héde et al., 2004; Yao et al., 2009). The latter do not analyze visual features[1], and focus on how to convey information embedded in the graphics (e.g., by choosing sentence patterns to describe the trend of a stock present in a financial graphics) (Mittal et al., 1998, 1995; Corio and Lapalme, 1999; Fasciano and Lapalme, 2000; Feiner and McKeown, 1990).

Generally speaking, a common theme across different formulations is their reliance on manually created resources such as a hand-crafted visual and textual correspondence dictionary, fine-gained knowledge bases, predefined sentence-templates or grammars. The development of such resources requires significant human involvement. For example, one has to manually label every word in the vocabulary with the visual features extracted from its corresponding image regions. Similarly, for the

---

[1] Approaches falling in this type usually assume that the data used to draw the graphics is already structured and available at hand.

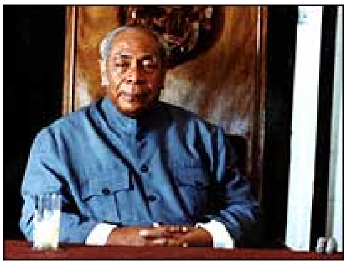
surface realization step, various sentence templates in the desired domain need to be manually constructed beforehand.

Recall that we aim to perform our image caption generation task with little human involvement. In the previous chapters, we have shown that it is possible to learn the correspondence between visual and textual modalities from a database that is not explicitly labeled by human annotators. We then presented a probabilistic image annotation model that learns to automatically annotate keywords for an image and its accompanying news document. We argued that the extracted annotation keywords can be considered as an approximation of the image's content. What remains is to render the extracted content into natural language sentences, again, in a knowledge-lean fashion.

In this chapter, we explore the feasibility of creating captions, using annotation keywords, for images associated with news documents. The availability of the accompanying news documents in our dataset enables us to formulate the generation module so that it resembles text summarization. We then propose both *extractive* and *abstractive* caption generation models. The backbone for both approaches is the probabilistic image annotation model presented in Chapter 5 that suggests content for an image given this image and its associated document. We can then simply identify (and rank) the sentences in the document that share these keywords or create a new caption that is potentially more concise but also informative and fluent. Specifically, for *extractive* models, we examine how to establish criterions for selecting sentences that are similar to the image content. We also present abstractive caption generation models that operate 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 both approaches generate readable captions with little human involvement. Our abstractive model defined over phrases yields more grammatical output than word-based methods.

In the remainder of the chapter, we first recap our task and motivate why we can perform image caption generation in a fashion similar to text summarization. Next, we present our extractive and abstractive models and finally discuss our results.

Thousands of Tongans have attended the funeral of King Taufa'ahau Tupou IV, who died last week at the age of 88. The ceremony in the capital, Nuku'alofa, combined Christian and traditional rituals. 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 IV ruled the Pacific nation for more than four decades, and was much loved by his people. But his death is likely to fuel calls for greater reform. Nuku'alofa came to a standstill as the people of Tonga said goodbye to their revered leader. Buildings, roadsides and palm trees were covered in the customary black...

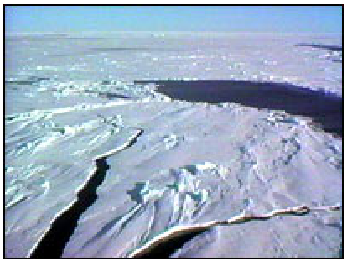


**King Tupou, who was 88, died a week ago.**

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. Even when parents had put controls on what youngsters could access, almost half the 1,003 children aged 11 to 16 surveyed said they could disable them. The NCH said families had to learn more about technology to protect children. 'Worldly wisdom' A tenth of the 11-year-olds who took part in the survey said their parents did not know about the people with whom they communicated online. And 13% revealed they were never supervised while using ...

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

A Nasa satellite has documented startling changes in Arctic sea ice cover between 2004 and 2005. The extent of "perennial" ice - thick ice which remains all year round - 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. The drastic shrinkage may relate partly to unusual wind patterns found in 2005, though rising temperatures in the Arctic could also be a factor. The research is reported in in the journal Geophysical Research Letters. The Arctic is average. Perennial decay Recent studies have shown that the area of the Arctic covered by ice each summer, and the ice thickness, have been shrinking. This latest study, from ...



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

Table 6.1: Each entry in the BBC News database contains a document, an image, and its caption (shown in boldface).

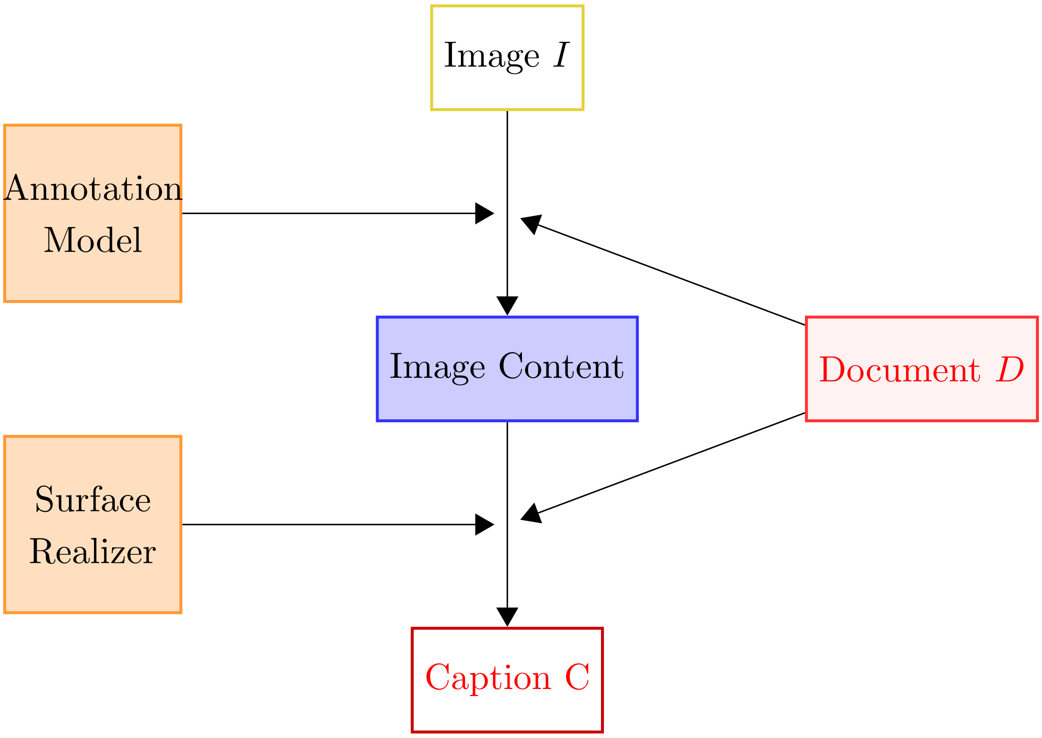


Figure 6.1: This figure shows a pipeline for the task of automatic caption generation for news images.

## 6.1 Background

Apparently, generating image captions is a challenging task even for humans, let alone computers. Journalists are given explicit instructions on how to write captions.[2] 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.

Here we repeat our formulation for the task of news image caption generation from Chapter 3:

**Definition.** *Given a news image I, and its associated news document D, create a natural language caption C that captures the main content of the image given D.*

---

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

Compared to our definition for generic image description generation (see Definition 1 or Figure 2.2 in Chapter 2), caption generation (as shown Figure 6.1) does not rely on a predefined knowledge base $\kappa$. In this setup, the training data consists of document-image-caption tuples like the ones shown in Table 6.1. During testing, we are given an image and its associated document, and we must generate a caption for the image without using manually created knowledge bases.

As we discussed before, our image annotation model can learn the multimodal correspondence between visual and textual information and suggest content keywords for a given image and its accompanying news document. Now, the remaining task is to render this content into human-readable sentences, again, in a knowledge-lean way.

Without access to the accompanying news documents, we would be exposed to a normal natual language generation paradigm, where we would unavoidably require fine-gained knowledge bases to assign semantic relations among keywords, and further to create natural language sentences. Fortunately, the availability of the accompanying news documents allow us to reuse document constituents (e.g., words, phrases or even sentences) to form a natural language caption. This bears some resemblance with text summarization, which condenses a source text into a shorter target text whilst preserving the gist of the source text. For example, an image caption, which is usually expected to be succinct and establish the picture's relevance to the events described in the document, is similar to news headlines[3]. The latter capture the most important points of the news stories using a few words. All these considerations indicate that we can perform our caption generation task in a fashion akin to text summarization. Importantly, under the text summarization framework, our caption generation module can produce expressive and flexible sentences with little reliance on manually created resources. As the generated captions must be faithful to the images' content, our models differ from most previous work in summarization which is solely text-based.

Following the main streams of work in text summarization discussed in Chapter 2, we can formulate the caption generation task both extractively and abstractively. However, before we proceed to present our caption generation models, we first talk about their common prerequisite, extracting image content, which is essentially our annotation model introduced in Chapter 5.

---

[3] Headline generation is considered as an instantiation of abstractive text summarization.

## 6.2 Content Extraction

As mentioned above, our image caption generation models will rely on an image annotation model to analyze the image content with annotation keywords. In Chapter 2, we have reviewed different paradigms of image annotation models ranging from supervised to unsupervised, probabilistic and non-probabilistic ones. The form of annotation outputs varies accordingly, e.g., non-probabilistic models usually output a ranked keyword list as the image content while probabilistic ones can output a distribution over the vocabulary or, sometimes, a distribution over a set of latent variables, e.g., topics (as the one in Chapter 5).

Here, we place emphasis on generative probabilistic models, specifically, we make use of the model presented in Chapter 5 which is well-suited to the caption generation 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.

As we discussed in Chapter 5, 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) and subsequently quantized into a discrete set of visual terms using the $K$-means clustering algorithm. 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:

$$\begin{aligned} W_I^* &= \arg\max_{W_t} P(W_t|I,D) && (6.1)\\ &\approx \arg\max_{W_t} \prod_{w_t \in W_t} P(w_t|d_{Mix}) && (6.2)\\ &= \arg\max_{W_t} \prod_{w_t \in W_t} \sum_{k=1}^{K} P(w_t|z_k)P(z_k|d_{Mix}) \end{aligned}$$

$W_t$ 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 for documents.

Given an unseen image-document pair and a trained model, it is possible to infer the posterior of topic proportions over the new data by maximizing the likelihood. The model will then output the probabilities over all textual words $\{w_t\}$. These probabilities can be naturally adjusted into a $n$-best keyword list to approximate the image content. Alternatively, we can treat this distribution over the whole vocabulary as the image content directly, which can also be considered as a unigram language model. This probabilistic representation is superior to the ranked keyword list. The latter is subject to the choices of $n$ and sometimes not robust enough under noisy conditions. annotations. The former can provide more complete information about the image content and its probabilistic nature makes it easier integrate with probabilistic generation models.

## 6.3 Extractive Image Caption Generation

Extractive models have received much attention in text summarization due to their simplicity and efficiency. A summary is formed simply by selecting and concatenating sentences from the source document. Thus, without a great deal of manual effort, summaries can be created for different languages and text genres. As the sentences are extracted verbatim from the source, they are grammatical (extracts can be, however, incoherent and contain much redundant information) and therefore extractive summarization methods place more emphasis on how to identify the most informative sentences.

With respect to our image caption generation task, we only need to extract a sentence, that is maximally similar to the image and its content as described by the image

**Cadbury will increase its contamination testing levels.**

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's recalled one million chocolate bars in June because of salmonella fears. The firm has blamed a leaking pipe at its Marlbrook plant in Herefordshire for the salmonella contamination. The seven brands affected by the recall were the 250g Dairy Milk Turkish, Dairy Milk Caramel and Dairy Milk Mint bars, the Dairy Milk 8 chunk and the 1kg Dairy Milk bar as well as the 105g Dairy Milk Buttons Easter Egg and the Freddo bar. People interviewed A cleaning ... The firm originally said it had recalled the bars purely as a precautionary measure. "The levels are significantly below the standard that would be any health problem, but we are taking this measure as a precaution," a spokesman said at the time.

Table 6.2: An example entry from the BBC dataset (caption is shown in boldface). Sentences with different types of underline indicate different extracts according to various models (double line for KL based model, single line for vector based model, waveline for word overlap).

annotation model. Given the probabilistic nature of our image annotation model, we are able to take advantage of two image content representations, 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 content representations.

**Word Overlap based Sentence Selection** Perhaps the most intuitive way of measuring the similarity between image keywords and document sentences is word overlap:

$$Overlap(W_I, S_d) = \frac{|W_I \cap S_d|}{|W_I \cup S_d|} \tag{6.3}$$

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

**Vector Space Model 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 (Salton and McGill, 1983) and compute the similarity between the two vectors, representing the image keywords and document sentences, respectively. Each entry in the image keyword vector is weighted by $tf-idf$ value. If a keyword does not appear in the document, then it will be assumed to appear once. Document sentences are represented with 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. 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:

$$sim(\overrightarrow{W_I},\overrightarrow{S_d}) = \frac{\overrightarrow{W_I}\cdot\overrightarrow{S_d}}{|W_I||\overrightarrow{S_d}|} \tag{6.4}$$

**Sentence Selection based on Topic Distribution** In probabilistic topic models, the similarity between two documents can be measured by the extent to which they share common topics (Steyvers and Griffiths, 2007). 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. For example, we may use the KL divergence to measure the difference between two distributions $p$ and $q$:

$$D(p,q) = \sum_{j=1}^{K} p_j \log_2 \frac{p_j}{q_j} \tag{6.5}$$

where $p$ and $q$ are shorthand for the image topic distribution $P_{d_{Mix}}$ and sentence topic distribution $P_{S_d}$, respectively. As described in Chapter 5, 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 $q$:

$$JS(p,q) = \frac{1}{2}\left[D(p,\frac{(p+q)}{2}) + D(q,\frac{(p+q)}{2})\right] \tag{6.6}$$

## 6.4 Abstractive Image 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. Secondly, 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.

### 6.4.1 Word-based Caption Generation

Banko et al. (2000) (see also Witbrock and Mittal 1999) propose a bag-of-words model for headline generation. Following the traditional NLG 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 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 output of reasonable length (around 5 words):

$$\begin{aligned} P(w_1, w_2, ..., w_n) = & \prod_{i=1}^{n} P(w_i \in H | w_i \in D) \\ & \cdot P(len(H) = n) \\ & \cdot \prod_{i=2}^{n} P(w_i | w_{i-1}) \end{aligned} \tag{6.7}$$

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. Banko et al. (2000) assume that headline length follows a normal distribution which they learn from a corpus (1997 Reuters News Stories).

The above model can be easily adapted to our image 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 6.2). In addition, we replace the bigram language model with a trigram one:

$$\begin{aligned} P(w_1, w_2, ..., w_n) = & \prod_{i=1}^{n} P(w_i \in C | I, D) \\ & \cdot P(len(C) = n) \\ & \cdot \prod_{i=3}^{n} P(w_i | w_{i-1}, w_{i-2}) \end{aligned} \tag{6.8}$$

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 (6.8) has a major drawback. Bear in mind that most image annotation models consider only content words — it does not make sense to output function words as they are not descriptive of the image content. This means that the content selection component will naturally tend to ignore these non-descriptive words. 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. Intuitively, we hope the new language model will prefer words that have high image annotation probabilities while are likely to appear in a sentence according to the background language model. Specifically, we use an *adaptive* language model (Kneser et al., 1997) that modifies an $n$-gram model with local unigram probabilities:

$$\begin{aligned} P(w_1, w_2, ..., w_n) = & \prod_{i=1}^{n} P(w_i \in C | w_i \in D) \\ & \cdot P(len(C) = n) \\ & \cdot \prod_{i=3}^{n} P_{adap}(w_i | w_{i-1}, w_{i-2}) \end{aligned} \tag{6.9}$$

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:

$$P_{adap}(w|h) = \frac{\alpha(w)}{z(h)} P_{back}(w|h) \tag{6.10}$$

$$\alpha(w) \approx (\frac{P_{adap}(w)}{P_{back}(w)})^{\beta} \tag{6.11}$$

$$z(h) = \sum_{w} \alpha(w) \cdot P_{back}(w|h) \tag{6.12}$$

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 $\beta$ a scaling parameter.

The model in (6.9) has three components. The conditional probability $P(w_i \in C|w_i \in D)$ captures the key points in the article, whereas the adapted language model $P_{adap}(w_i|w_{i-1}, w_{i-2})$ is responsible for the grammatical issue, meanwhile, highlights a subset of the words related to the image content. The length component $P(len(C) = n)$, modeled as a normal distribution, modulates the caption length.

### 6.4.2 Phrase-based Caption Generation

The model outlined in equation (6.9) 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 summarization as the building blocks of abstracts. For example, Zhou and Hovy (2003), first, extract keywords according to word frequency, position feature and the conditional probability of a word appearing in the headline given that the same word appears in the document (the ones used by Jin and Hauptmann (2002)), then glue each keyword with its left and right neighbors in original text and further define a phrase as a word window that covers a series of keyword chunks in the document. This could give pieces of text that are locally grammatical while containing much topic-related information. However, they have to rely on a set of hand-written rules in order to obtain globally coherent headlines. Soricut and Marcu (2006) use the same algorithm as Zhou and Hovy (2003) to extract keywords, then collect phrases from lexical dependencies that the keywords have, and assign each phrase

| Heads | Dependencies |
|---|---|
| Nouns | amod, det, nn, poss |
| Verbs | dobj, iobj, nsubj, nsubjpass, aux, auxpass, xsubj |
| Prepositions | prep, pobj |

Table 6.3: Selected dependencies[1] for phrase extraction.

with an associated probability according to its frequency in the document. Again, the phrases are treated in isolation without paying attention to the relations among phrases.

In our word-based models, words are the basic units for both content selection and surface realization. As previous work show (Zhou and Hovy, 2003; Soricut and Marcu, 2006), it is relatively straightforward to extend content selection from words to phrases. Using multi-words as the basic content unit poses, however, difficulties for surface realization since language model-based realizers are built from individual words rather than phrases, where constraints between phrases are not modeled. As a result, the coherence and grammaticality of the resulting sentences may suffer. A phrase-based caption generation model in the following extend it so as to take phrase dependency constraints into account.

**Phrase Extraction** Our model relies on phrases which we obtain 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). We also record the dependent that a given head co-occurs with. This gives us an idea of the type of syntactic relations that the head words are involved in.

For example, from the first sentence in the top block of Table 6.1, we would extract the phrases: *thousands*, *of Tongans*, *attended*, *the funeral*, *King Taufa‘ahau Tupou IV*, *died*, *last week*, *at the age*, and so on. We only consider dependencies whose heads are nouns, verbs, and prepositions (shown in Table 6.3), 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

---

[1] We used the Stanford Parser (Klein and Manning, 2003) (`http://nlp.stanford.edu/software/lex-parser.shtml`) to extract the typed dependencies. The Stanford typed dependency representation is designed to describe the grammatical relations between pairs of words, e.g., in Figure 6.2, the direct object of *attended* is *funeral*.

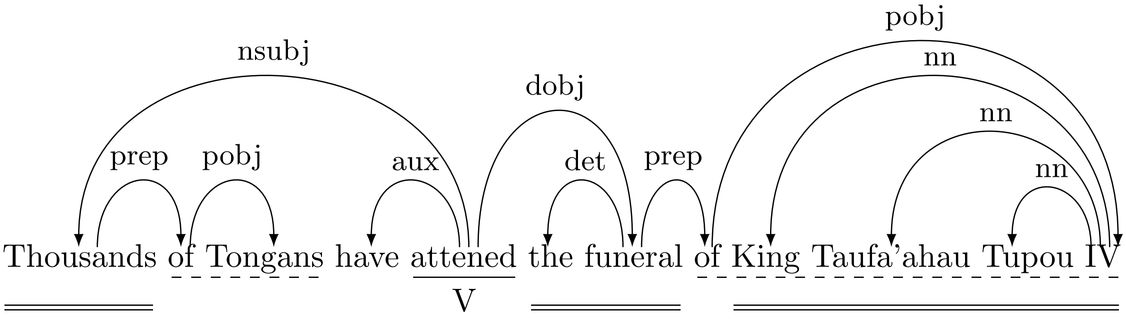

nsubj
prep
pobj
aux
dobj
det
prep
pobj
nn
nn
nn
Thousands of Tongans have attened the funeral of King Taufa'ahau Tupou IV
V

Figure 6.2: Dependencies for sentence: “*Thousands of Tongans have attended the funeral of King Taufa'ahau Tupou IV.*“ We can extract phrases from the sentence fragments: dashed underline for preposition, double underline for noun, and single underline for verb.

selection and caption length components in equation (6.9) as follows:

$$\begin{aligned} P(\rho_1,\rho_2,...,\rho_m) \approx & \prod_{j=1}^{m} P(\rho_j \in C|\rho_j \in D) \\ & \cdot P(len(C) = \sum_{j=1}^{m} len(\rho_j)) \\ & \cdot \prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i|w_{i-1},w_{i-2}) \end{aligned} \tag{6.13}$$

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

$$P(\rho_j \in C|\rho_j \in D) = \prod_{w_j \in \rho_j} P(w_j \in C|w_j \in D) \tag{6.14}$$

where $w_j$ is a word in the phrase $\rho_j$.

**Phrase Adjacency** The content component of our model rests on the hypothesis that words or phrases are independent of each other, and it is therefore up to the trigram language model to enforce coarse ordering constraints. This 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 $\rho_j$ attaching to the right of phrase $\rho_i$ as:

$$\begin{aligned} P(\rho_j|\rho_i) &= \sum_{w_i \in \rho_i} \sum_{w_j \in \rho_j} p(w_j|w_i) \\ &= \frac{1}{2} \sum_{w_i \in \rho_i} \sum_{w_j \in \rho_j} \{ \frac{f(w_i,w_j)}{f(w_i,-)} + \frac{f(w_i,w_j)}{f(-,w_j)} \} \end{aligned} \tag{6.15}$$

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.[4]

After integrating the attachment probabilities into equation (6.13), the caption generation model becomes:

$$\begin{aligned} P(\rho_1,\rho_2,...,\rho_m) \approx & \prod_{j=1}^{m} P(\rho_j \in C|\rho_j \in D) \\ & \cdot \prod_{j=2}^{m} P(\rho_j|\rho_{j-1}) \\ & \cdot P(len(C) = \textstyle\sum_{j=1}^{m} len(\rho_j)) \\ & \cdot \prod_{i=3}^{\sum_{j=1}^{m} len(\rho_j)} P_{adap}(w_i|w_{i-1},w_{i-2}) \end{aligned} \tag{6.16}$$

On the one hand, the model in equation (6.16) 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 (6.9)) and $P(\rho_1,\rho_2,...,\rho_m)$ for the phrase-based model (equation (6.16)). 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 6.3, we may rank its sentences in terms of their relevance to the image contents 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.

---

[4] Equation (6.15) is smoothed to avoid zero probabilities.

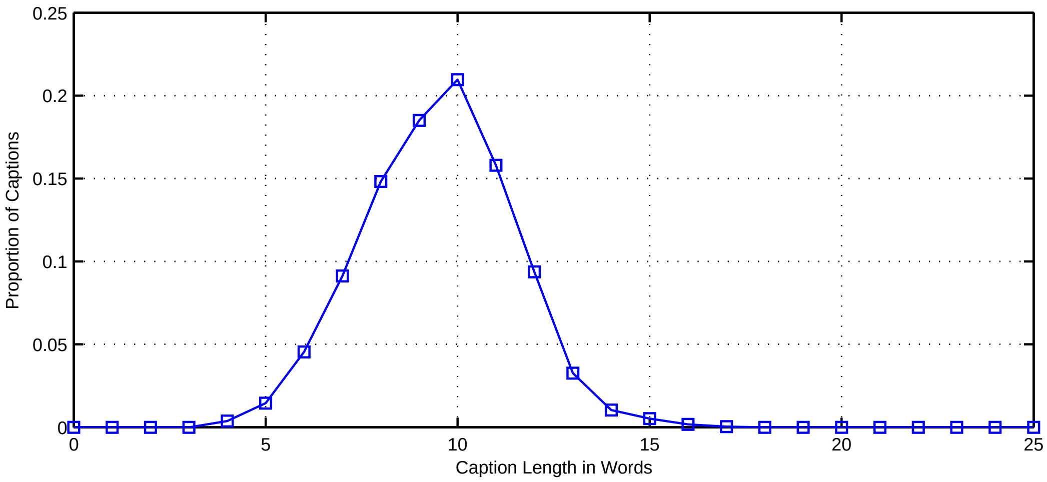


Figure 6.3: Distribution of caption length in the BBC dataset

## 6.5 Experiments

### 6.5.1 Setup

In this section we will 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 how we evaluate our models both automatically and manually.

#### Data

All our experiments were conducted on the BBC dataset ( 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. These parameters are tuned in our development set. For the abstractive models, we used a trigram model trained with the SRI language model toolkit[5] on a newswire corpus consisting of BBC and Yahoo! news documents (6.9 M words). The attachment probabilities (see equation (6.15)) 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 Gaussian distribution (shown in Figure 6.3). The scaling parameter $\beta$ for the adaptive language model was also tuned on the development set using a range of [0.5,0.9]. We report results with $\beta$ set to 0.5, which is expected to balance the image annotation model and the background language model. 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 5 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. We chose to use TER over other metrics with similar properties such as BLEU (Papineni et al., 2002) since it can account for word reordering and be applied to individual sentences without any adjustments. 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:

$$\mathrm{TER}(E,E_r) = \frac{\mathrm{Ins}+\mathrm{Del}+\mathrm{Sub}+\mathrm{Shft}}{N_r} \tag{6.17}$$

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 alignment is usually found through beam search. We used

---

[5] `http://www.speech.sri.com/projects/srilm/`

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 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 (KL divergence based), 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 (Keller et al., 2009) experimental software.

### 6.5.2 Results

Table 6.5 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 (6.9)) and phrases (see equation (6.16)). 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, $p < 0.01$).[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 varia-

---

[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).

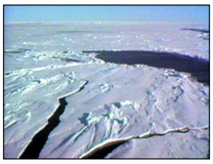
| | |
|---|---|
| | G : King Tupou, who was 88, died a week ago.<br>KL: Last year, thousands of Tongans took part in unprecedented deonstrations to demand greater democracy and public ownership of key national assets.<br>$A_W$: King Toupou IV died at the age of Tongans last week.<br>$A_P$: King Toupou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

Table 6.4: Captions written by humans (G) and examples produced by our systems: KL: KL divergence based extractive model, $A_W$: word-based abstractive model, and $A_P$: phrase-based abstractive model.

| 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 6.5: TER results for extractive, abstractive models, and lead sentence baseline; $^{*}$: significantly different from lead sentence; $^{\dagger}$: significantly different from KL and JS divergence.

| 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 6.6: Mean ratings on caption output elicited by humans; $^{*}$: significantly different from word-based abstractive system; †: significantly different from phrase-based abstractive system.

tion 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 6.6 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 6.6 (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 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 6.1 are given in Table 6.4.

## 6.6 Summary

In this chapter, we have presented a novel task, automatic caption generation for news images, and proposed extractive and abstractive 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.

# Chapter 7

# Conclusions and Future Directions

In this chapter, we conclude the thesis with our major findings and contributions, and discuss possible directions for future research.

## 7.1 Findings

In this thesis, we have focused on the task of automatically generating captions for news images. As a departure from previous work, we have approached this task in a knowledge-lean way that relies on little human involvement. This is manifest in the dataset we employed, and the way we extract image content and render it into natural language form. We summarize the main contributions of our work below:

1. We have introduced a new task, automatic caption generation for news images, that fuses insights from computer vision and natural language processing. We approached the task in a learning-from-data fashion. We built our dataset from resources that are publicly available on the internet without manual post-processing. Moreover, during modeling we did not make use of fine-gained knowledge bases. We extracted the image content by building a probabilistic image annotation model whose output we then use to generate captions with the help of the news documents accompanying the image. Importantly, our generation model does not rely on manually created sentence templates or grammars. Our experiments showed that it is possible to create a caption generation model from such a noisy dataset and to perform the task without much human involvement.

2. We addressed the data acquisition bottleneck associated with image related ap-

plications by exploiting data resources where images and their textual descriptions co-occur naturally. Specifically, we built the BBC news dataset consisting of news articles, images, and their captions. Our dataset differs from traditional image datasets in several aspects. It contains real-world images and employs 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. The latter are admittedly noisy yet can serve as a gold standard for caption generation; and lastly, our dataset contains a unique component, the news document, which provides both information regarding to the image's content and how it should be rendered. We argued that this dataset is suitable as a testbed for our task based on the observation that news images, their captions and corresponding documents are closely related in content. In addition, news documents can provide rich linguistic information required for the generation procedure.

3. We have adapted the continuous relevance image annotation model (Lavrenko et al., 2003) to our dataset by taking into account the accompanying news documents in two ways. We first smoothed the probability of a keyword given an image with its likelihood of appearing in the associated document, and further pruned the model's output with a topic model that is trained on the document collection. Our experimental results showed that it is possible to learn the correlations between visual and textual information from our dataset even if it is not explicitly annotated in any way.We also found that the associated news documents are able to benefit the annotation model, beyond the captions and images.

4. In order to extract the image content, we have presented a probabilistic image annotation model that exploits the synergy between visual and textual modalities. Specifically, it assumes that visual terms and annotation keywords are generated by a set of latent variables (topics) which are captured probabilistically by Latent Dirichlet Allocation (Blei et al., 2003). We thus represented visual and textual meaning probabilistically as a distribution over topics. By utilizing these topic distributions, we computed the predictive word probabilities given an image and its accompanying document. Our experiments have demonstrated that this probabilistic model is robust to the noise inherent in our dataset and is useful on its own right, not only for the caption generation task. Our model improved upon competitive image annotation approaches including our exten-

tion of the continuous relevance model and other topic model-based approaches, such as CorrLDA (Blei and Jordan, 2003) and PLSA-based models (Monay and Gatica-Perez, 2007). We also demonstrated how this framework can be straightforwardly modified to perform automatic text illustration with encouraging results.

5. Finally, with the help of the accompanying news documents, we demonstrated that the caption generation task can be formulated in a fashion akin to summarization. We proposed both extractive and abstractive models that do not rely on fine-gained sentence-templates or grammars. A key aspect of our approach is to allow both the visual and textual information to influence the generation task. In this sense, our approach differs from vanilla text summarization since the visual information plays an important role in content selection. Visual information is represented through our probabilistic annotation model that characterizes the content of the image in the form of annotation keywords or distributions over keywords or latent topics, which are subsequently used to guide the generation process. Given the suggested image content, our extractive models select a sentence from the accompanying document as the image caption, while our abstractive models create a new sentence from scratch. Our experiments showed that extracting a sentence from the document often yields an inferior caption, while a probabilistic abstractive model defined over phrases yields promising output which manages to capture the gist of the image.

## 7.2 Future Research Directions

In this thesis, we explored the feasibility of automatically generating captions for news images in a knowledge-lean way. We exploited resources from the BBC news website and formulated the task in a two-step fashion, namely image content extraction and surface realization. Avenues for future research are many and varied.

The BBC news dataset discussed here is only one instantiation of similar publicly available resources. These include Yahoo! News, CNN News, Wikipedia, and so on. These are all examples where images and their textual descriptions co-occur naturally. These resources can be used to prompt image-related research, such as text-based image retrieval, content-based image retrieval, story picturing, image browsing support, etc. Besides these multimedia resources, it is also worth looking at other venues where images and their corresponding textual information are domain specific and in a clear structure. For example, Ahmed et al. (2009) focus on figures and their captions in the biological literature and propose a structured correspondence LDA model that take into account the inherent structure of the figures and their captions in life science publications to capture the correspondence between the visual and textual modalities. As video processing usually involves processing key frames (images) from streaming video data, it is also possible to adapt existing models and applications from images to video (e.g., automatic video summarization).

The dataset discussed in this thesis can be further refined according to specific applications to eliminate some of the noise. For example, if the dataset is collected for the purpose of object recognition or image retrieval (both tasks require more accurate and complete annotations), a possible refinement would be to use the news document to increase the annotation keywords by identifying synonyms or even sentences that are similar to the image caption. In addition, one would make use of the characteristics of specific domains (e.g., captions of figures in the bio-science literature usually mention protein names, tissue labels, etc., related to the figure and described in the article), or filter the data with other discriminative models (Schroff et al., 2007). Currently, our analysis of the accompanying document is only limited to part of speech tagging, there are good reasons to expect that more sophisticated processing (e.g., named entity recognition, parsing, word sense disambiguation, etc.), would improve the quality of the dataset.

In this work, images were preprocessed by extracting primarily local feature representations (e.g., color, texture, corners, SIFT features, etc.), without considering more global representations, such as spatial relationship among different regions. An obvious extension would be taking spatial information into account when dealing with image representations. Currently, we treat the image regions or detected regions of interest as bags-of-words, which could be extended to bigrams according to their spatial relations.

In Chapter 5, we presented a topic based image annotation model where the number of topics is a parameter to be optimized experimentally, a procedure which must be repeated for different image collections. The model could be improved to allow an infinite number of topics and evolve to a nonparametric version that *learns* how many topics are optimal. Furthermore, currently the model is based on word unigrams, and thus takes 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 (Boyd-Graber and Blei, 2009).

In Chapter 6, we demonstrated how we performed the generation task both extractively and abstractively in a summarization framework, by considering the image annotation probabilities, a language model and shallow syntactic information. Our approach would potentially 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).

The image caption generation task has been formulated as a two-step approach, where the image content extraction 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.

# Appendix A

# Human Evaluation for Image Caption Generation Systems

This appendix includes the instruction presented to our subjects in the human evaluation studies for comparing our image caption generation systems (see Chapter 6).

## A.1 Evaluation Instructions

In this experiment you will be presented with a news image, an article associated with the image, and a caption describing the image. Your task is to judge how well the caption describes the content of the image given the accompanying article and how grammatical the caption is. Some captions will seem appropriate to you, but others will not. You will make your judgement by choosing a rating from 1 (the caption is not appropriate, or the caption is not grammatical at all) to 7 (the caption is appropriate, or the caption is grammatical). All captions were generated automatically by a computer program.

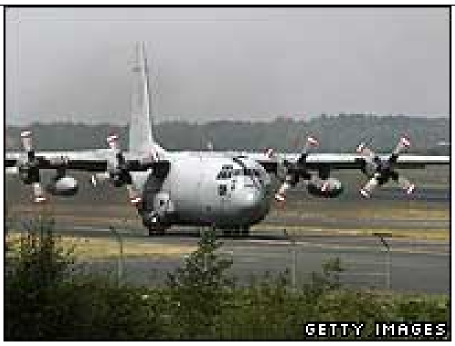
## A.2 Example

For example, if you were presented with the document, image, and caption shown in Figure A.1:

You would probably give the caption in bold a higher content rating (e.g., 6 or 7) since it is relevant both for the image and the document. Indeed, the image shows a plane in an airport and the article discusses US planes landing in UK airports with bombs. Even though the words bombs, US and UK are not explicitly depicted in the

**Document:** The British government could face claims it violated international humanitarian laws by allowing US arms flights to Israel to use UK airports.

The Islamic Human Rights Commission (IHRC) is seeking permission to contest government bodies over what it says are crimes against the Geneva Convention. A number of US planes said to be carrying bombs to Israel refueled in the UK during the Lebanon conflict. The IHRC said it received complaints from Britons with families in Lebanon. The commission is accusing the government of ”grave and serious violations” of international humanitarian law. It is seeking permission to bring its case against the Civil Aviation Authority, the Foreign and Commonwealth Office and Defence Secretary Des Brown in the High Court. The IHRC said it is bringing the case after receiving ”many complaints ... from British citizens whose family members are in Lebanon and facing grave danger as well as acts of terror”. The US aircraft believed to have refueled in the UK are said to have been carrying supplies including ”bunker buster bombs”.



**Caption:** A US plane landed in a UK airport with bunker buster bombs

Figure A.1: Example of a document, image and caption presented to the subjects.

image, they are related to the accompanying article which discusses how the British government allowed US planes to refuel in the UK while carrying bombs to Israel. If a caption is neither related to the image nor to the article, then it should receive a lower content rating. If the caption is grammatical, then you should rate it a higher content score. If it is not fluent and reads like word salad, then you should give it a lower rate.

## A.3 Interface

You will be presented with the document, the image, and the caption. Once you read the document, look at the picture and read its caption, please make your judgement by selecting a number between 1 and 7. Each number will be represented by a button, all you have to do is click the button corresponding to your judgement. There are no ’correct’ answers, so whatever numbers seem appropriate to you are a valid response. While you are deciding the rating, try to ask the following questions:

- Does the caption describe information present in the image and the document?
- Does the caption represent the main topic of the document?
- Does the caption depict an object present in the picture?
- Does the caption seem fluent? Is it understandable?

Use high numbers if the answer to the above questions is ’yes’, low numbers if it is ’no’, and intermediate numbers for captions that represent peripheral aspects of the image and document. Try to make up your mind quickly, base your judgments on your first impressions. The experiment will last less than 10 minutes.

# Bibliography

Abella, A., Kender, J. R., and Starren, J. (1995). Description generation of abnormal densities found in radiographs. In *Proceedings of the Symposium on Computer Applications in Medical Care (SCAMC)*, pages 542–546. American Medical Informatics Association.

Ahmed, A., Xing, E. P., Cohen, W. W., and Murphy, R. F. (2009). Structured correspondence topic models for mining captioned figures in biological literature. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 39–48, New York, NY, USA. ACM.

Amazon (2009). https://www.mturk.com/mturk.

Banko, M., Mittal, V., and Witbrock, M. (2000). Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 318–325, Morristown, NJ. Association for Computational Linguistics.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., and Jordan, M. (2002). Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Barnard, K., Fan, Q., Swaminathan, R., Hoogs, A., Collins, R., Rondot, P., and Kaufhold, J. (2008). Evaluation of localized semantics: Data, methodology, and experiments. *International Journal of Computer Vision*, 77(1-3):199–217.

Barnard, K. and Forsyth, D. (2001). Learning the semantics of words and pictures. In *Proceedings of the 8th International Conference on Computer Vision*, pages 408–415, Vancouver, BC.

Barzilay, R. and McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Blei, D. (2004). *Probabilistic Models of Text and Images*. PhD thesis, University of Massachusetts Amherst.

Blei, D. and Jordan, M. (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134, Toronto, ON.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bonnie, D. Z., Dorr, B., and Schwartz, R. (2004). Bbn/umd at duc-2004: Topiary. In *Proceedings of the HLT-NAACL 2004 Workshop on Text Summarization and Document Understanding Conference (DUC 2004)*, pages 112–119.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pêcheux, M. G., Ruel, J., Venuti, P., and Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development*, 75(4):1115–1139.

Bosch, A. (2007). *Image Classification for a Largre Number of Object Categories*. PhD thesis, Universitat de Girona.

Boyd-Graber, J. and Blei, D. (2009). Syntactic topic models. In *Proceedings of the 22nd Conference on Advances in Neural Information Processing Systems*, Vancouver, BC.

Brown, P., Pietra, S. D., Pietra, V. D., and Mercer, R. (1993). The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Carberry, S., Carberry, R., Elzer, S., Green, N., Mccoy, K., and Chester, D. (2004). Extending document summarization to information graphics. In *Proceedings of the ACL Workshop on Text Summarization*.

Carneiro, G. and Vasconcelos, N. (2005). Formulating semantic image annotation as a supervised learning problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 163–168, Washington, DC. IEEE Computer Society.

Chai, C.-F. and Hung, C. (2008). Automatically annotating images with keywords: A review of image annotation systems. *Recent Patents on Computer Science*, 1:55–68.

Clarke, J. (2008). *Global Inference for Sentence Compression: An Integer Linear Programming Approach*. PhD thesis, the University of Edinburgh.

Corio, M. and Lapalme, G. (1999). Generation of texts for information graphics. In *Proceedings of the 7th European Workshop on Natural Language Generation*, pages 49–58.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Survey*, 40(2):1–60.

Datta, R., Li, J., and Wang, J. Z. (2005). Content-based image retrieval – approaches and trends of the new age. In *Proceedings of the International Workshop on Multimedia Information Retrieval*, pages 253–262, Singapore.

Daumé III, H. and Marcu, D. (2006). Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, Morristown, NJ, USA. Association for Computational Linguistics.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the 15th International Conference on Computer Vision and Pattern Recognition*.

Deschacht, K. and Moens, M.-F. (2007). Text analysis for automatic image annotation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1000–1007, Prague, Czech Republic. Association for Computational Linguistics.

Dorr, B., Zajic, D., and Schwartz, R. (2003). Hedge trimmer: a parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 2003 Text Summarization Workshop and Document Understanding Conference (DUC 2003)*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Duygulu, P., Barnard, K., de Freitas, J., and Forsyth, D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision*, pages 97–112, Copenhagen, Danemark.

Elzer, S., Carberry, S., Zukerman, I., Chester, D., Green, N., , and Demir, S. (2005). A probabilistic framework for recognizing intention in information graphics. In *Proceedings of the 19th International Conference on Artificial Intelligence*, pages 1042–1047, Edinburgh, Scotland.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Fan, J., Gao, Y., and Luo, H. (2007). Hierarchical classification for automatic image annotation. In *Proceedings of the 30th annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 111–118, New York, NY. ACM.

Fan, J., Luo, H., and Guo, Y. (2005). Learning the semantics of images by using unlabeled samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 704–710, Washington, DC. IEEE Computer Society.

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

Fei-Fei, L., Fergus, R., and Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Workshop on Generative-Model Based Vision*.

Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531, Washington, DC. IEEE Computer Society.

Feiner, S. and McKeown, K. (1990). Coordinating text and graphics in explanation generation. In *Proceedings of National Conference on Artificial Intelligence*, pages 442–449.

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

Feng, Y. and Lapata, M. (2008). Automatic image annotation using auxiliary text information. In *Proceedings of the 46th Annual Meeting of the Association of Compu-*

*tational Linguistics: Human Language Technologies*, pages 272–280, Morristown, NJ, USA. Association for Computational Linguistics.

Feng, Y. and Lapata, M. (2010a). How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1239–1249, Uppsala, Sweden. Association for Computational Linguistics.

Feng, Y. and Lapata, M. (2010b). Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839, Los Angeles, California. Association for Computational Linguistics.

Feng, Y. and Lapata, M. (2010c). Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99, Los Angeles, California. Association for Computational Linguistics.

Fergus, R., Fei-Fei, L., Perona, P., and Zisserman, A. (2005). Learning object categories from google's image search. In *Proceedings of International Conference on Computer Vision*, pages 1816–1823, Washington, DC, USA. IEEE Computer Society.

Ferres, L., Parush, A., Roberts, S., and Lindgaard, G. (2006). Helping people with visual impairments gain access to graphical information through natural language: The *igraph* system. In *Proceedings of 11th International Conference on Computers Helping People with Special Needs*, pages 1122–1130, Linz, Austria.

Filippova, E. (2009). *Dependency Graph Based Sentence Fusion and Compression*. PhD thesis, EML Research gGmbH.

Griffin, G., Holub, A., and Perona, P. (2007). Caltech 256 object category dataset. Technical Report 7694, California Institute of Technology.

Griffiths, T., Tenenbaum, J., and Steyvers, M. (2007). Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Gupta, S., Kim, J., Grauman, K., and Mooney, R. (2008). Watch, listen & learn: Co-training on captioned images and videos. In *Proceedings of the 2008 European*

*Conference on Machine Learning and Knowledge Discovery in Databases - Part I*, pages 457–472, Berlin, Heidelberg. Springer-Verlag.

Haghighi, A. and Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Morristown, NJ, USA. Association for Computational Linguistics.

Héde, P., Moëllic, P. A., Bourgeoys, J., Joint, M., and Thomas, C. (2004). Automatic generation of natural language descriptions for images. In *Proceedings of Recherche dInformation Assiste par Ordinateur (RIAO)*, Avignon, France.

Herbert Bay, Andreas Ess, T. T. and Gool, L. V. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–259.

Hofmann, T. (1998). Learning and representing topic. A hierarchical mixture model for word occurrences in document databases. In *Proceedings of the Conference for Automated Learning and Discovery*, pages 408–415, Pittsburgh, PA.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 41(2):177–196.

Holub, A. (2007). *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. PhD thesis, California Institute of Technology.

Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 119–126, New York, NY, USA. ACM Press.

Jeon, J. and Manmatha, R. (2004). Using maximum entropy for automatic image annotation. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, pages 24–32, Dublin City, Ireland.

Jin, R., Chai, J., and Si, L. (2004). Effective automatic image annotation via a coherent language model and active learning. In *Proceedings of the 12th annual ACM International Conference on Multimedia*, pages 892–899, New York, NY. ACM Press.

Jin, R. and Hauptmann, A. (2001a). Automatic title generation for spoken broadcast news. In *Proceedings of the 1st International Conference on Human Language Technology Research*, pages 1–3, Morristown, NJ, USA. Association for Computational Linguistics.

Jin, R. and Hauptmann, A. (2002). A new probabilistic model for title generation. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Jin, R. and Hauptmann, A. G. (2001b). Title generation for machine-translated documents. In *Proceedings of the 17th International Joint Conference on Artificial intelligence*, pages 1229–1234, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jing, H. and McKeown, K. R. (2000). Cut and paste based text summarization. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics conference*, pages 178–185, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jones, S. S., Smith, L. B., and Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development*, 62(3):499–516.

Jordan, M. (1999). *Learning in Graphical Models.* MIT Press, Cambridge, MA.

Joshi, D., Wang, J., and Li, J. (2006). The story picturing engine—a system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):68–89.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence).* Prentice Hall, 1 edition. neue Auflage kommt im Frhjahr 2008.

Keller, F., Gunasekharan, S., Mayo, N., and Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, 41(1):1–12.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics*, pages 423–430, Morristown, NJ, USA.

Kneser, R., Peters, J., and Klakow, D. (1997). Language model adaptation using dynamic marginals. In *Proceedings of5th European Conference on Speech Communication and Technology*, volume 4, pages 1971–1974, Rhodes, Greece.

Knight, K. and Hatzivassiloglou, V. (1995). Two-level, many-paths generation. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, pages 252–260, Morristown, NJ, USA. Association for Computational Linguistics.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, 139:91–107.

Kojima, A., Izumi, M., Tamura, T., and Fukunaga, K. (2000). Generating natural language description of human behavior from video images. In *International Conference on Pattern Recognition*, volume 4, page 4728, Los Alamitos, CA, USA. IEEE Computer Society.

Kojima, A., Takaya, M., Aoki, S., Miyamoto, T., and Fukunaga, K. (2008). Recognition and textual description of human activities by mobile robot. In *Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control*, pages 53–56, Washington, DC, USA. IEEE Computer Society.

Kojima, A., Tamura, T., and Fukunaga, K. (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.

Landau, B., Smith, L., and Jones, S. (1998). Object perception and object naming in early development. *Trends in Cognitive Sciences*, 2(1):19–24.

Langkilde, I. and Knight, K. (1998). The practical value of n-grams in generation. In *In International Natural Language Generation Workshop*, pages 248–255.

Lavrenko, V. (2004). *A Generative Theory of Relevance*. PhD thesis, University of California Berkeley.

Lavrenko, V. and Croft, W. (2001). Relevance-based language models. In *Proceedings of the 24th ACM SIGIR Conference on Research and development in Information Retrieval*, pages 120–127, New Orleans, LA. ACM Press.

Lavrenko, V., Feng, S., and Manmatha, R. (2004). Statistical models for automatic video annotation and retrieval. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 1044–1047.

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

Li, J. and Wang, J. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1075–1088.

Li, J. and Wang, J. (2006). Real-time computerized annotation of pictures. In *Proceedings of the 14th annual ACM International Conference on Multimedia*, pages 911–920, New York, NY. ACM.

Li, L.-J., Socher, R., and Fei-Fei, L. (2009). Towards total scene understanding:classification, annotation and segmentation in an automatic framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2036 – 2043, Los Alamitos, CA, USA. IEEE Computer Society.

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

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Mani, I. (2001). *Automatic Summarization*. John Benjamins, Amsterdam.

Marcus, M., Santorini, B., and Marcinkiewicz, M. (1994). Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Maron, O. and Ratan, A. (1998). Multiple-instance learning for natural scene classification. In *Proceedings of the 15th International Conference on Machine Learning*, pages 341–349, San Francisco, CA. Morgan Kaufmann Publishers Inc.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8th International Conference on Computer Vision*, pages 416–423, Vancouver, BC.

Metzler, D., Lavrenko, V., and Croft, W. B. (2004). Formal multiple-bernoulli models for language modeling. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 540–541, New York, NY, USA. ACM.

Mikolajczyk, K. and Schmid, C. (2003). A performance evaluation of local descriptors. In *Proceedings of the 9th International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 257–263, Nice, France.

Miller, G. (1995). Wordnet: a lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41. ACM Press.

Mittal, V. O., Moore, J. D., Carenini, G., and Roth, S. (1998). Describing complex charts in natural language: A caption generation system. *Computational Linguistics*, 24:431–468.

Mittal, V. O., Roth, S., Moore, J. D., Mattis, J., and Carenini, G. (1995). Generating explanatory captions for information graphics. In *IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1276–1283, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Monay, F. and Gatica-Perez, D. (2003). On image auto-annotation with latent space models.

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

Moran, S. (2009). Automatic image tagging. Master's thesis, The University of Edinburgh.

Mori, Y., Takahashi, H., and Oka, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of the 1st International Workshop on Multimedia Intelligent Storage and Retrieval Management*, Orlando, FL.

Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarization. Technical report, Microsoft Research.

Noreen, E. W. (1989). *Computer-intensive Methods for Testing Hypotheses: An Introduction*. John Wiley & Sons Inc.

Pan, J., Yang, H., Duygulu, P., and Faloutsos, C. (2004). Automatic image captioning. In *Proceedings of the 2004 International Conference on Multimedia and Expo*, pages 1987–1990, Taipei.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pedersen, T. and Bruce, R. (1998). Knowledge lean word-sense disambiguation. In *Proceedings of the fifteenth national/tenth conference on Artificial Intelligence/Innovative applications of artificial intelligence*, pages 800–805, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 275–281, New York, NY.

Qi, X. and Han, Y. (2007). Incorporating multiple svms for automatic image annotation. *Pattern Recognition*, 40(2):728–741.

Quinn, P. C., Eimas, P. D., and Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, 22(4):375–463.

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173.

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Schroff, F., Criminisi, A., and Zisserman, A. (2007). Harvesting image databases from the web. In *Proceedings of the 11th International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T. (2005). Discovering object categories in image collections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. (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, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (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*, pages 223–231, Cambridge.

Soricut, R. and Marcu, D. (2006). Stochastic language generation using widl-expressions and its application in machine translation and summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 1105–1112, Morristown, NJ, USA. Association for Computational Linguistics.

Spärck Jones, K. (1999). Automatic summarizing: factors and directions. In Mani, I. and Maybury, M. T., editors, *Advances in automatic text summarization*, chapter 1, pages 1 – 12. The MIT Press.

Spärck Jones, K. (2007). Automatic summarising: The state of the art. *Information Processing and Management: an International Journal*, 43:1449–1481.

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

Tang, J. and Lewis, P. H. (2007). A study of quality issues for image auto-annotation with the Corel data-set. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):384–389.

Ulusoy, I. and Bishop, C. M. (2005). Generative versus discriminative methods for object recognition. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, CVPR '05, pages 258–265, Washington, DC, USA. IEEE Computer Society.

Vailaya, A., Figueiredo, M., Jain, A., and Zhang, H. (1999). Content-based hierarchical classification of vacation images. In *Proceedings of the International Conference on Multimedia Computing and Systems*, pages 518–523, Los Alamitos, CA. IEEE Computer Society.

Vailaya, A., Figueiredo, M., Jain, A., and Zhang, H. (2001). Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10:117–130.

Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. (2007). Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *International Journal on Information Processing and Management*, 43(6):1606–1618.

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

Wang, C., Blei, D., and Fei-Fei, L. (2009). Simultaneous image classification and annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1903–1910, Los Alamitos, CA, USA. IEEE Computer Society.

Wang, J. and Li, J. (2002). Learning-based linguistic indexing of pictures with 2-d MHMMs. In *Proceedings of the 10th ACM International Conference on Multimedia*, pages 436–445. ACM Press.

Wei, X. and Croft, B. W. (2006). LDA-based document models for ad-hoc retrieval. In *Proeedings of the 29th Annual International ACM SIGIR Conference on Research and development in Information Retrieval*, pages 178–185, Seattle, WA.

Westerveld, T. and de Vries, A. P. (2003). Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data. In *In Proceedings of the SIGIR Multimedia Information Retrieval Workshop*, Toronto, ON.

Weston, J., Bengio, S., and Usunier, N. (2010). Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, 81(1):21–35.

Witbrock, M. and Mittal, V. (1999). Ultra-summarization : a statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of the 22nd annual international ACM SIGIR Conference on Research and development in Information Retrieval*, pages 315–316, New York, NY, USA. ACM.

Yao, B., Yang, X., Lin, L., Lee, M. W., and chun Zhu, S. (2009). I2t: Image parsing to text description. *Proceedings of IEEE (invited for the special issue on Internet Vision)*.

Zajic, D., Dorr, B., and Schwartz, R. (2002). Automatic headline generation for newspaper stories. In *Proceedings of the ACL 2002 Workshop on Text Summarization and Document Understanding Conference (DUC 2002)*, pages 78–85, Morristown, NJ, USA. Association for Computational Linguistics.

Zhao, R. and Grosky, W. I. (2003). Video shot detection using color anglogram and latent semantic indexing: From contents to semantics. In Furht, B. and Marques, O., editors, *Handbook of Video Databases: Design and Applications*, pages 371–392. CRC Press.

Zhou, L. and Hovy, E. (2003). Headline summarization at isi. In *Proceedings of the HLT-NAACL 2003 Text Summarization Workshop and Document Understanding Conference (DUC 2003)*, pages 174–178, Morristown, NJ, USA. Association for Computational Linguistics.