A Template-based Model for Automatic Image Description

Clemens Wolff

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2014

Abstract

We propose a novel system to generate simple textual descriptions for images. An example of such a description could be: *Mike is happy to see the orange cat*. We formulate the image description task as a three-stage process involving content planning, content selection and surface realisation. The first phase learns grammatical templates that provide a salient structure to describe image semantics. The second phase learns a classifier to predict words appropriate for a given image and grammatical constraints. The third phase re-ranks the model's output to ensure fluency and adequacy.

One of our model's simplifying assumptions is the use of clip-art pictures instead of photo-realistic images. This enables us to focus on the vision–language interplay and natural language processing aspects of the image description task.

We tune our model by optimising for BLEU and METEOR scores and we use an information theoretic measure to select classifier features (visual and linguistic). We find that strongly regularised Logistic Regression produces the best results. We evaluate our model against two non-trivial baselines that are inspired by related work in the literature. Our approach outperforms both baselines. Human evaluation further confirms the high quality of the descriptions generated by our model.

Acknowledgements

I would like to thank my supervisor, Prof. Mirella Lapata, for her consistent help throughout the course of the project.

Special thanks are furthermore due to Larry Zitnick for providing me with the data that this work is based on and for answering all of my questions about his methodology.

I am also obliged to Dr. Victor Lavrenko for reality-checking my ideas for the keyword baseline in Section 2.3.2.1 as well as to Dr. Iain Murray and Dr. Charles Sutton for helping me understand the theory behind Equation (2.1) and how to implement it.

Table of Contents

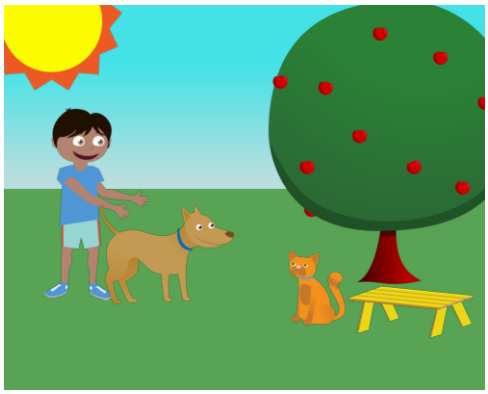
Chapter 1

Introduction

Images are a rich source of information merging the explicit (What objects are in the picture? What attributes do the objects have?) with the implicit (How do the objects in the picture relate? How do they interact?). "A picture is worth a thousand words" — and yet humans can describe images succinctly, even when they are of poor quality (Bachmann, 1991; Oliva and Schyns, 2000) or if they are only visible for an instant (Potter, 1976; Fei-Fei et al., 2007).

A human can *see* what objects are in an image, *recognise* the relationships between them, *understand* their connections and *summarise* the most salient aspects of the image, passing over less important parts. For instance, seeing the image in Figure 1.1, a human prioritised the boy and the interaction he has with the cat but selectively chose to not describe the clothes the boy is wearing, the dog, table, sun, tree, grass or sky: the image is about a boy enjoying meeting a cat, the other aspects are only side concerns.

Describing images *automatically* is a formidable task at the intersection of computer vision and natural language processing, involving many components such as object recognition, image segmentation, attribute extraction, content planning, content selection and surface realisation.



Mike is happy to see the orange cat.

Figure 1.1: Sample image and description.

While challenging (perhaps as involved as the larger image-understanding problem itself) automated image description has a number of powerful and immediate engineering applications. For instance:

- Web-based image search-engines retrieve pictures in response to textual queries. This is achieved by exploiting meta-data such as the image's file-name and surrounding text (noisy), human tagging (expensive) or simple visual features like the image's predominant colour (low-level). Generating automatic image descriptions will enable a semantic component within the retrieval model, thus enabling the user to find images matching a concept rather than a bag of low-level features. This has direct applications in the domain of illustration (Barnard et al., 2001): for instance, a journalist could use a semantically-aware image search engine to find a photo to accompany an article (Feng and Lapata, 2013).

- Automatic image description techniques will furthermore allow users to search collections of images without any annotation or meta-data (such as private photo archives). Some progress has already been made in this area, for instance, a system by Krizhevsky et al. (2012) allows users to retrieve unannotated images by object-occurrence. However, this system can only be thought of as a first step given the evidence that more semantic information (e.g. relationships between objects) is necessary for effective image retrieval (Armitage and Enser, 1997).

- Summarising image contents in text also has applications in the domain of accessibility. Currently, the visually impaired can use screen-readers to access images. However, this relies on content creators to provide manually created semantic annotations for each image. Automatic image description techniques will remove the need for laborious manual labelling and make the vast amount of existing but unannotated data accessible (Ferres et al., 2006).

In addition, studying the vision–language interplay will lead to a deeper understanding of how humans describe images, working towards answering the question of "what is important in an image." Insights thus-gained can be used to guide and focus research in computer vision, giving insight on issues such as which object-detectors might be the most important to build.

1.1 Background

A large amount of work towards building automated image understanding systems has focused on annotating images with keywords (Duygulu et al., 2002; Lavrenko et al., 2003; Feng et al., 2004; Guillaumin et al., 2009; Makadia et al., 2010). In this way, a keyword-generation system might annotate the image in Figure 1.1 with the words: *cat*, *orange* and *table*. This conveys some notion about the objects present in the picture, however the informative details are ambiguous (is the cat orange or the table?) and do not communicate what the picture really is about (how are the cat and table related?). On the other hand, a full-sentence description such as *The orange cat sits next to the table* can explicitly encode the contents of the scene (using words) and implicitly

communicate the relationships between the objects (via the relationships between the words).

Prior work attempting to describe images using higher-level concepts such as phrases, sentences or even multi-sentence free text can roughly be divided along two main paradigms. *Transfer-based* or retrieval-based approaches exploit large collections of parallel images and human-generated descriptions. New images are described by finding closely matching images in a collection and transferring (parts of) their human-generated descriptions onto the unseen image. *Generative* approaches on the other hand, combine the responses of object-detectors with a linguistic model in order to generate completely novel descriptions for unseen images.

The transfer-based approach has the advantage that it leverages the human's skill at describing images: descriptions are always grammatical and fluent, sometimes even poetic and intricately crafted. However, all retrieved descriptions share the property that they were originally written for images distinct from the one that they are now being retrieved to describe. It is therefore unlikely that retrieved descriptions will match a new image as well as novel descriptions custom-made for the pertinent image. The generative approach fixes this flaw, but at the cost of having to deal with natural language generation (an open field with on-going research in and of itself) in addition to image understanding.

Notable representatives of the transfer-based and generative paradigms include the following.

Farhadi et al. (2010) represent images and descriptions in a shared "meaning-space" where semantics are represented by {object, action, scene}-triplets. A Markov Random Field (Li, 1995) is used to map image-features (e.g. object detector responses (Felzenszwalb et al., 2008), scene classification responses (Everingham et al., 2009) and the gist global image descriptor (Oliva and Torralba, 2006)) into the meaning-space. Sentences are mapped into the meaning-space by producing a dependency parse and lifting verbs to actions, subjects to objects and prepositions to scenes (while taking word similarities into account). New images are described by retrieving close sentences in the meaning-space. The system is trained on 1,000 realistic images and 5,000 corresponding descriptions.
The authors note the difficulty of quantitatively evaluating the quality of the generated image descriptions and settle on a human study (which finds the model to perform admirably). An interesting observation of the paper is that the intermediate meaning-space representation allows descriptions to contain information about image-elements which are present and visible but were not picked up by the computer vision systems — image descriptions in natural language can thus be used to guide and correct computer vision detections.
The paper's definition of semantics is rather crude and will lead to image–description mismatches due to the many degrees of freedom that are left open by only constraining three dimensions of the image descriptions. Subtler notions such as spatial relationships for example will only figure in the produced descriptions by chance.

Ordonez et al. (2011) use a simpler approach but a much larger data-set (1 million images and descriptions found on the web). Given an unseen image, related images in

the data-set are found using crude vision features such as the gist global image descriptor and image thumbnails (carrying visual information such as predominant colours). The captions of the related images are then re-ranked by taking image-contents into account (objects, stuff, people and scenes).
The authors evaluate their work using BLEU (Papineni et al., 2002) and report that their model performs better given more data (i.e. performance is improved by using a larger knowledge-base of images and captions with which to compare unseen images and from which to retrieve candidate descriptions). The model also performs better when taking image contents into consideration (i.e. making sure that captions and images are at least somewhat related improves performance).
The model produces good results but requires very large amounts of data to work well. The lack of an underlying semantic model of images and descriptions makes it hard to reason about the efficacy of the system and glean wider applications.

Kuznetsova et al. (2012) build on Ordonez et al. (2011)'s approach and data, addressing some of the criticism levelled towards the transfer-based image description paradigm. Once again, candidate descriptions are lifted from images containing similar visual elements. However, the candidates are constrained to contain a noun phrase, a verb phrase and two prepositional phrases (referring to regions and stuff). The phrases within the descriptions and the words within the phrases are then reordered using integer linear programming (ILP) (Karlof, 2005) in order to satisfy visual constraints (e.g. objects that are prominently described by humans should be mentioned first) and linguistic constraints (e.g. discourse coherence and ngram cohesion should be satisfied). The ILP approach mixes and matches human-generated descriptions, thus producing at least partially novel descriptions for new images. This reduces the image–description mismatch of many transfer-based image description systems.
The model is evaluated (using BLEU and a human study) against Ordonez et al. (2011)'s system and a baseline. The baseline replaces the ILP post-processing with a simpler approach based on Hidden Markov Models (Baum and Petrie, 1966). The authors find that the ILP system outperforms the other models due to its ability to enforce linguistic discourse constraints, even if it introduces more grammatical errors and sometimes produces nonsensical results.
The system is interesting in that it highlights the benefits of using linguistic processing to increase description relevance. Ultimately however, the model still struggles because the text it uses as inputs may or may not be relevant to the image at hand.

Kulkarni et al. (2011) present a system that creates novel descriptions for images by explicitly enumerating all the elements in the picture and their relationships. This produces descriptions that are exhaustive but not overly natural sounding. The approach is based on a Conditional Random Field (CRF) (Lafferty et al., 2001) that is trained to extract meaning-representation pairs from images. The meaning representation consists of pairs of nouns with attributes, linked by a spatial relation (preposition). The meaning-pairs are used for surface realisation via an ngram language model and manually constructed templates. The CRF uses standard vision inputs (object detectors, scene classifiers, attribute detectors, spatial relationship detectors) and simple linguistic inputs mined from text corpora (frequency of attributes, frequency of positional language, frequency of object–attribute pairs).

The system is evaluated using BLEU and human judges. The paper argues that BLEU is a poor metric to evaluate image descriptions, stating that the metric does not correlate very well with human judgement. The paper also finds that using templates helps to generate more realistic descriptions. Due to the simplicity of the natural language processing techniques used, the quality of the generated descriptions is heavily reliant on good vision detections.

Mitchell et al. (2012) treat the vision problem in a similar way as Kulkarni et al. (2011) but use a more flexible description synthesis approach, taking into account co-occurrence statistics and syntactic structure. The language generation system introduced in the paper starts with a set of vision-generated object-, attribute- and relationship-detections and combines them into natural language description by using a probabilistic approach growing syntactic trees bottom-up (i.e. reordering words, combining groups of words into phrases, grouping and de-duplicating similar words, etc.).
Due to the flexible nature of the syntactic structure imposed on the generated descriptions, the proposed system outperforms Kulkarni et al. (2011) in a human study. However, the system does not handle incorrect vision detections and uses a relatively simple maximum likelihood formulation for the language generation process.

Krishnamoorthy et al. (2013) are primarily interested in understanding videos (not images) but their approach shares many facets with image description, whence it is of interest here. Like other works, the paper uses a simple {subject, verb, object} meaning-representation for vision detections and a template-based approach for surface realisation. The novelty of their approach is the use of a strong linguistic component. A language-model prior is used to ground vision detections; Synonym expansion of candidate words is used to increase the space of descriptions that are generated and considered for images.
Evaluation (using BLEU, METEOR (Denkowski and Lavie, 2011) and humans) reveals that the strong linguistic grounding helps the model outperform more vision-oriented approaches.

Beyond transfer-based and generative approaches, full-sentence image description generation is more rarely treated as a summarisation task (using abstractive or extractive methods (Feng and Lapata, 2013) or incorporating domain knowledge (Aker and Gaizauskas, 2010)).

In sum, both transfer-based and generative image description approaches are found to produce reasonable results. It is somewhat unfortunate that (due to the fragmentation of data-sets and evaluation methods) there is no standard way to evaluate and compare systems. This makes in-depth analysis of the existing literature difficult. Some trends do however emerge: more linguistic grounding generally tends to produce more relevant and more grammatical descriptions. This motivates the emphasis on language processing of the work presented here.

1.2 Contributions

The work presented by this report ties in with the generative image description paradigm. Our primary contribution is a novel template-based system that generates sentence-length descriptions for simple images. The novelty of our approach is that we learn template structures and how to fill them in from (potentially noisy) data. Our model outperforms a non-trivial transfer-based baseline and generally produces descriptions that are not only adequate but also highly fluent.

Our image description system is trained on the "Abstract Scenes Data-Set" of 10,020 semantically similar clip-art images and 60,396 corresponding descriptions (Zitnick and Parikh, 2013). The images in the data-set are (trivially) fully labelled — the choice of data thus enables us to focus on the semantic understanding of scenes and on the natural language generation aspects of the image description task, without having to worry about object recognition and image segmentation.

As a secondary contribution, we introduce a methodology (based on prior work by Zitnick and Parikh (2013)) to perform feature selection of visual features and word features.

We also introduce two baseline systems for the image-description task on the "Abstract Scenes Data-Set" and evaluate our template-based model against them.

1.3 Related Work

Our work closely ties in with the existing literature in the automated image description community, taking inspiration and learning lessons most specifically from many authors, including the following.

Kulkarni et al. (2011) argue that imposing a grammatical structure on image descriptions (e.g. requiring that every description has to follow a certain template) produces more readable results than relying on a language model to ensure well-formedness. Mitchell et al. (2012) find that using grammatical structures that can adapt to the image being described (instead of forcing all descriptions to follow the same rigid rule) is an important factor in the generation of relevant descriptions. Therefore, our approach performs content planning using grammatical structures that are learned from data and matched to images on a case-by-case basis.

Farhadi et al. (2009) find that attributes of objects are important for visualness. Gupta and Davis (2008) corroborate this statement by arguing for the importance of adjectives and prepositions in the image description task. Aker and Gaizauskas (2010) take the previous results one step further and state that some grammatical relations in general (not just adjectives and prepositions) provide useful support to generate very readable textual descriptions. Gupta et al. (2012) build on this and report that using richer semantic relations (e.g. subject, object, etc.) helps in creating highly coherent and adequate descriptions. Given these findings, our approach uses the full range of grammatical

structure supported by the Stanford Dependency Parser (De Marneffe and Manning, 2008b) in order to inform content selection and content planning.

Yang et al. (2011) use language models in order to guide text generation and find that this eliminates the creation of nonsensical descriptions. Similarly, Krishnamoorthy et al. (2013) use language models as a post-processing step in order to select the most grammatical and most realistic description from a pool of candidates. As a result, our approach uses a language model during surface realisation to increase the grammaticality of generated image descriptions.

Further relevant literature will be introduced in the main text whenever appropriate.

1.4 Outline

The remainder of this report is structured as follows.

Chapter 2 introduces our main contribution, i.e. the development of a template-based model for automatic image description. The chapter also presents the baselines against which our model is compared and the evaluation metrics used to judge the quality of the models. Additionally, the chapter introduces the data-set used to train the models and our methods for feature selection.

Chapter 3 is a companion to Chapter 2 in that the former provides experimental validation for the choices of the latter. Chapter 3 also reports the results of the feature-selection process previously described as well as the results of tuning the models introduced in the previous chapter.

Chapter 4 evaluates our models using automatic techniques and a human study. The chapter also discusses the respective performances of the different approaches to automatic image description we presented in earlier chapters.

Finally, Chapter 5 synthesises our findings, presents conclusions, suggests how to improve our template-based model and explores avenues for future work based on the broader results of this report.

Chapter 2

Methodology

This chapter presents details on the data, algorithms and evaluation methods employed by this work. The chapter should be read together with Chapter 3 which provides experimental justification and validation for the techniques introduced below.

Section 2.1 introduces the data we use: the "Abstract Scenes Data-Set" (Zitnick and Parikh, 2013) of clip-art images and descriptions. The section gives a brief overview of how the "Abstract Scenes Data-Set" was created and justifies why the data-set is relevant to this report.

Section 2.2 introduces a heuristic for finding the most salient words to describe the images in the domain we are considering. The heuristic will be used to reduce the size of the vocabulary targeted during description generation. The heuristic will also be used as a feature-selection method in order to reduce the dimensionality of our data's image representation.

Section 2.3 proposes and motivates three ways of modelling the image description task (these models are the main novel contribution of this work). The section is broken into two main parts. First, Section 2.3.1 shows a way to reduce the problem of generating textual descriptions for images to a classification task (achieving high description adequacy). Then, Section 2.3.2 compares this model to two baseline approaches where description generation is replaced with simple description retrieval.

Finally, Section 2.4 introduces the methods used (throughout the remainder of the report) to evaluate, compare and contrast the models presented in this chapter.

2.1 The Abstract Scenes Data-Set

This section introduces the "Abstract Scenes Data-Set" created by Zitnick and Parikh (2013). The section summarises the process that was used to create the data-set and explains why the process makes the data-set particularly interesting for this project.

Zitnick and Parikh state that the "Abstract Scenes Data-Set" was created by recruiting workers on the crowd-sourcing platform Amazon Mechanical Turk to create simple

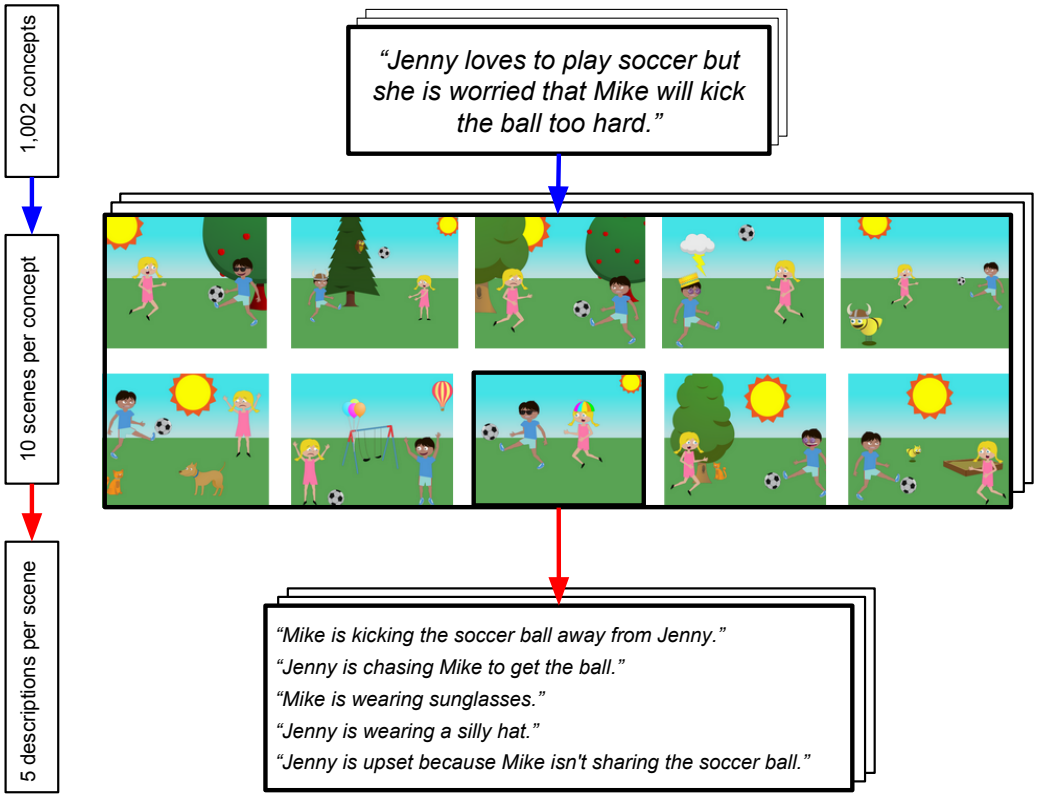


Figure 2.1: Overview of the creation process of the "Abstract Scenes Data-Set."

clip-art scenes based on 1,002 textual concepts such as *Jenny loves to play soccer but she is worried that Mike will kick the ball too hard* or *Lightning is about to strike at a park*. Workers created the scenes using a graphical program displaying a limited number of objects (trees, toys, hats, food and so on) and two agents (children) with a small inventory of poses and expressions. For each of the 1,002 "seed" concepts, 10 scenes were created (10,020 images in total). A separate set of Amazon Mechanical Turk workers was instructed to produce simple textual descriptions of the scenes. Every scene was described by 3 to 9 different workers (60,396 descriptions in total). Figure 2.1 above summarises the data-set creation process. The figure also shows a sample of the images, concepts and descriptions in the data-set.

In order to make the data-set instantaneously usable, Zitnick and Parikh bundle a comprehensive array of visual features extracted from the clip-art scenes with the data-set. These features can be split into two main categories. Discrete visual features measure object occurrence, object co-occurrence, expression of the agents, etc. Continuous visual features cover information such as absolute and relative distances between objects, or positioning of the head and hands of the agents (see Zitnick and Parikh (2013) for full detail). Figure 2.2 shows a sample image from the "Abstract Scenes Data-Set" and some discrete and continuous features extracted from the image.

The "Abstract Scenes Data-Set" is of interest for us for three main reasons. Firstly, the choice of clip-art scenes instead of photo-realistic pictures means that the difficult (and as of yet very much unsolved) problems of object-recognition and image-segmentation

Instance absolute location	
Boy-Bennie:	$9.9e^{-7}$
...	...
Girl-Bennie:	$9.9e^{-1}$
Girl-Fire:	0.0
...	...
Boy-Dog:	0.0

Instance co-occurrence	
Sun-Boy:	1
...	...
Girl-Bennie:	1
HotDog-Fire:	0
...	...
Pail-Shovel:	0

Category occurrence	
Weather:	1
...	...
Hat:	1
Airship:	0
...	...
Food:	0

Agent attributes	
Boy-Expression-Angry:	1
...	...
Girl-Pose-Run:	1
Boy-Pose-Run:	0
...	...
Girl-Expression-Smile:	0

Figure 2.2: Sample image-features included in the "Abstract Scenes Data-Set."

are avoided. It is noteworthy that this gain comes at little to no cost since there is evidence to suggest that semantics do not require photo-realism (Heider and Simmel, 1944; Oatley and Yuill, 1985; Zitnick and Parikh, 2013). Any findings of this project should thus likely translate to "real", photo-realistic images. In the meantime, the use of abstract images enables a focus on the vision-semantics interplay which is relatively novel in the automated image description community. To-date a lot of effort in the field has focused on (or had to focus on) the computer vision aspect of the problem.

Secondly, the methodology underpinning the creation of the "Abstract Scenes Data-Set" is not only thorough but representative. While the data-set covers relatively few concepts, each concept is realised with different images and descriptions that are the result of a back-and-forth process involving humans creating images based on a shared concept and different humans describing these images. This methodology — unlike others in the community (e.g. Rashtchian et al. (2010) or Berg et al. (2010)) — is bound to capture the creativity inherent in the image description task by highlighting different facets of every image's semantics.

Lastly, the "Abstract Scenes Data-Set" with its more than 60,000 descriptions and 10,000 images is relatively large and refined compared to other data-sets used in the literature. Some of the other data-sets used as a basis for work on the image understanding task are significantly smaller than the "Abstract Scenes Data-Set". For example, the "Pascal Sentences" data-set (Rashtchian et al., 2010) contains realistic images and carefully selected captions, however, its size is limited to only 1,000 pictures and 5,000 descriptions. On the other hand, data-sets larger than the "Abstract Scenes Data-Set" often can't guarantee matching descriptions for images (i.e. the descriptions are noisy). For instance, the "Attribute Discovery Data-Set" (Berg et al., 2010) contains 37,795 images and descriptions but the image-captions often contain words or entire phrases that are not related to the images at hand. Another example of a "larger-but-noisier" data-set is the "SBU Captioned Photo Data-Set" (Ordonez et al., 2011) that contains one million images and descriptions but the captions are low-quality, weakly filtered and directly scraped from the web.

This section introduced the "Abstract Scenes Data-Set" of clip-art images and descriptions. The data-set is interesting because it enables our work to focus on the language–vision relations underpinning the image description process, without having to deal with noisy computer vision inputs and detections. The data-set is also noteworthy because of its relatively large size and because it explores a range of concepts using multiple distinct but semantically related images and descriptions.

2.2 Analysing Image Descriptions

One of the main challenges of automated image description is how to chose which words to include in an image description. *"An image is worth a thousand words"* — but exactly which ones? Trying to generate image descriptions using an unrestricted vocabulary (e.g. the entire English language) would be a tremendous challenge for reasons such as data sparsity and engineering considerations (e.g. model building time).

The goal of this section is to find a small subset of words that is sufficient to describe the images in the "Abstract Scenes Data-Set." In order to tackle the task, we detail an information-theoretic heuristic (first proposed in Zitnick and Parikh (2013)) to judge the importance of word features relative to image features.[1] The restricted vocabulary thus-found will be used throughout the remainder of the report, making the description generation task more tractable.[2] Additionally, the heuristic can (by definition) be used in order to identify the visual features that are most predictive of word features. This allows for less predictive visual features to be pruned, therewith further reducing the complexity of the models proposed in future sections.

Equation (2.1) below introduces the metric. First, a word feature w is extracted from image descriptions and a set of visual features V is extracted from the corresponding images. Then, the mutual information $I(v,w)$[3] is computed between w and every visual feature v in V. The sum of all these mutual information scores is the final measure of importance ψ of the word feature w.

$$\psi(w) = \sum_{v \in V} I(v,w) \tag{2.1}$$

Word features can be occurrences of lemmas,[4] occurrences of words in certain parts-of-speech or grammatical positions and so forth — all of which are discrete.[5]

Recall that the "Abstract Scenes Data-Set" represents images using both discrete and continuous visual features. For the discrete subset of these features, Equation 2.2[6] can be used to compute the mutual information between some particular word feature and some particular visual feature.

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.2}$$

For the continuous visual features in the "Abstract Scenes Data-Set", kernel density estimation (Rosenblatt et al., 1956; Parzen et al., 1962) with a discrete kernel[7] is used in order to compute the $I(\cdot)$ term in Equation (2.1).

[1] Information-theoretic metrics have previously been found to be effective for this task e.g. see Lin and Hauptmann (2006) or Yang and Pedersen (1997).

[2] Other work in the literature also restricts the target vocabulary prior to content selection (Yang et al., 2011; Yao et al., 2010; Li et al., 2011; Kulkarni et al., 2011).

[3] Mutual information is a measure of dependence between two random variables X and Y. The quantity can be thought of as the divergence between X and Y's marginal distributions from their joint distribution: $I(X,Y) = D_{KL}(p(x,y)||p(x)p(y))$, where D_{KL} is the Kullback-Leibler divergence (MacKay, 2003, Sections 2.6,9.5).

[4] A lemma is the canonical form of a word. For example, the words *go*, *going*, *goes*, *went* and *gone* all share the lemma *go*.

[5] All word features are extracted using the Stanford CoreNLP tools (De Marneffe et al., 2006; Klein and Manning, 2003; Finkel et al., 2005; Toutanova and Manning, 2000).

[6] As implemented by Brown et al. (2012).

[7] As implemented by Lizier (2013).

This section presented a metric to isolate the words most descriptive of images and the image features most indicative of descriptive words. The findings of analysing the "Abstract Scenes Data-Set" with the metric are described in Section 3.1: we combine the metric presented in this section with frequency analysis and use the k-best words as the target vocabulary of our template-based image description system of Section 2.3.1.

2.3 Modelling Image Descriptions

This section presents three ways of modelling the image description task. Section 2.3.1 below describes the main contribution of this report: a model that uses templates to guide the process of generating novel descriptions for images. Section 2.3.2 proposes two baselines that will be used to bound the performance of the template-based model. The baselines do not generate novel descriptions but rather describe new images by retrieving appropriate descriptions from a corpus.

2.3.1 A Template-based Model

This section describes a model (the "template model") that creates image descriptions in a three step process. Firstly, given an image, a generative model predicts grammatical structures that are typical for the image's description. Then, another generative model uses these structures to construct descriptions for the image. Lastly, the best generated description is selected by a discriminative model.

From a natural language generation perspective, the first phase of our model is the content planing stage, the second phase is content selection and the third phrase constitutes surface realisation.

The method outlined above reduces the image description task to a two-fold classification problem, followed by a re-ranking step. First, the model predicts which structure is appropriate for an image (i.e. which class of structures the image belongs to). Then, the model predicts how to use this structure to generate a description (i.e. which concrete realisation of the abstract structure the image belongs to).

In spirit, the process underlying our template model is not unlike the "overgenerate-and-select" approach proposed by Langkilde and Knight (1998) and adapted for the image description task by Mitchell et al. (2012). The main inspiration for the model, however, is Barzilay and Lee (2003) who mine grammatical structures from text data and use these templates for language generation.

The following sections give detail on the three components of the model.

2.3.1.1 Template Acquisition

This section introduces a data-driven approach to induce templates that describe the underpinning structure of an image description. The templates are acquired by looking

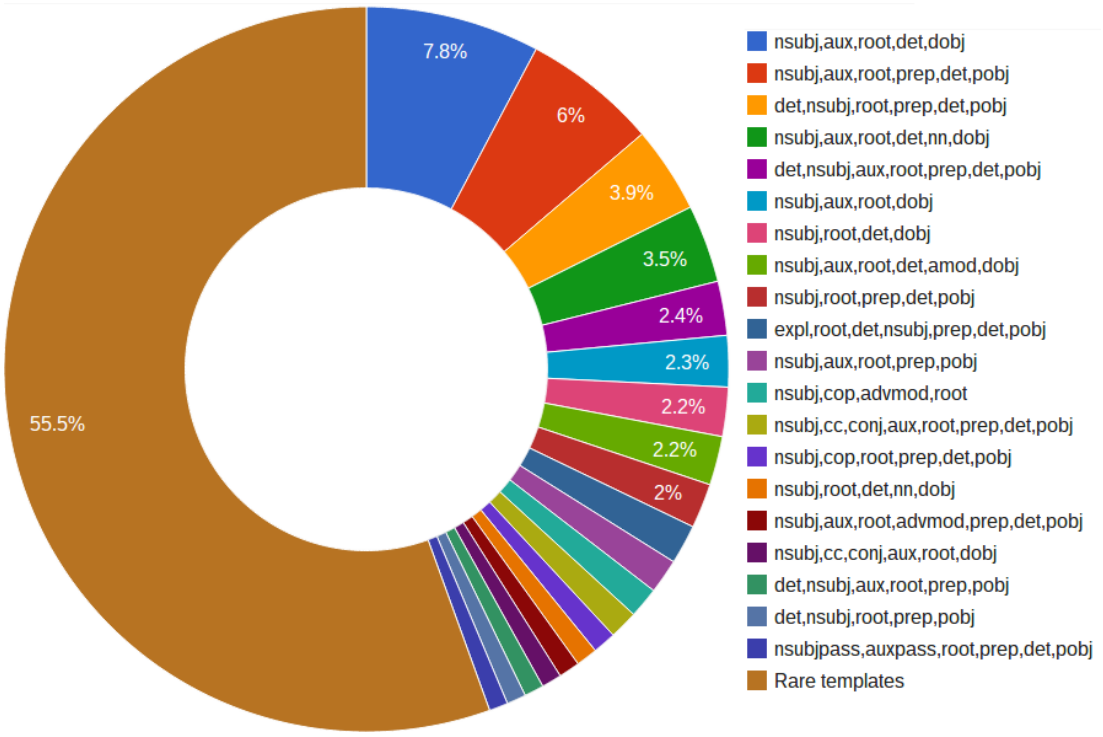


Figure 2.3: Most frequent typed-dependency-based templates.

at the grammatical structure of the human-generated image descriptions in the "Abstract Scenes Data-Set" and finding patterns to emulate therein.

The Stanford "typed dependencies" representation (De Marneffe et al., 2006; De Marneffe and Manning, 2008a) is a standard way to model the grammatical structure of sentences (De Marneffe and Manning, 2008b). The notation assigns a tag to every word in a sentence depending on which grammatical function that word exercises. For example, the phrase *Jenny is wearing a silly hat* will be tagged as *nsubj,aux,root,det,amod,dobj* — "Jenny" is the subject of the sentence, the main action or verb is "wearing", "hat" is the object and so forth. A full description of the typed dependency notation can be found in De Marneffe and Manning (2008a).

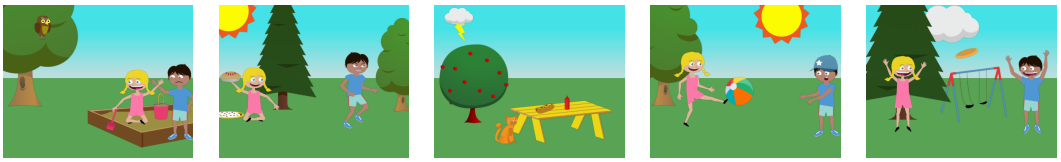
Analysing the descriptions in the "Abstract Scenes Data-Set" in terms of the Stanford typed dependencies shows that most image descriptions follow similar grammatical patterns. Most descriptions are short, in active voice and talk about a single action or fact; sometimes, descriptions are illustrated using a single adjective or adverb; etc.

Reducing each description in the data-set to a sequence of typed dependencies reveals that only a very limited number of grammatical structures — or "templates" — model a large part of the data. Case in point: the 20 templates that occur more than 500 times in the corpus are enough to cover the grammatical structure of more than 44% of all image descriptions in the "Abstract Scenes Data-Set" (Figure 2.3). Figures 2.4–2.9 exemplify images and human generated descriptions matching the six most frequent instances of these templates. Sample sentences for all of the 20 aforementioned templates are given in Appendix A.

Mike is wearing a pirate hat.	Mike is wearing a wool cap.	Jenny is kicking the beach ball.	Mike is throwing the soccer ball.	Jenny is wearing a pirate hat.

nsubj,aux,root,det,nn,dobj

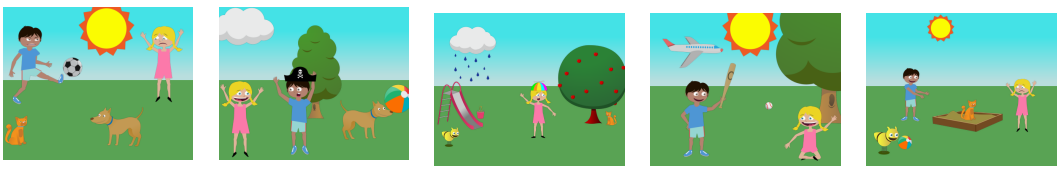
Figure 2.4: Sample images and human generated descriptions realising the *nsubj,aux,root,det,nn,dobj* template.



Jenny is playing in the sandbox.	Jenny is sitting on the grass.	Apples are growing on the tree.	Mike is waiting on the ball.	Jenny is jumping in the air.

nsubj,aux,root,prep,det,pobj

Figure 2.5: Sample images and human generated descriptions realising the *nsubj,aux,root,prep,det,pobj* template.



The dog is looking at the cat.	The dog is playing with the ball.	The cat is sitting under the tree.	The plane is flying in the sky.	The cat is sitting in the sandbox.

det,nsubj,aux,root,prep,det,pobj

Figure 2.6: Sample images and human generated descriptions realising the *det,nsubj,aux,root,prep,det,pobj* template.

Mike is wearing a crown.	Mike is kicking the ball.	Jenny is holding a pie.	Lightning is striking the tree.	Mike is enjoying a hamburger.

nsubj,aux,root,det,dobj

Figure 2.7: Sample images and human generated descriptions realising the *nsubj,aux,root,det,dobj* template.

An owl sits in the tree.	The owl is on the tree.	A thunderstorm moves into the park.	The cloud is behind the tree.	The pizza is on the table.

det,nsubj,root,prep,det,pobj

Figure 2.8: Sample images and human generated descriptions realising the *det,nsubj,root,prep,det,pobj* template.

Mike is wearing sunglasses.	Jenny is wearing pink.	Jenny is holding balloons.	Jenny is wearing sunglasses.	Mike is grilling hamburgers.

nsubj,aux,root,dobj

Figure 2.9: Sample images and human generated descriptions realising the *nsubj,aux,root,dobj* template.

Analysis of the data-set (Section 3.1.1) will show that only a handful of words are commonly used in the grammatical sub-components (or "slots") in these templates. Adverbs, for example, mostly describe spatial relations, nouns function as objects and so forth. This provides evidence that the here-described grammatical-relation based formalism to represent image descriptions is apt at reducing the complexity and variety of the language that humans use to describe images.[8]

Having finalised this template-based formalism for image descriptions, a classifier is trained.[9] This "template-predictor" predicts which template $\hat{t}^{\lambda}$ out of the most common template structures T is adequate for some input image λ (Equation (2.3)).

$$\hat{t}^{\lambda} = \arg\max_{t \in T} p(\lambda, t) \tag{2.3}$$

Details of the types of classifiers used to implement the template-predictor (and performance achieved) are given by Section 3.2.1.

2.3.1.2 Description Generation

In order to transform the grammatical templates of Section 2.3.1.1 into sentences in natural language, a "word-predictor" classifier is trained. Given an image λ, this classifier predicts the most likely word $\hat{w}_s^{\lambda}$ for every grammatical function slot s in a given template $t = s_1, s_2, \cdots, s_n$. In an attempt to deal with the variety of natural language and reduce data sparsity, the candidates for $\hat{w}_s^{\lambda}$ are drawn from a limited set W_s of salient words (see Section 3.1.1 for a discussion of how this set is generated). A formal specification of the word-predictor is given by Equation (2.4).

$$\hat{w}_s^{\lambda} = \arg\max_{w \in W_s} p(w, \lambda \mid s) \tag{2.4}$$

Using this model, a set of candidate descriptions can be generated for any image λ as follows.

- Let $T^{\lambda} = \{t_1, \cdots, t_N\}$ be the set of the N most likely templates for image λ as generated by the template-predictor of Section 2.3.1.1.
- Then, have the word-predictor described in Equation (2.4) predict the set $c_{i,j}^{\lambda}$ of the M most likely candidate words for some grammatical function s_i in some template $t_j \in T^{\lambda}$.
- Now represent $c_{1,j}^{\lambda}, c_{2,j}^{\lambda}, \cdots, c_{M,j}^{\lambda}$ as a directed graph G_j^{λ} where all the elements of $c_{i,j}^{\lambda}$ are connected to all the elements of $c_{i+1,j}^{\lambda}$.

[8] Note that this finding gives empirical evidence supporting Kulkarni et al. (2011)'s assumption that descriptive language only uses a handful of syntactic patterns.

[9] All classifiers described in this report are trained using the machine learning toolkit scikit-learn (Pedregosa et al., 2011).

- Let P_j^{λ} be the ordered set of all words that lie on those paths through G_j^{λ} that start at some element of $c_{1,j}^{\lambda}$ and end at some element of $c_{M,j}^{\lambda}$:

$$P_j^{\lambda} = \left\{ p = n_1 \rightarrow n_2 \rightarrow \ldots \rightarrow n_M, \mid n_1 \in c_{1,j}^{\lambda} \wedge n_M \in c_{M,j}^{\lambda} \wedge p \text{ is a path in } G_j^{\lambda} \right\}.$$

- The set S^{λ} of candidate descriptions for image λ then is the union of all of these paths for all templates:

$$S^{\lambda} = P_1^{\lambda} \cup P_2^{\lambda} \cup \cdots \cup P_N^{\lambda}.$$

Details of the types of classifiers used to realise the word-predictor (alongside with performance achieved) are given by Section 3.2.2.

2.3.1.3 Description Selection

Finally, the best description $\hat{d^{\lambda}}$ for image λ is selected from the candidate descriptions in S^{λ} using the "re-ranker" described by Equation (2.5).[10] The model defines the "goodness" of a description d with m words $d = w_1, w_2, \cdots, w_m$ as a weighted linear combination of the following:

- the probability p_2 of d under a bigram language model,
- the probability p_3 of d under a trigram language model,
- a factor p_l to normalise the language model scores for description length.

$$\hat{d^{\lambda}} = \arg\max_{d \in S^{\lambda}} \prod_{i \in \{2,3,l\}} c_i p_i(d) \tag{2.5}$$

The bigram and trigram language models are simple statistical models that disambiguate well formed descriptions from poorly formed ones by assigning high probabilities the the former and low probabilities to the latter. The two models can be explained by Equation (2.6), substituting n for 2 and 3 respectively.[11]

$$p_n(d) = p_n(w_1, w_2, \cdots, w_m) = \prod_{i=i}^{m} p(w_i | w_{i-(n-1)}^{i-1}) \tag{2.6}$$

Note that as the length of a description d increases, the probability $p_n(d)$ will decrease under any n-gram language model because of the product term in Equation (2.6). A corollary of this observation is that short descriptions score highly under any language model, regardless of quality, just by virtue of being short. This is why the length-normalising factor p_l is introduced in Equation (2.5): to penalise overly short descriptions. For simplicity's sake, p_l is taken to be the maximum likelihood estimate

[10] Note that for reasons of numeric stability, a log-transformed version of Equation (2.5) is used in the actual implementation.

[11] The language models are trained using the SRILM language modelling toolkit (Stolcke et al., 2002).

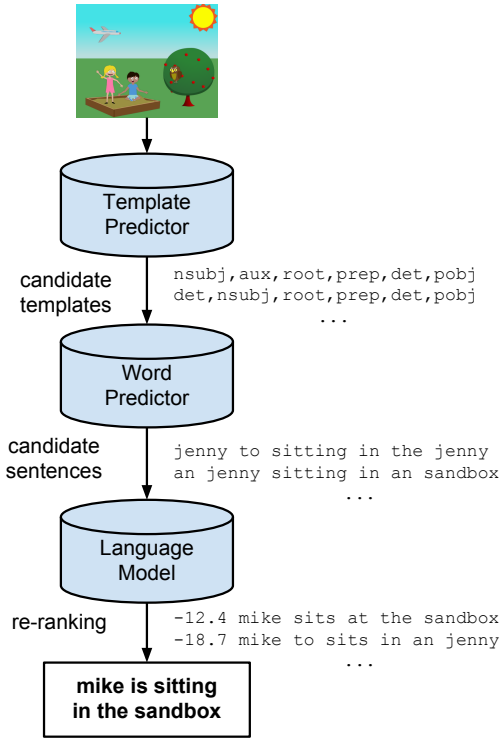


Figure 2.10: Overview of the template-based image description model.

of the description length over all sentences d' in the entire corpus of image descriptions K:

$$p_l(d) = \frac{\|\{d' \in K \mid \text{Len}(d') = \text{Len}(d)\}\|}{\|K\|}. \tag{2.7}$$

Here $\|\cdot\|$ is the norm and $\text{Len}(\cdot)$ is the description length in words.

Details on how the bigram, trigram and length language models influence the performance of the re-ranker are given in Section 3.2.2.

This section presented a novel three-phase system that uses templates learned from data to guide the generation of image descriptions. Figure 2.10 illustrates and summarises the template-based model.

2.3.2 Search-based Baseline Models

This section suggests two baseline models for the image description task. These simple models will be used in Chapter 4 to provide a lower bound on the performance of the more involved template model presented earlier in this chapter. Unlike the template model, the baseline models do not generate novel image descriptions. Instead, they simply retrieve the most appropriate human generated descriptions seen at model training time. The following sections introduce the baselines in detail.

2.3.2.1 Keyword baseline

The first baseline model is keyword based. The model finds a description $\hat{d}^{\lambda}$ for an image λ using a two step process. First, a word-predictor (as described in Section 2.3.1.2) is used to predict the bag-of-words B^{λ} that contains the root word, object, subject, adverb, noun, adjective and so forth that most saliently describe λ.[12] In essence, we use the word-predictor as a simple image-tagging system.

The image-keywords in B^{λ} are then used as a term-frequency-inverse-document-frequency (TFIDF)[13] search query[14] against the set H of human generated image descriptions seen during training of the word-predictor. Equation (2.8) details the TFIDF similarity metric used for the search:

$$\begin{aligned} \text{TFIDF}(q,d) &= \sum_{w \in q} \text{TF}(w,d)\,\text{IDF}(w), \\ \text{TF}(w,q) &= \sqrt{\sum_{w' \in q} \mathbb{1}_{w=w'}}, \\ \text{IDF}(w) &= 1 + \log \frac{\|H\|}{1 + \sum_{d \in H} \sum_{w' \in d} \mathbb{1}_{w=w'}}. \end{aligned} \tag{2.8}$$

Here H is the set of all human generated image descriptions seen at model training time, $\|\cdot\|$ is the set-norm, q is a search query, d is any description in H and $\mathbb{1}_{w=w'}$ is an indicator variable that is 1 if w and w' are the same word and 0 otherwise.

As such, the description $\hat{d}^{\lambda}$ returned by the keyword baseline model is given by Equation (2.9). It is simply the human generated description (i.e. the element of H) that maximises the TFIDF similarity with the keywords B^{λ} generated for the image.

$$\hat{d}^{\lambda} = \underset{d \in H}{\arg\max}\ \text{TFIDF}(d, B^{\lambda}). \tag{2.9}$$

Note that the keyword baseline model is conceptually similar to the model proposed by Farhadi et al. (2010). In the paper, images and descriptions seen at model training time are mapped into a shared meaning space M using a function f. Given an unseen image λ, the description closest to $f(\lambda)$ in M is retrieved and returned by the model. In essence, the keyword baseline model proposed in this section can be seen as an instantiation of Farhadi et al.'s shared-space model (using keywords as the meaning space and TFIDF as a distance measure in that space).

Figure 2.11 illustrates and summarise the keyword baseline model.

[12] A more involved baseline could use the k-best such words and weigh the impact of the words on the retrieval-process according to their k-rank. For simplicity's sake, we only consider the case $k = 1$ here.

[13] TFIDF is a relevance or similarity measure for documents (Wu et al., 2008).

[14] All searches are realised using the Apache Lucene search engine (Cutting, 2014).

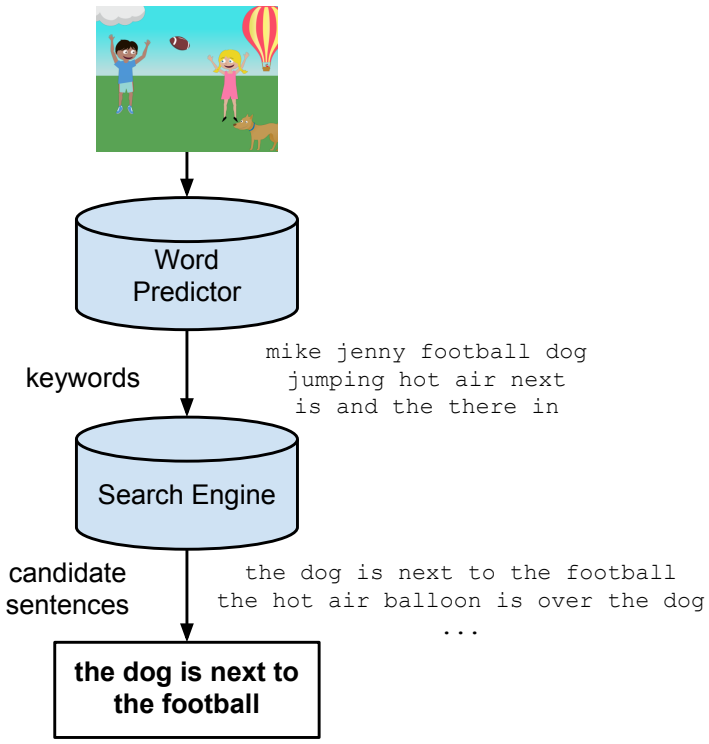


Figure 2.11: Overview of the keyword-based information retrieval baseline.

2.3.2.2 Image-Similarity Baseline

Where the baseline model of Section 2.3.2.1 was based on description similarity, this section's baseline model exploits image similarity.

The model finds a description $\hat{d}^{\lambda}$ for an image λ using a two step process. First, I, the set of images on which the baseline model was trained, is searched in order to find the image λ' that is most similar to λ. Then, the model selects one of the human generated descriptions of λ' at random: this is the predicted description $\hat{d}^{\lambda}$.

In order to implement the image-similarity model in a tractable way, locality sensitive hashing is used to reduce the search space when looking for λ'. Instead of finding λ' by comparing λ to all images in I, λ is only compared to all images that map into the same region of space as λ under some hashing function LSH. Here, LSH is the function that partitions the space of images along the number of their active visual features. Equation (2.10) introduces the locality sensitive hashing function formally:

$$\mathrm{LSH}_I(\lambda) = \left\{ \lambda' \in I \mid \sum_{i=1}^{n} \lambda'_{\mathrm{VF}^i} = \sum_{i=1}^{n} \lambda_{\mathrm{VF}^i} \right\}. \tag{2.10}$$

For this equation, λ_{VF^i} is the i^{th} visual feature of image λ and n is the number of visual features of the image.

Ad-hoc analysis shows that the locality sensitive hashing function in Equation (2.10)

reduces the search space for every image by a factor of at least 10.

After using Equation (2.10) to find a subset I^{λ}_{LSH} of candidate images that are similar to image λ, the Cosine similarity metric[15] (Equation (2.11)) is used to search I^{λ}_{LSH} to find λ'.

$$\text{Sim}(\lambda,\lambda') = \frac{\sum_{i=1}^{n} \lambda_{\text{VF}^i}\lambda'_{\text{VF}^i}}{\sqrt{\sum_{i=1}^{n}(\lambda_{\text{VF}^i})^2}\sqrt{\sum_{i=1}^{n}(\lambda'_{\text{VF}^i})^2}}. \tag{2.11}$$

Performing ad-hoc analysis justifies the choice of using Cosine similarity as the distance metric in I^{λ}_{LSH}. Cosine similarity gives slightly better results than other metrics such as, for instance, Hamming distance[16] (the gain in BLEU score is about 4%).

Equation (2.12) combines Equations (2.10) and (2.11) to give a summary overview of the image-similarity baseline:

$$\hat{d}^{\lambda} = d^{\lambda'}_{\text{Rnd}}, \text{ such that } \lambda' = \underset{\lambda' \in \text{LSH}_I(\lambda)}{\arg\max}\ \text{Sim}(\lambda,\lambda'). \tag{2.12}$$

In this case, $d^{\lambda'}_{\text{Rnd}}$ is a one of the human generated descriptions of image λ' selected at random, $\text{LSH}_I(\lambda)$ is a function that returns those images in I that have the same locality sensitive hash as the image λ (Equation (2.10)) and Sim is a measure of the similarity of two images (Equation (2.11)).

Note that the generation of $\hat{d}^{\lambda}$ involves a random component. This means that the image-similarity baseline may produce different outputs given the same inputs on subsequent runs. In order to counteract the influence of this non-determinism, the models' scores will always be reported as averages over ten runs.

Figure 2.12 illustrates and summarise the image-similarity baseline model.

2.4 Evaluation Techniques

This section discusses the methods that will be used in Chapters 3 and 4 to evaluate the description-generation models presented in the previous section.

The literature in the domain of automated image description distinguishes between two main forms of evaluation techniques: automated and involving humans. Section 2.4.1

[15] The cosine measure is a standard way to compute the similarity between two vectors that is frequently used in information retrieval (Croft et al., 2010, Section 7.1.2), natural language processing (Mihalcea et al., 2006) and computer vision (Nguyen and Bai, 2011). Cosine similarity defines two vectors to be similar if the angle between them is small.

[16] The Hamming distance between two N-dimensional vectors A and B is the number of elements that the vectors have in common: $\text{Hamming}(A,B) = \sum_{i=1}^{N} \mathbb{1}_{A_i=B_i}$, where $\mathbb{1}_{A_i=B_i}$ is an indicator variable that takes on the value 1 if A_i and B_i are the same and 0 otherwise.

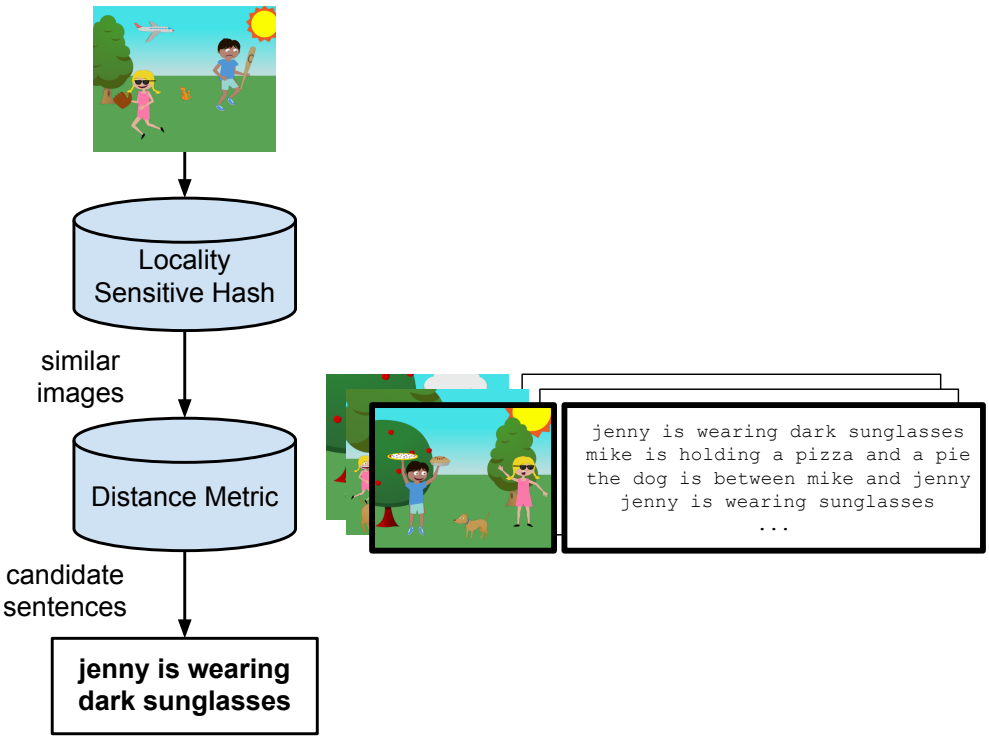


Figure 2.12: Overview of the image-similarity-based information retrieval baseline.

gives a brief overview of automated image description evaluation techniques and proposes two metrics that will be used to validate implementation choices during the model selection process described in Chapter 3. The automated metrics will also be used as a part of the overall model evaluation process (Chapter 4). The best models with regards to the automated metrics will additionally be subjected to human evaluation using the technique outlined in Section 2.4.2.

2.4.1 Automatic Evaluation

Automated image-description evaluation-techniques are (generally) supervised, (generally) end-to-end, computationally efficient metrics that aim to approximate human judgement. Unlike human evaluation (which is slow and expensive to aggregate), automatic evaluation can be used during the development phase of an image description model (e.g. for model selection or parameter optimisation).

Popular automatic evaluation metric include keyword precision and recall (Gupta and Davis, 2008; Berg et al., 2010; Yang et al., 2011; Feng and Lapata, 2013) or keyword accuracy (Gupta et al., 2008a; Farhadi et al., 2009; Krishnamoorthy et al., 2013). Keyword-based metrics give a high-level overview of the quality of a description-generation model. The metrics will capture whether generated descriptions roughly talk about the same topics as the gold-standard descriptions. Keyword-based metrics are thus useful to evaluate the performance of object recognition or image-segmentation systems. However, grammaticality and well-formedness are not taken into account.

Therefore, keyword-based evaluation techniques are not very interesting for us.

Research more interested in the natural language generation aspect of image description frequently borrows automatic evaluation metrics from statistical machine translation. The image description problem is re-formulated as a translation task from image-content to natural language and can then be evaluated using standard metrics from the machine translation literature. The BLEU[17] score first introduced by Papineni et al. (2002) is one of these oft-borrowed metrics (Farhadi et al., 2010; Ordonez et al., 2011; Kulkarni et al., 2011; Kuznetsova et al., 2012; Elliott and Keller, 2013). The METEOR[18] score first introduced by Denkowski and Lavie (2011) is another metric from the machine translation community that is sometimes used to evaluate image description systems (Krishnamoorthy et al., 2013). More detail on both metrics is given in Sections 2.4.1.2–2.4.1.3.

Unlike the other models described in Section 2.3, the template acquisition sub-component (Section 2.3.1.1) of the template model (Section 2.3.1) can easily be evaluated in isolation. It therefore deserves special treatment during evaluation. Whence Section 2.4.1.1 introduces the Levenshtein distance which will be used to evaluate the quality of the templates generated by the sub-component in isolation.

2.4.1.1 Levenshtein Distance

The Levenshtein distance first introduced by Levenshtein (1966) is a measure of sequence similarity. Informally, the metric defines the similarity of two sequences as the number of insertions, deletions and substitutions necessary to transform one sequence into the other. Equation 2.13 defines the metric formally.

$$\mathrm{Lev}(a,b) = \mathrm{Lev}(a,b,\|a\|,\|b\|) \tag{2.13}$$

$$\mathrm{Lev}(a,b,i,j) = \begin{cases} \max(i,j) & \text{if } i=0 \text{ or } j=0 \\ \min \begin{cases} \mathrm{Lev}(a,b,i-1,j)+1 \\ \mathrm{Lev}(a,b,i,j-1)+1 \\ \mathrm{Lev}(a,b,i-1,j-1)+\mathbb{1}_{a_i \neq b_i} \end{cases} & \text{otherwise.} \end{cases} \tag{2.14}$$

Here a and b are two sequences, $\|\cdot\|$ is the sequence length and $\mathbb{1}_{a_i \neq b_i}$ is an indicator variable that is 1 when the i^{th} element in a and the i^{th} element in b are different and 0 otherwise.

In order to evaluate the quality of the templates generated by the template acquisition model of Section 2.3.1.1, Equation (2.13)[19] is used to quantify how different the predicted templates are from the gold-standard (i.e. the templates extracted from the human generated descriptions). The advantage of evaluating with the Levenshtein

[17] BLEU stands for "Bilingual Evaluation Understudy."
[18] METEOR stands for "Metric for Evaluation of Translation with Explicit Ordering."
[19] As implemented by Bird et al. (2009).

distance over other metrics is that it takes into account order (unlike precision and recall) and allows for slight variations on the gold-standard (unlike accuracy). The latter is a very nice property to have since many of the templates can be used interchangeably with little to no loss in semantics. For instance, adding a noun to transform *nsubj,aux,root,det,dobj* into *nsubj,aux,root,det,nn,dobj*[20] or pre-pending a determiner to *nsubj,aux,root,prep,det,pobj*[21] will only marginally change the semantics implied by the template.

2.4.1.2 BLEU

The BLEU score introduced by Papineni et al. (2002) is a metric that measures the ngram precision between a set of machine generated "hypothesis" sentences and set of gold-standard human generated "reference" sentences. In order to avoid giving good scores to sentences that repeat "easy" words that are likely to appear in the references (such as "the"), BLEU modifies the definition of ngram precision. The counts for any ngram in the machine-generated sentences are truncated to the maximum number of occurrences of that ngram in any single reference sentence. Often this ngram precision-score combines with a brevity penalty and is interpolated over various values of n ($n = 1,2,3,4$ and interpolation smoothed via exponential decay are common). The BLEU metric can be summarised by Equation (2.15):

$$\text{BLEU}_N = \xi \cdot \exp(\sum_{n=1}^{N} \omega_n \log P_n), \text{ subject to } \sum_{n=1}^{N} \omega_n = 1, \tag{2.15}$$

where ω_n are positive weights tuned to maximise the correlation of BLEU with human judgement, ξ is the brevity penalty and P_n is the modified ngram precision.

Using BLEU as a metric to evaluate the quality of image descriptions is not without problems. BLEU was originally designed to evaluate the performance of machine translation systems — a domain where the agreement within the gold-standard is usually quite high: there are only so-and-so many ways in which humans translate a given sentence. The variability in image descriptions, however, is much higher — humans describe the same image in very different ways, choosing to describe different components of the image or simply interpret the same elements differently (Kulkarni et al., 2011; Gupta et al., 2012). The immediate effect of this is that the intra-corpus human-human agreement as measured by BLEU is quite low. In other words, the human-human agreement does not necessarily represent an upper-bound on the performance that can be achieved by an image description system. Table 4.1 in Chapter 4 (page 50) gives evidence that this observation holds for the "Abstract Scenes Data-Set." The unfortunate side-effect of this fact is that the BLEU score is less interpretable and may not correlate well with human judgement.

Despite these shortcomings, BLEU is a standard way to evaluate the quality of image descriptions, used widely throughout the literature. In order to maintain compatibility

[20] Compare: *Mike is kicking a soccer ball* and *Mike is kicking the ball*.
[21] Compare: *The cat is sitting in the tree* versus *Jenny is sitting on the ground*.

and comparability with other work, we will therefore present BLEU scores ($N = 4$) smoothed with exponential decay[22] wherever appropriate.

2.4.1.3 METEOR

Section 2.4.1.2 raised the issue that BLEU might not be the most appropriate metric to evaluate systems producing textual image descriptions because of its limited ability to capture the diversity of the language humans use to describe images. The METEOR metric (Denkowski and Lavie, 2011) aims to address this shortcoming by matching hypothesis and reference sentences in a fuzzy way. In addition to exact agreement, stemmed-, synonym- and paraphrase-matches are also given partial credit when computing the ngram overlap between hypotheses and references. Furthermore, METEOR also takes fuzzy recall into account (whereas BLEU is only concerned with precision). The METEOR metric can be summarised by Equation (2.16):

$$\text{METEOR} = (1 - \xi) \cdot F_{\text{mean}}, \tag{2.16}$$

where ξ is a brevity penalty and F_{mean} is the harmonic mean between fuzzy precision and recall.

METEOR focuses on computing the semantic similarity between two sentences rather than getting caught up in the details of how these semantics are expressed. The metric's fuzzy matching approach is likely to counter-act some of the variability and subjectivity inherent in the image description task. METEOR should thus correlate more strongly with the human judgement than BLEU. Considering this strength, we will include METEOR scores[23] in addition to BLEU scores wherever appropriate.

2.4.2 Human Evaluation

Image description systems are commonly evaluated using humans to measure the quality of the produced descriptions. Performing human evaluation is more arduous than automated evaluation but produces more directly interpretable and relatable results. A common human evaluation technique is to ask a panel of judges to quantify measures such as description relevance (adequacy) or readability (fluency) on a Likert scale (Barzilay and Lee, 2003; Yang et al., 2011; Kulkarni et al., 2011; Mitchell et al., 2012).[24] Alternatively, humans are sometimes asked to assign a preference order to generated and gold-standard descriptions (Berg et al., 2010; Ordonez et al., 2011).

[22] As implemented by Clark et al. (2011).

[23] See Footnote 22.

[24] A Likert scale (Likert, 1932) is a tool used to collect human answers to questionnaires. Subjects are presented with statements and asked to quantify their level of agreement or disagreement on the scale: *agree strongly*, *agree mostly*, *disagree mostly* and so forth.

This project uses the Likert scale approach. Human judges were recruited via Amazon Mechanical Turk.[25] Every judge was shown 20 sets of an image and three descriptions:

- A description selected at random from the image's gold-standard descriptions
- A description generated by the keyword baseline as per Equation (2.9)
- A description generated by the template model as per Equation (2.5)

The order of the descriptions was randomised for every image in order to avoid biases. The keyword baseline was chosen over the image-similarity baseline of Section 2.3.2.2 because it achieved better results during automatic evaluation (see Table 4.1 in Chapter 4, page 50).

The judges were asked to rate the adequacy of the three descriptions for every image on a five-point Likert scale. Scores were collected for 200 images (10 sets of 20 images, every set was shown to 10 distinct human judges). Every judge was paid \$0.20 to complete one set of judgements. The scores thus gathered can be used to create a ranking of the models or to perform statistical analysis on the adequacy of each model individually. Figure 2.13 shows a screen-shot of the evaluation setup on Amazon Mechanical Turk. The full instructions given to the human evaluators can be found in Appendix B.

The human evaluators were *not* asked to judge the grammaticality or fluency of the descriptions. This is because almost all of the image descriptions used in the human evaluation experiment are well formed. The gold-standard and baseline descriptions are grammatical by construction (they were all written by humans) and the template model produces consistently highly grammatical sentences.[26]

This section introduced two automatic methods to evaluate the quality of image description systems: BLEU and METEOR. In order to complement these automatic evaluation techniques, the section proposed an experiment to gather the human perspective on the quality of the image descriptions generated by our models.

[25] Amazon Mechanical Turk is a crowd-sourcing platform that allows human workers to complete small tasks for a small monetary reward. The platform has been used extensively to evaluate image description tasks (Ordonez et al., 2011; Yang et al., 2011; Mitchell et al., 2012; Elliott and Keller, 2013; Krishnamoorthy et al., 2013).

[26] Like other authors (Yang et al., 2011; Kulkarni et al., 2011) this work finds that the combination of a template-based description generation process and language model produced very readable descriptions.

	Description	Relevance
1.	Mike is running away from the bear.	**Low** ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 **High**
	Mike is wearing a hat.	**Low** ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 **High**
	Mike cooked hot dogs.	**Low** ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 **High**

Figure 2.13: Screen-shot of the evaluation setup on Amazon Mechanical Turk. Every human evaluator is asked to rate twenty of the here depicted image-description pairs.

Chapter 3

Implementation

This chapter details how the methodology outlined in Chapter 2 is implemented.

Section 3.1 discusses the techniques used to reduce the dimensionality of the visual and linguistic features in the "Abstract Scenes data-set" and presents the condensed feature sets.

Section 3.2 gives detail on the classifiers and parameters used to implement the template model of Section 2.3.1.

3.1 Feature Selection

The "curse of dimensionality" (data sparsity resulting from overly high dimensional data representation) is a real problem for the "Abstract Scenes data-set." The image representation proposed by the data-set has 7,000+ dimensions. The data-set's descriptions contain over 2,700 different words with more than 2,000 different lemmas — arguably overwhelming dimensionality for the data-set's mere 10,000 images and 60,000 descriptions. This section posits a solution to this "curse of dimensionality" for the visual and word representations in the data-set.

Section 3.1.1 analyses the "Abstract Scenes Data-Set" using the word importance heuristic defined in Equation (2.1) to identify the most descriptive words in the data-set.

Section 3.1.2 uses the same heuristic in order to identify the most important visual features for the image-description-generation task.

3.1.1 Selecting Word Features

This section uses the word-importance heuristic defined in Section 2.2 in order to find the words that most saliently describe images in the "Abstract Scenes Data-Set."

Note that the metric utilised in this section was first introduced by Zitnick and Parikh (2013) in order to analyse the 1,002 concept image descriptions of the "Abstract Scenes

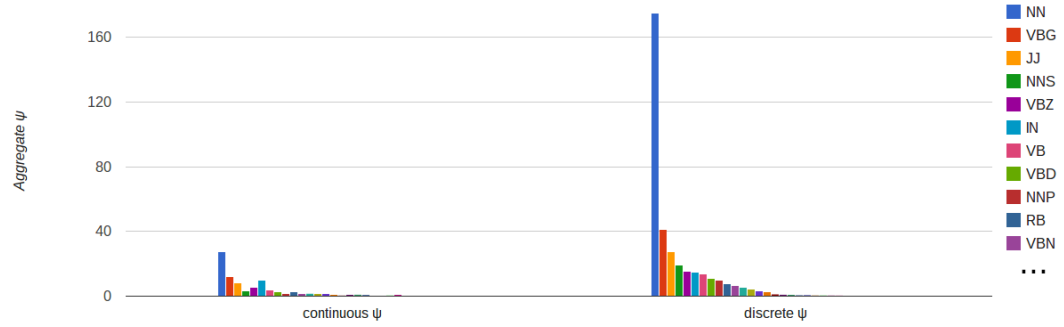


Figure 3.1: Most important parts of speech according to Equation (2.1)

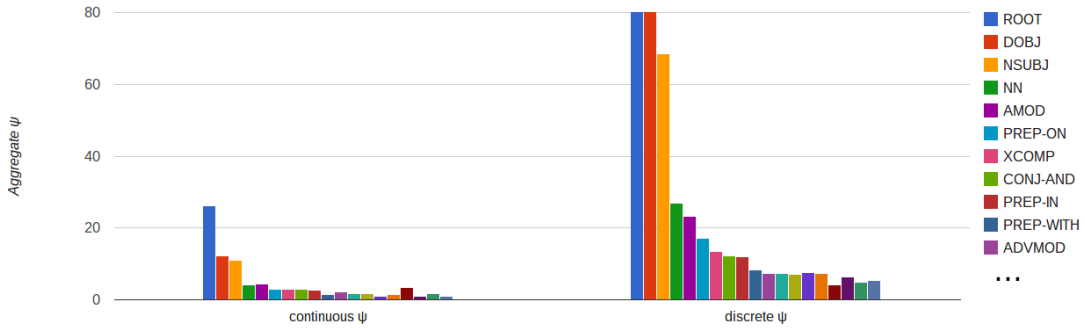


Figure 3.2: Most important grammatical functions (typed dependencies) according to Equation (2.1)

Data-Set." This section reports the findings of analysing the full 60,396 image descriptions with the metric, thus generalising the results of in Zitnick and Parikh (2013).

Table 3.1 gives the 10 most important words, the 10 least important words and 10 words of middling importance according to Equation (2.1). The table largely agrees with the similar analysis of (Zitnick and Parikh, 2013, Fig. 4,6).

Figures 3.1–3.2 use Equation (2.1) to find the most important grammatical relations (dependency types) and parts of speech in image descriptions. Both figures agree that active word types (nouns and verbs or subjects and objects) are most likely to be informative parts of image descriptions.

Combining the knowledge of diverging word and word-function importance (Table 3.1 and Figure 3.2 respectively), a quick analysis of how often some lemma occurs in some grammatical function further narrows the space of important words:

- Adjectival modifiers can be grouped into two dominant categories: colours (e.g. *blue*, *yellow*, *purple*, *colourful*, etc.) and weather (e.g. *hot*, *sunny*). These two categories respectively make up 31% and 24% of all adjectives used. Consistent with Levinson (1983, p. 101) we find that any other aspects of scenes that merit adjectival quantification have to be extremely distinctive (11% — e.g. *silly*, *funny*, *viking*, *pirate*).

Lemma	$\Psi_{discrete}$	$\Psi_{continuous}$
ball	6.0895	0.823699
dog	5.5865	0.745349
wearing	5.3177	0.947201
table	5.6530	0.517763
hat	5.3550	0.814296
cat	5.2840	0.578014
bear	4.8296	0.500458
soccer	4.2520	0.446080
tree	3.6440	0.781586
sandbox	3.9969	0.415286
...	...	...
eats	0.17733	0.026883
love	0.1353	0.065019
alone	0.16442	0.03435
built	0.17503	0.023432
got	0.13052	0.067804
spring	0.17254	0.025169
telling	0.15837	0.039141
burned	0.17201	0.024303
friend	0.13452	0.060785
joy	0.15839	0.035962
...	...	...
eyes	0.074665	0.011388
clear	0.074932	0.010772
shorts	0.073215	0.012376
pointing	0.071192	0.012376
annoyed	0.070486	0.011388
disk	0.07351	0.007274
well	0.0693	0.011388
things	0.068755	0.011388
yard	0.066302	0.010386
raise	0.061954	0.011388

Table 3.1: Most, middling and least important words according to Equation (2.1)

- Direct objects represent 20% clothing (e.g. *hat*, *sunglasses*, *cap*) and 22% action-items (e.g. *ball*, *football*, *frisbee*, etc.).
- Passive objects and subjects are mostly agents (e.g. *Mike*, *Jenny*, *bear*, *dog*, *snake*; 25% of cases) or the environment (e.g. *tree*, *sky*, *ground*, *table*, *sandbox*, etc.; 25% of cases).
- Adverbial modifiers involve predominantly *next* (24%), *very* (21%) and *away* (13%).

Note that these commonly occurring words also obtain scores in the top 5–10% when evaluated with the ψ metric of Equation (2.1).

Similarly, function words[1] also have a rather limited vocabulary in the "Abstract Scenes Data-Set":

- Lemmas *be* and *to* cover 84% and 12% of all auxiliaries used. Additionally, the lemma *be* also accounts for 97% of all passive auxiliaries and 99% of all copulas.
- Nearly all coordinations (99%) are the token *and.*
- 65% of all determiners are the token *the.* The lemma *a* covers another 35% of all determiners.

The combination of word-frequency (in some given grammatical function) and ψ score can now be used as a tool to extract the most important words to describe images. Table 3.2 shows the most common words for every non-root[2] grammatical function used in the model's templates. Figure 3.3 verifies that these words not only represent the most common but also the most descriptive words in the "Abstract Scenes Data-Set." The figure computes the average ψ score of every candidate word in Table 3.2 and contrasts the value with the score for the non-candidate words. Figure 3.3 shows that the small set of lemmas in Table 3.2 indeed contains the bulk of the descriptive power of the language used to describe images in the "Abstract Scenes Data-Set" (as measured by the ψ metric).

This section analysed the descriptions in the "Abstract Scenes Data-Set" using a metric of word importance based on mutual information. The results indicate that the small subset of less than 70 words listed in Table 3.2 (plus some root words) is enough to cover most aspects of how humans describe images. Thus, these words will be used hereafter as the target vocabulary when generating image descriptions.

3.1.2 Selecting Visual Features

The previous section used the ψ word-importance heuristic defined in Formula (2.1) to find the most important words (or word features) in the "Abstract Scenes Data-Set." This section employs the same heuristic to find the most important visual features in the data-set.

[1] Function words are words with no semantic content that act as grammatical glue in a phrase.
[2] The class of root words is too diverse to be restricted in any meaningful way.

Typed Dependency	Candidate Lemmas
advmod	next, very, away
amod	blue, purple, pink, yellow, colourful, green, red, hot, sunny, viking, silly, funny, pirate
aux	be, to
auxpass	be
cc	and
conj	jenny, mike, dog, cat, bear, duck, mustard, ketchup, pie
cop	be
det	the, a
dobj	hat, sunglass, cap, ball, football, frisbee, balloon, kite, baseball, mike, jenny, dog, hamburger, pie, pizza
expl	there
nn	soccer, baseball, tennis, ball, apple, swing, rocket, air, sand, sun, beach, picnic, pirate, witch, chef, mike, jenny
nsubj	mike, jenny, dog, bear, cat, owl, snake, duck, balloon, helicopter, sun, it
nsubjpass	jenny, she, mike, he, bear, owl, cat, dog
pobj	mike, jenny, table, sandbox, slide, bear, dog, snake, tree, sky, ground, park, grass
prep	in, on, to, at, of, with, near, by

Table 3.2: Most informative lemmas per typed dependency position.

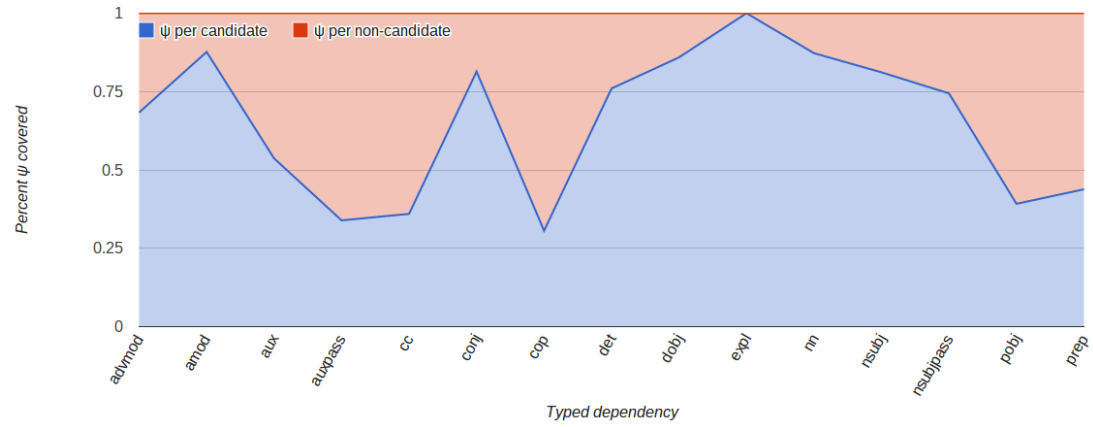


Figure 3.3: Comparison of average per-token ψ score of candidate words versus non-candidate words. Candidate words have one of the lemmas listed in Table 3.2. Non-candidate words are all other words.

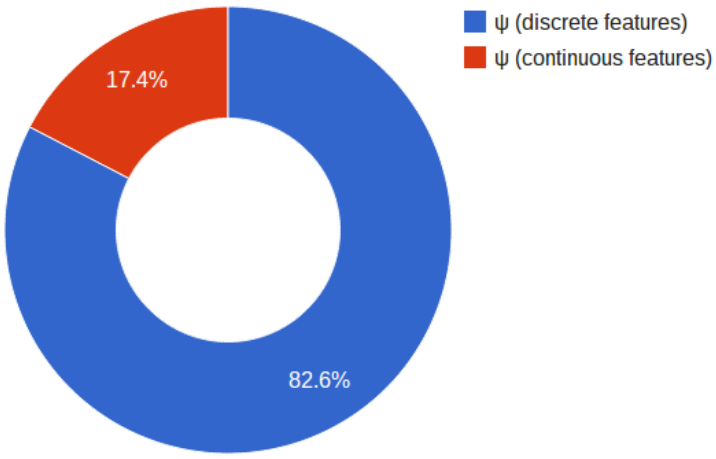


Figure 3.4: Analysis of the relative importance of the two classes of features used in the "Abstract Scenes Data-Set"'s image representation (discrete and continuous). The ψ score of Equation (2.1) is used to compare the importance of the feature classes.

An important finding of analysing the "Abstract Scenes Data-Set" with the ψ-metric is given by Figure 3.4. Summing ψ over all word features and restricting V first to the set of discrete visual features and then to the continuous counterparts shows that the discrete visual features (e.g. object occurrences) are almost five times more predictive of salient word features than the continuous ones (e.g. relative distance between objects).

Figures 3.1–3.2 and Table 3.1 in the previous section verify the statement above: continuous visual features are much less informative generally than discrete visual features. This is likely caused by the fact that the continuous visual features capture the more subtle relationships between the entities in the images (e.g. distance between objects as opposed to object occurrence). However, subtle relations are less likely to be noted in the image descriptions,[3] whence the poor score of the continuous visual features under the ψ metric.

This section presented evidence that the discrete visual features (that capture aspects such as object occurrence) in the "Abstract Scenes Data-Set" are more predictive of salient words than the data-set's continuous visual features (that capture more subtle aspects such as distance between objects). Since this report is only interested in generating simple descriptions, the discrete visual features should provide a good basis for the description-generation process. The images in the "Abstract Scenes Data-Set" will therefore be characterised by their discrete visual features alone from here onwards. This labelling reduces the complexity of the proposed models with little loss in word-feature-predictive power.

[3] After all, the human descriptions in the "Abstract Scenes Data-Set" are rather short (the average description length is 6 words) and thus necessarily simplistic.

3.2 Building the Template Model

This section discusses the implementation choices and parameters used to instantiate the three sub-components of the template model described in Section 2.3.1 (template acquisition, description-generation and description-selection). The effect of the choices is quantified using the automatic evaluation techniques of Section 2.4.1.

The models presented in this section are trained on 70% of the "Abstract Scenes Data-Set" (i.e. 7,014 images, 42,276 descriptions) and tested on a validation set spanning 10% of the data-set (901 images, 5,430 descriptions). The remaining 20% of the data (2,004 images, 12,081 descriptions) is not touched in this section and will be used as an evaluation set in Chapter 4.

3.2.1 Template Acquisition

The template-acquisition component of the template model uses a classifier to predict the grammatical structure for a given image (Section 2.3.1.1). Subsequently, this structure will be used as the basis for descriptions that the template model generates.

This section investigates two main questions. Firstly, is this complex, prediction-driven approach really necessary? And if so, what implementation of the method maximises the quality of template acquisition?

In order to tackle these questions, the performance of the template-acquisition strategy of Section 2.3.1.1 is compared to a naive "fixed template" baseline. The baseline simply predicts the most common grammatical structure in the corpus. Conversely, the more involved strategy of Section 2.3.1.1 uses a classifier to predict salient grammatical structures based on the visual features of the image at hand.

In order to get an impression of which type of model works best for the template acquisition task, we present results for a broad range of standard classifiers:

- Naive Bayes (Rish, 2001; Zhang, 2004)
- Decision Tree (Olshen and Stone, 1984)
- Random Forest (Breiman, 2001)
- Logistic Regression (Yu et al., 2011)

It is beyond the scope of this report to give explicit detail on the workings of the classifiers. The interested reader is invited to consult the references provided above and refer to Caruana and Niculescu-Mizil (2006) for an in-depth comparison of the models.

Table 3.3 shows that the baseline template-acquisition strategy depreciates the BLEU and METEOR scores of the template model. The more involved acquisition strategy of Section 2.3.1.1 outperforms the baseline by up to 16% in BLEU score and up to 7% in METEOR score. The maximal improvement over the baseline performance is achieved

Template Acquisition Strategy	BLEU score	METEOR score
Baseline (fixed template)	32.6	27.4
Logistic Regression	39.9	30.0
Random Forest	37.8	29.4
Decision Tree	37.8	28.3
Gaussian Naive Bayes	11.3	22.9
Multinomial Naive Bayes	15.1	23.0

Table 3.3: BLEU and METEOR scores (higher is better) for a template model with different template-acquisition components. The components use different classifiers to predict templates for images.

when basing the template-acquisition model on a Logistic Regression classifier.[4,5] Given the difference in competence, the complexity added by the predictive template acquisition strategy is justified.[6]

The observed drop in BLEU and METEOR implies that the grammatical structure used to describe images relates to the image's semantics. Not all grammtical structures are adequate to describe all images. For instance, images featuring multiple agents are likely to be described better with a grammatical structure containing a conjunction:

Mike and Jenny are playing baseball.

Images containing unusual features require grammatical structures that allow for more specificity with the descriptive language (e.g. adverbs or adjectives):

Mike is wearing a silly hat.

This section discussed the implementation of the template-predictor proposed in Section 2.3.1.1 and justified the component's inclusion in the template model. The section found that using a Logistic Regression classifier to predict the grammatical structure of image descriptions improves performance by up to 16% over a naive fixed-structure strategy.

[4]In order to verify this result, the different classifiers used to perform template-acquisition are evaluated in isolation of the rest of the template generation model, using the Levenshtein distance metric (Levenshtein distance is used to quantify how close the predicted templates are to gold-standard templates extracted from human generated descriptions). Table 3.4 shows the results of the evaluation: the Logistic Regression classifier produces templates that are closest on average to the gold-standard.

[5] Naive Bayes methods likely do not work well because the visual features used for prediction violate the classifier's independent assumption (e.g. occurrence of hats and sunglasses is not independent of occurrence of people). The Decision Tree classifier performs poorly due to the high dimensionality of the visual features. Random Forest and Logistic Regression are generally strong classifiers and likely work well in this scenario due to their robustness to noise as well as high dimensional and correlated data.

[6]Note that this finding (using templates specific to the described images rather than a one-template-fits-all approach improves performance) is consistent with Mitchell et al. (2012).

Classifier	Levenshtein Distance			
	Mean	Standard Deviation	Maximum	Minimum
Baseline (fixed template)	2.70	0.00	2.70	2.70
Logistic Regression	2.55	1.38	5.65	0.00
Random Forest	2.60	1.33	5.75	0.10
Decision Tree	2.78	1.42	5.55	0.00
Gaussian Naive Bayes	3.29	1.46	5.80	0.00
Multinomial Naive Bayes	2.70	1.56	5.75	0.00

Table 3.4: Isolated evaluation of how well different classifiers perform the template-acquisition task. Given an image, a template is generated and compared against templates extracted from the 5 gold-standard descriptions of the image. The reported Levenshtein metrics (lower is better) are averages of these comparisons over the 5 standard gold-standard templates.

3.2.2 Description Generation

The description-generation component of the template model (Section 2.3.1.2) can be summarised as follows. For any given image, M candidate image description templates are acquired. Then, the templates are expanded to candidate descriptions by predicting N words relevant to the image for every slot in every template. The description-generation component thus has three parameters: M, the number of candidate templates, N, the number of candidate words for every template slot and the type of the word-prediction classifier used to generate words to fill the templates. This section presents optimal values for all three parameters.

Table 3.5 shows the performance of the template model for different values of N and M and a variety of word-prediction classifiers. We find that regardless of the values of N and M, using a Logistic Regression classifier to predict salient words for images gives the best results.[7] Figure 3.5 therefore plots the subset of information in Table 3.5 pertaining to the Logistic Regression model in order to facilitate interpretation.

The table and figure reveal the following:

1. When increasing M (the number of candidate templates), the performance of the template model increases, but only up to a certain point ($M = 4$ is optimal). Note that, consistent with the findings of Section 3.2.1, Figure 3.5 verifies that a fixed template acquisition strategy produces worse results than a flexible template acquisition strategy.
   Some images are best described using less common grammatical structures. Using too many candidate templates will lead to the language model rejecting the less common grammatical structures (the more common grammatical structures "swamp" the less common ones). This leads to poorly fitted image descriptions and the observed drop in performance for large values of M.

[7] The competence of the different classifiers explored can be explained by an argument similar to the one in Footnote 5.

2. If increasing N (the number of candidate words per candidate template slot), the performance of the template model increases with $\log N$. Due to engineering considerations (memory consumption, run-time), only values of N less than five are considered in here. Arguably, $N = 4$ is a good value anyway because the trend for $1 \leq N \leq 4$ implies that the performance improvement resulting from larger values of N will plateau for $N > 4$.

Figure 3.6 shows that tuning the regularisation penalty and regularisation strength[8] of the Logistic Regression classifier adds a further margin of improvement. L1 regularised regression[9] outperforms L2 regularisation.[10] This can be explained by the observation that there is a relatively high number of features to be learned compared to training examples. This traditionally means that a L1 regularisation norm is more appropriate than L2 regularisation (Ng, 2004). Furthermore, decreasing the inverse regularisation strength parameter (i.e. using stronger regularisation) improves performance because it keeps the words used in the image descriptions relevant to the image contents (i.e. stronger regularisation prevents the language model prior from taking over).

This section presented different options for the implementation of the description-generation component of the template model proposed in Section 2.3.1.1. Generating image descriptions using four candidate templates, four candidate words per template slot and a Logistic Regression classifier to predict the candidate words produces the best results.

3.2.3 Description Selection

The description-selection component of the template model (Section 2.3.1.3) takes a list of candidate image descriptions and selects the most human-like. This is achieved by linearly combining the scores of three language models (bigram, trigram and description length) to score every candidate description. Thus, the description-selection component of the template model has three parameters: c_2, c_3 and c_l. These three parameters are the influence of the three language models on the selection procedure. This section presents optimal values for all three parameters.

The influence of the three language model components weights on the performance of the template model is shown in Figure 3.7. The values $c_2 = 0.3$, $c_3 = 0.3$ and $c_l = 0.4$ achieve a good balance between high BLEU and high METEOR scores.

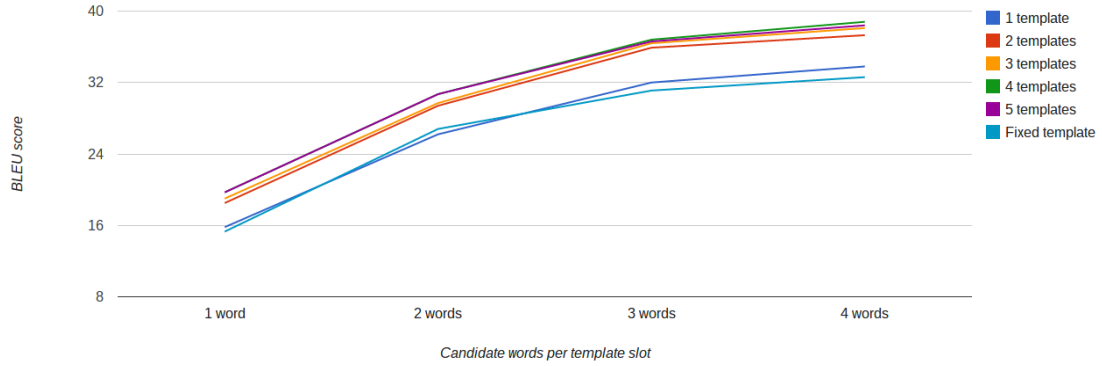
This section illustrated the implementation of the description-selection component of the template model proposed in Section 2.3.1.3. The component works best when the

[8] A simple formulation of Logistic Regression is $\arg\max_w P_w - \gamma R_w$ where P_w is the likelihood of the training data given some model features w. In this formulation, R_w is the regularisation penalty that determines how the model's complicated features are penalised and γ is the regularisation strength that determines how much weight is given to the regularisation penalty.

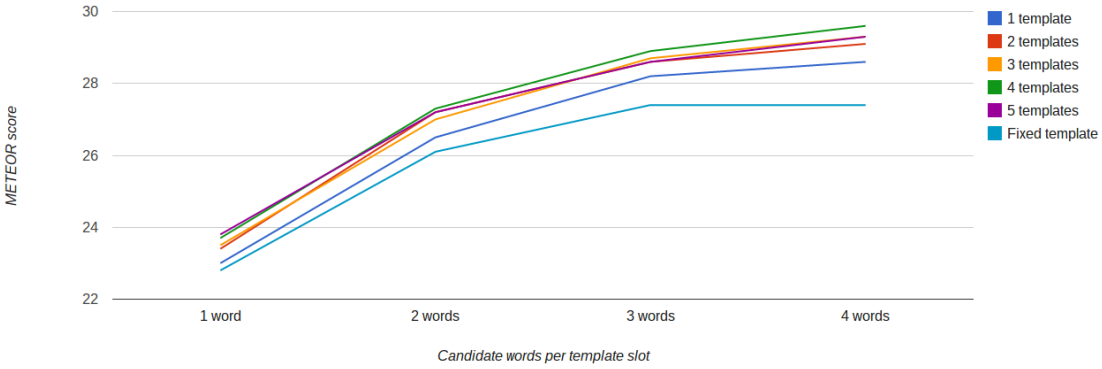
[9] L1 regularised regression penalises large feature weights with $\sum_{i=1}^{n} |\theta_i|$, where n is the number of features θ_i of the model and $|\cdot|$ is the absolute value.

[10] L2 regularised regression uses the following formula to penalise large feature weights: $\sum_{i=1}^{n} \theta_i^2$, where n is the number of features θ_i of the model.

three aforementioned language models have about equal weights (30% bigram model weight, 30% trigram model weight and 40% description length model weight).



(a) BLEU score



(b) METEOR score

Figure 3.5: BLEU (top) and METEOR (bottom) scores (higher is better) for a template model with different description-generation components. The components use a Logistic Regression word-predictor, up to $M = 5$ candidate templates and up to $N = 4$ candidate words per template slot.

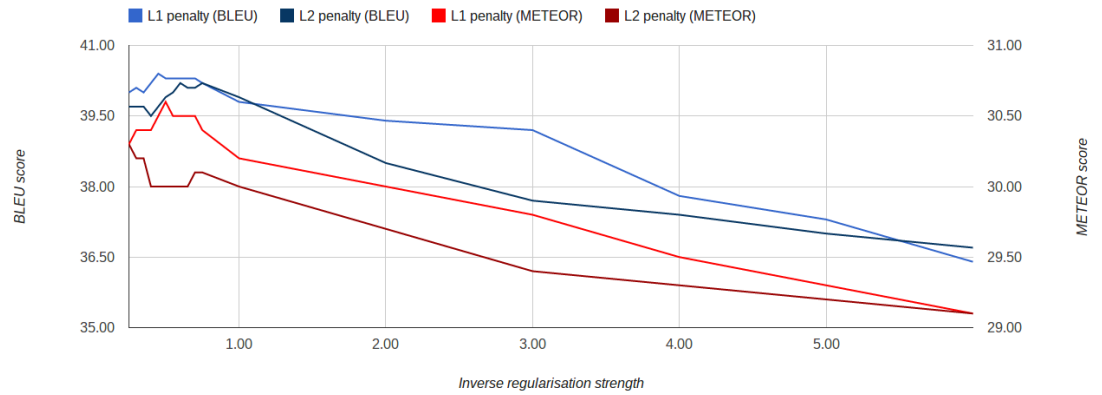
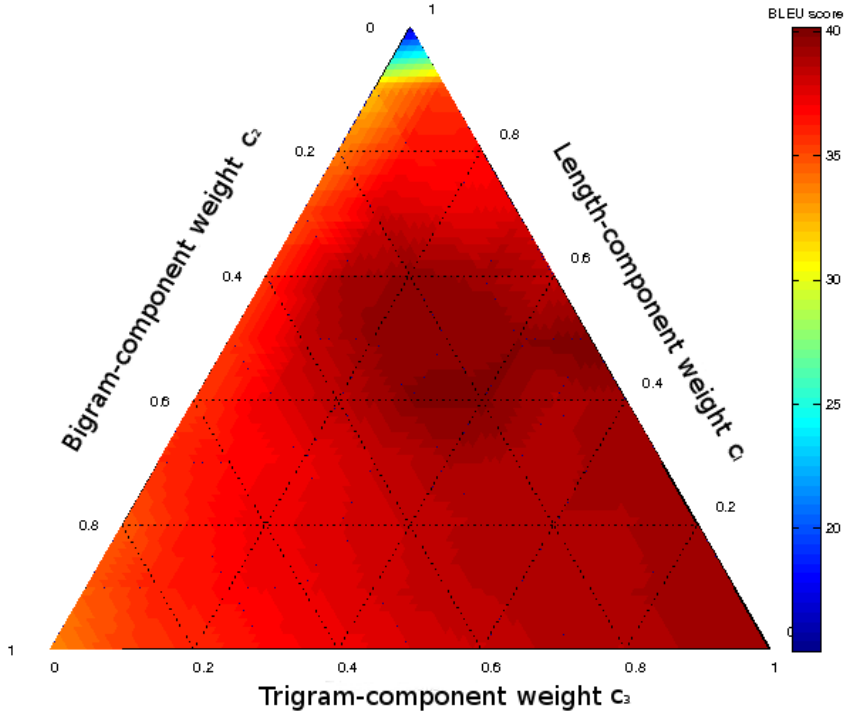


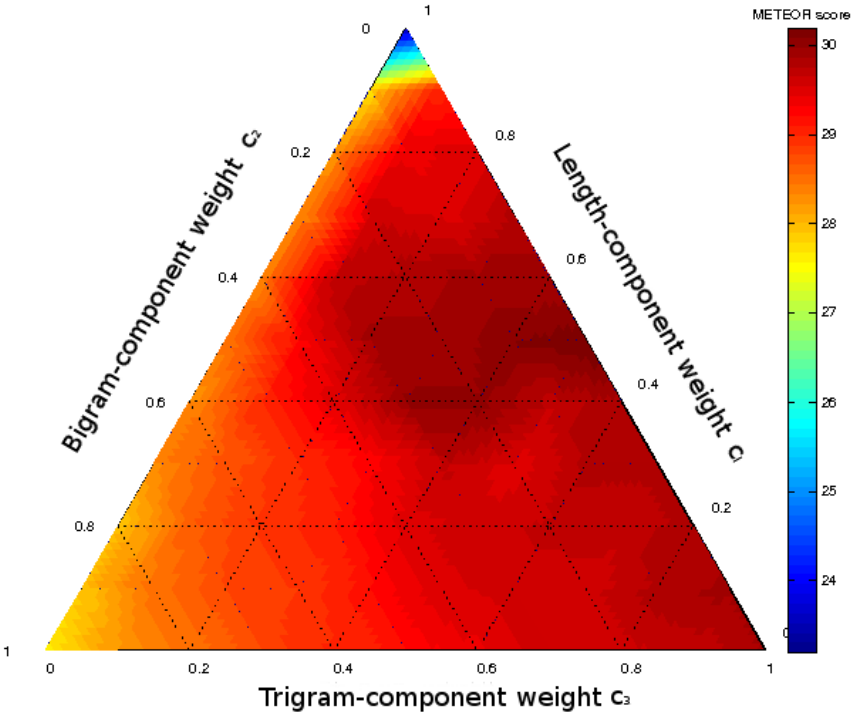
Figure 3.6: BLEU and METEOR scores (higher is better) for a template model with different description-generation components. The description-generation components vary the regularisation penalty and inverse regularisation strength of the word prediction model.

Gaussian Naive Bayes	1 word		2 words		3 words		4 words	
1 template	0.60	0.60	2.90	11.7	4.90	13.1	7.10	13.7
2 templates	0.70	0.70	3.60	11.9	4.90	13.1	7.10	13.7
3 templates	0.90	0.90	3.40	11.9	5.20	13.1	7.30	14.0
4 templates	0.90	0.90	3.80	12.1	5.40	13.1	7.10	14.0
5 templates	0.90	0.90	4.20	12.0	6.00	13.2	7.70	14.1
Multinomial Naive Bayes	1 word		2 words		3 words		4 words	
1 template	11.6	20.1	23.8	25.1	27.7	26.9	28.2	27.0
2 templates	13.7	20.4	25.6	25.7	30.1	27.6	30.3	27.6
3 templates	14.1	20.5	26.6	26.0	31.8	28.1	31.6	27.8
4 templates	14.8	20.5	27.6	26.2	32.3	28.2	32.1	28.1
5 templates	15.1	20.7	27.5	26.1	32.1	28.0	32.0	28.0
Decision Tree	1 word		2 words		3 words		4 words	
1 template	4.70	18.2	13.1	22.5	18.6	24.6	23.0	25.6
2 templates	5.00	18.2	15.1	22.8	21.4	25.0	26.0	26.0
3 templates	5.20	18.4	16.5	23.0	22.8	25.2	27.4	26.2
4 templates	5.40	18.3	16.5	23.0	23.3	25.4	27.9	26.4
5 templates	5.40	18.2	16.8	22.9	23.6	25.4	27.8	26.3
Random Forest	1 word		2 words		3 words		4 words	
1 template	13.1	22.7	24.8	24.8	29.3	29.3	30.9	30.9
2 templates	14.4	23.0	29.6	26.8	33.6	28.0	34.2	28.1
3 templates	15.5	23.2	30.3	26.8	35.3	28.3	35.4	28.5
4 templates	15.9	23.2	30.5	26.8	35.8	28.4	35.8	28.5
5 templates	15.7	23.2	30.6	26.8	36.1	28.4	35.7	28.5
Logistic Regression	1 word		2 words		3 words		4 words	
1 template	15.8	23.0	26.2	26.5	32.0	28.2	33.8	28.6
2 templates	18.5	23.4	29.4	27.2	35.9	28.6	37.3	29.1
3 templates	19.0	23.5	29.7	27.0	36.4	28.7	38.1	29.3
4 templates	19.7	23.7	30.7	27.3	36.8	28.9	38.8	29.6
5 templates	19.7	23.8	30.7	27.2	36.6	28.6	38.4	29.3

Table 3.5: BLEU (left) and METEOR (right) score pairs for a template model with different description-generation components. The components use different word-predictors, up to $M = 5$ candidate templates and up to $N = 4$ candidate words per template slot. Good scores are highlighted with a dark shade of red, bad scores are highlighted with a dark shade of blue.



(a) BLEU score



(b) METEOR score

Figure 3.7: BLEU (top) and METEOR (bottom) scores for a template model with different description-selection components. The description-selection components vary the relative importance of the bigram, trigram and length language models. Better scores are denoted in dark shades of red. Worse scores are coloured with dark shades of blue.

Chapter 4

Results

This chapter reports the performance of the models described in Chapter 2 as implemented using the methods described in Chapter 3. The models are trained on the union of the training and validation set mentioned in Chapter 3 and evaluated on the remaining data. 80% of the "Abstract Scenes Data-Set" (8,016 images, 48,315 descriptions) is thus used for model training and 20% of hitherto unseen data (2,004 images, 12,081 descriptions) is used for evaluation.

The remainder of this chapter splits performance evaluation into three main sections.

First, Section 4.1 uses the automatic evaluation techniques described in Section 2.4.1 to compare and contrast the performance of the template model with the baseline models.

In order to complement the automatic evaluation, we analyse the performance of the models using manual inspection and report the findings in Section 4.2.

Section 4.3 then completes the evaluation of the models by presenting and analysing the results of the human evaluation experiment described in Section 2.4.2.

4.1 Automatic Evaluation

This section gives an overview of the performance of the template model, keyword baseline model and image similarity baseline model described in Sections 2.3.1–2.3.2. The BLEU and METEOR metrics described in Section 2.4.1 are used to obtain a high-level perspective of the performance of the models.

Section 2.4.1 mentioned that one of the drawbacks of the BLEU and METEOR metrics is their lack of interpretability. Specifically, the metrics do provide a way to compare the performance of two systems (if system X has a higher score than system Y, X likely is the better system), however, the metrics do not give any insight on what it *means* to obtain a score of Z. In order to address this lack of interpretability, the human performance in the image description task is measured using BLEU and METEOR. The measurements are then used to ground the performance of the automatic image description models.

Model	BLEU score	METEOR score
Image Similarity Baseline	12.80	21.77
Keyword Baseline	14.70	26.60
Template Model	40.30	30.40
Human Agreement	21.17	25.52

Table 4.1: Comparison of image description correctness (as measured by BLEU and METEOR scores) for a template-based image description model and two baselines. Human performance[2] is included as a reference to enable better interpretation of the scores.

Human performance in the image description task is approximated by measuring inter-annotator agreement on the "Abstract Scenes Data-Set." Said inter-annotator agreement can be computed as follows. Let I be the set of all images in the "Abstract Scenes Data-Set." Every image λ in I has n^λ descriptions: $d_1^\lambda, d_2^\lambda, \ldots, d_{n^\lambda}^\lambda$. Let D_i be the set of all the i^{th} descriptions of all the images in the data-set. Now recall that BLEU and METEOR scores are obtained by computing ngram overlaps between a set of hypothesis sentences and a set of reference sentences. The inter-annotator agreement A_M for a metric M[1] can therefore be computed by using the set of all i^{th} descriptions (D_i) as the hypotheses and the set of all other descriptions (D_{-i}) as the reference. Then, the result is averaged over all values of i. Equation (4.1) summarises the computation of the human agreement.

$$\begin{aligned} A_M &= \frac{1}{n_{\min}} \sum_{i=1}^{n_{\min}} M(D_i, D_{-i}), \\ D_i &= \left\{ d_i^\lambda \mid \lambda \in I \right\}, \\ D_{-i} &= \bigcup_{j \in \{1,2,\ldots,n_{\min} \mid j \neq i\}} D_j, \\ n_{\min} &= \min \left\{ n^\lambda \mid \lambda \in I \right\} \end{aligned} \tag{4.1}$$

Table 4.1 (above) uses Equation (4.1) to measure how well the human performs the image description task. The table also shows the performances of the models described in Section 2.3. Note that the human performance is relatively low. As a matter of fact, one of our automated image description models even out-performs the human annotators! This highlights a further pitfall of using BLEU and METEOR to judge the quality of image descriptions. The image description task is highly subjective: humans will often describe very different things about the same image, or at least describe similar things using different language. A good example for this is the image in Figure 4.1: five out of the six human-generated descriptions of the image refer to entirely separate aspects of the image. Human agreement is thus necessarily low:

[1] Here, M is either BLEU or METEOR.

[2] As measured by inter-annotator agreement on the "Abstract Scenes Data-Set."

the subjectivity of image description and the way that BLEU and METEOR measure "goodness" do not match well. Figure 4.2 further illustrates this shortcoming. The template model, keyword baseline and image similarity baseline describe the image as follows: *Mike is wearing a witch hat*, *A helicopter is flying by Mike and Jenny* and *There is a big cloud in the sky* — all perfectly fine descriptions. Note, however, that the first and third of these descriptions will be heavily penalised by BLEU and METEOR because they happen to deviate in focus from the human annotators. Thus, Table 4.1 highlights the need for human evaluation as back-up and complement to automatic evaluation via BLEU and METEOR.

Besides human agreement, Table 4.1 also gives BLEU and METEOR scores for the template model (Section 2.3.1), keyword baseline (Section 2.3.2.1) and image similarity baseline (Section 2.3.2.2). We discuss these scores below.

The performance of the two baseline models is bounded by the inter-annotator agreement. Both baselines describe images by retrieving appropriate human generated descriptions. It is therefore difficult for the models to outperform the inter-annotator agreement. "Difficult", not "impossible." Table 4.1 shows that the keyword baseline's METEOR score is about 4% relative (or 1.08 points absolute) higher than the inter-annotator agreement. This can be explained observing that the keyword baseline may retrieve a description that has a higher overlap with the set of human descriptions of an image than any singular gold-standard description of that image. For example, the retrieved description might synthesise multiple gold-standard descriptions of the image. However, the chances that such a great description exists and is retrieved is low since a sufficiently similar image must have been seen during model training time and that image must have been described in a sufficiently exhaustive manner. Hence the statement holds: the performance of the baseline models is bounded by the inter-annotator agreement.

The performance of the template model on the other hand, quite substantially exceeds the baselines and inter-annotator agreement (+90% relative BLEU score and +19% relative METEOR score as per Table 4.1). As mentioned previously, however this does not mean that the template model produces better descriptions than humans. The good performance of the model under the automatic evaluation metrics is simply an artifact of the subjectivity of the image description task (i.e. the template model is being compared against a human agreement that is artificially low). Nevertheless, the template model's high scores and specifically the fact that the template model outperforms both retrieval-based baselines does highlight one of the model's strengths. Unlike the baseline models, the template model generates novel image descriptions instead of retrieving human-generated ones. This means that the template model is more flexible with the content of its image descriptions. For instance, if the template model thinks that *Jenny*, *Mike*, *owl* and *ball* are important words for an image, the model prioritises these words to fashion a new description using all of them. A human generated description, on the other hand, is likely to only focus on one or two of the aspects. Thus created, the template model's description is likely to encompass multiple gold-standard descriptions, meaning that the

Figure 4.1: Sample image from the “Abstract Scenes Data-Set.” The human-generated descriptions for the image are:

d_1 : *Jenny is pretending to be a witch.*
d_2 : *Jenny is wearing the hat.*
d_3 : *Mike is eating a hot dog.*
d_4 : *Mike is eating the hot dog.*
d_5 : *The pie is on the table.*
d_6 : *The snake is using balloons to float.*

Figure 4.2: Sample image from the “Abstract Scenes Data-Set.” The human-generated descriptions for the image are:

d_1 : *A big helicopter flies above Jenny.*
d_2 : *Jenny almost burned her hand on the fire.*
d_3 : *Jenny doesn’t like the fire.*
d_4 : *Jenny sat close to the fire.*
d_5 : *Mike is trying to scare Jenny.*
d_6 : *Mike jumped when he saw the helicopter.*

construction will score well under BLEU and METEOR.[3]

This section quantified the performance of the template model, keyword baseline model and image similarity baseline model using the BLEU and METEOR automatic evaluation metrics. The template model outperformed both baselines. The performance of all three models was further compared against human performance in the image description task (as measured by inter-annotator agreement on the "Abstract Scenes Data-Set"). It was found that the template model's performance exceeded the inter-annotator agreement due to the highly subjective nature of the image description task (leading to low inter-annotator agreement) and due to the template model's ability to construct descriptions that synthesise image contents well.

4.2 Performance Analysis

The previous section gave a high-level overview of the performance of Section 2.3's models. This section aims to complement and concretise that impression by analysing the performance of the three models via providing sample images and descriptions generated by the models.

The overarching aim of the section is to introduce a better feel for the models' performance, before subjecting the models to human evaluation in Section 4.3. The section should be read in conjunction with Appendix C which provides additional images illustrating the points argued below.

Investigating the image descriptions produced by the three image description systems of Section 2.3 suggests that some observations hold for all three models alike. For instance:

- Figure 4.5 and Figures C.1–C.2 in Appendix C suggest that all three models can perform quite admirably, producing descriptions that match a broad range of images very well.
- Figure 4.6 suggests that the models are also able to produce reasonable descriptions for strange and out-of-the-ordinary images.
- Figure 4.7 suggests that some images are difficult to describe for all models.

Manually inspecting the descriptions produced by the model-trio generally confirms the observations of the last section. The template model consistently produces the best results, the image similarity baseline is the worst model and the keyword baseline is somewhere in between, ranging from excellent to terrible.

The remainder of this section investigates the performance of each of the three models individually.

[3]The observation that generation-based image description models can produce descriptions more specifically matching images than retrieval-based models is consistent with Kulkarni et al. (2011) who also observe this.

4.2.1 Template Model

The content of the descriptions generated by the template model (unsurprisingly) follows a Zipfian distribution:[4] a few phrases are used to describe lots of images and lots of phrases are generated for only a handful of images. It is therefore prudent to check whether the most commonly generated descriptions are appropriate for their corresponding images. Figures 4.8–4.10 and Figures C.3–C.8 in Appendix C show a selection of images that were described by the template model using some of the most frequent phrases in that model's vocabulary. The descriptions all match the images (or at least parts of them) well.

Note that the commonly used "description building blocks" introduced in the last paragraph[5] often cover the main action in the described image. The template model thus has the beneficial property that its most frequently generated descriptions capture the most salient parts of the described images instead of generic, always-applicable aspects (e.g. describing elements of the image that are always present such as the sky or the grass).

More broadly speaking, the template model often produces very adequate descriptions or descriptions that are at least relevant to a subset of the described image (refer to any of the figures in this section or Appendix C to find examples verifying this statement).

Additionally, the template model rarely generates descriptions that are truly wrong. When the model produces incorrect descriptions, usually only one or two words in the description are mistaken (e.g. switching *Jenny* for *Mike* or using *in* instead of *near*). This observation further explains the template model's good performance under BLEU and METEOR: one or two incorrect words will not substantially harm ngram overlap scores whence the model will get partial credit for somewhat-correct descriptions.

A more in-depth investigation of the errors systematically committed by the template model reveals that the model suffers from three main problems.

First and most importantly, certain words have a strong prior in certain grammatical functions. For instance, humans overwhelmingly describe images that contain the little boy by using *Mike* as the active subject of their description. Therefore, the template model has a strong bias for using *Mike* in a subject role. This leads to descriptions where *Mike* is erroneously used as the subject of a sentence whose focus is on a completely different part of the image or even to descriptions where *Mike* is used as subject despite the boy not being present in the image. Figure 4.11 gives example image descriptions exhibiting this default.

Secondly, unlike other image descriptors in the literature (e.g. Yang et al. (2011)), the template model has no explicit notion of semantic grounding. This means that the template model does not penalise descriptions that are clearly nonsensical such as *Mike*

[4] A Zipfian distribution is a discrete power law distribution where rank is roughly inversely proportional to frequency. The distribution explains many phenomena in natural language (Manning and Schütze, 1999, Section 1.4.3).

[5] These building blocks are frequently occurring description fragments such as *Mike is wearing*, *Mike is sitting*, *Mike and Jenny* and so forth.

is sitting in the sky (Figure 4.12). Arguably, the description generator's language model should obstruct these sentences, but evidently it fails to do so. This is likely due to the dearth of data that the template model's language model was trained on.

The final systematic error committed by the template model stems from the very foundations on which the model is built. The model generates descriptions by filling in templates consisting of subjects, objects, verbs and so forth. This means that the model has difficulties dealing with collocations. Sometimes, descriptions are short a word due to a lack of an available and appropriate template-slot. An example for this is the generated description *Mike is wearing a baseball ~~glove~~*. Here, the template model did not provide for a collocated word at the sentence end. More examples of this phenomenon can be found in Figure 4.13.

Note however, that in the grand scheme of things, the template model commits relatively few errors. The down-side of this is that the template models' descriptions are never truly exciting. The model factually describes objects and relations that are explicitly present in the image, leaving little to no room for interpretation or imagination. While, arguably, objectivity is a good quality for an automated image description system, it also makes the model's productions seem slightly dull to the human.

4.2.2 Keyword Baseline

Now compare the "correct-but-boring" template model to the keyword baseline. The later can produce descriptions that, to the human, sound surprisingly intricate because they are open to interpretation (e.g. see Figure 4.3) or relatively complex (e.g. see Figure 4.4). However, this predisposes the model to frequent failure by generating descriptions that are not even remotely relevant to the image at hand (e.g. see Figures 4.14–4.15).

There are two main types of errors that the keyword baseline commits systematically.

Firstly, the baseline is prone to "false associations" i.e. the model retrieves a description that matches some of the keywords for an image but not the contents of the image itself (see Figure 4.14: a pink dress or a blue shirt in the image lead to descriptions about a pink bucket or a blue hat being retrieved). In the template model, selecting less salient keywords to describe the image leads to only one or two incorrect words — with the keyword baseline, on the other hand, the entire description risks being inadequate. This is a nice demonstration of one of the chief advantages of the generation-based image description paradigm over the retrieval-based one.

Secondly, the baseline struggles to deal with unusual elements in images. Unusual image elements lead to high inverse-document-frequency (IDF) keywords being generated. Consequentially, the high IDF keywords bias the description retrieval process, leading to the selection of overly specific or nonsensical descriptions (see Figure 4.15: clouds or balloons in an image lead to descriptions about rain or hot air balloons being retrieved).

Thus, the keyword model is simple, sometimes wrong, but still often effective. Note that the model can only get better as the size of the available data (and therewith the pool

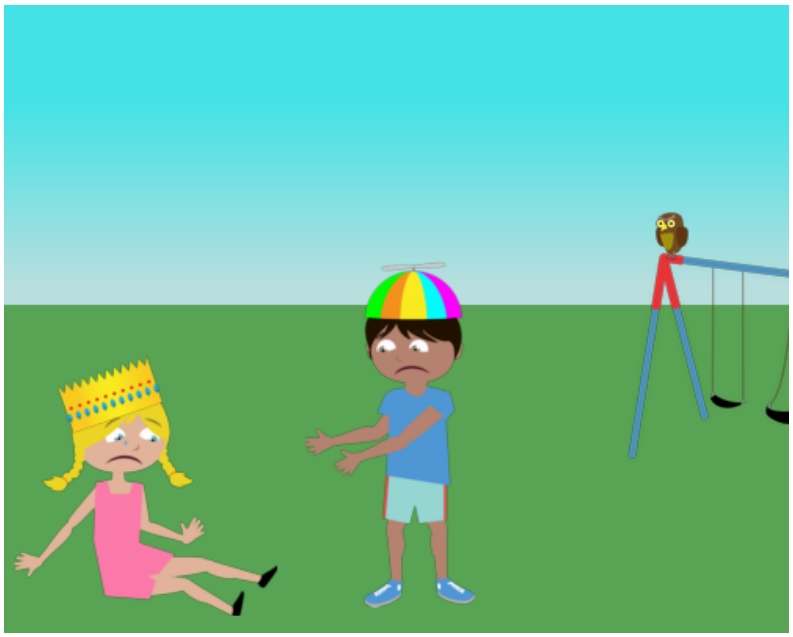


Figure 4.3: Sample image from the “Abstract Scenes Data-Set.”
Our automatic image description systems describe this image as follows:
Keyword baseline: *Mike and Jenny are sad because the kite flew away.*
Image Similarity baseline: *Mike is reaching his arms forward.*
Template model: *Mike is wearing a hat.*

Figure 4.4: Sample image from the “Abstract Scenes Data-Set.”
Our automatic image description systems describe this image as follows:
Keyword baseline: *Mike is holding Jenny’s hot-dog.*
Image Similarity baseline: *The baseball is in the baseball glove.*
Template model: *Mike is holding a hamburger.*

of retrievable descriptions) grows (Ordonez et al., 2011; Krishnamoorthy et al., 2013). As a matter of fact, models similar to the keyword baseline have been proposed in the literature and were found to work remarkably well given a large enough collection of images and descriptions to work with (refer to Ordonez et al. (2011) and Kuznetsova et al. (2012) for directly comparable approaches or refer to Berg et al. (2010), Kulkarni et al. (2011) and Mitchell et al. (2012) for related approaches).

4.2.3 Image Similarity Baseline

Lastly, consider the image similarity baseline. Figure 4.16 (and numerous other examples in this chapter and Appendix C) suggests that the baseline often fails to produce adequate descriptions for images. This is likely caused by the two major sources of noise that the model imposes onto the description retrieval process by design. In order for the image similarity model to produce a valid description for an unseen image, it needs to find a sufficiently similar image to the unseen one and then, additionally, select a similarly well matching description from the retrieved image.

The bulk of this chapter has argued extensively that humans describe very different aspects of an image. This implies that some of the descriptions of one image will not apply to another, even if the images are somewhat similar. This means that the description selection aspect of the image similarity baseline is noisy. Additionally, examining the way that the image similarity baseline restricts the search space to find similar images (see Figure C.9 in Appendix C) suggest that the model is also rather poor at performing visual matches. It is hardly surprising then, that the image similarity baseline performs poorly and produces lots of mismatched descriptions.

This section analysed the performance of the template model, keyword baseline and image similarity baseline through manual inspection of the descriptions generated by the models. The section also investigated some of the re-occurring sources of errors for all three models. The findings of this section mirror the results of the automatic evaluation: the template model produces better image descriptions than either of the two baseline models.

4.3 Human Evaluation

This section presents the results of the human evaluation study described in Section 2.4.2. The study recruited 100 human judges on Amazon Mechanical Turk to evaluate the quality of 200 descriptions generated by the template model and keyword baseline using a five-point Likert scale. In order to ground the ratings of the two systems, the judges were also asked to evaluate the quality of a randomly selected human-generated gold-standard description for every image.

Inspection of the judges' ratings led to the rejection of 17 participants: their scores did

not look natural (e.g. uniformly distributed ratings). A one-way ANOVA[6] with post-hoc Tukey HSD test[7] was used to analyse the remaining human judges' ratings.

Table 4.2 shows the results of the study. The full HSD matrix of the tests is in Table 4.3. Both models (keyword- and template-based) are significantly different from the human gold-standard ($p < 0.01$). There is no statistically significant difference between the template model and the keyword baseline.

There are several conclusions to be drawn from this study. Firstly, note that the human judges were unable to differentiate between the performance of the template model and keyword baseline despite the former model's descriptions being machine generated to the latter model's human generated descriptions. The humans who wrote the descriptions of the keyword baseline presumably highlighted salient aspects of the described images. This study thus implies that the template model is also able to do this.

Secondarily, the discrepancy between the results of the human evaluation study and the automatic evaluation of Section 4.1 is of interest. How can the template model vastly outperform the keyword baseline from BLEU and METEOR's perspective and yet be indistinguishable from a human's point of view? One way to explain the difference is the observation that, unlike the automated evaluation metrics, humans penalise the types of errors made by the template model heavily (e.g. confusing the subject or primary object of a description). This behaviour was initially observed when running the human evaluation study informally with a small amount of face-to-face participants and holds for the study run at scale on Amazon Mechanical Turk. The results of this section can thus inform future work in automated image description: getting the main word of a sentence right is important — automated metrics such as BLEU or METEOR will not necessarily catch this.

This section reported the results of a study using human judges to evaluate the performance of the keyword baseline and template model. The study found no statistically significant difference between the two systems. The section argued that this result reflects favourably on the quality of the machine-generated descriptions produced by the template model. The section further argued that the finding is also indicative of a discrepancy in what humans deem important in an image description and what automated evaluation metrics measure.

[6] ANOVA (short for "analysis of variance") is a family of statistical models to analyse studies with three or more groups. ANOVA computes the probability of the null hypothesis (i.e. the samples in all groups are drawn from the same population) by comparing two estimates of the population standard deviation: one based on sample variances and the other based on differences between sample means. If the null hypothesis is true, both of these estimates should be the same, otherwise the estimate based on sample variances should be higher. The larger the difference between the two mean-estimates, the higher the probability that the samples stem from different populations. Refer to (Howell, 2012, Chapter 11) for a more thorough introduction to the procedure.

[7] Tukey HSD (short for "honest significant difference") is a statistical procedure to check if the means of different groups are significantly different from one-another. The test compares the mean of every group to the means of all other groups and identifies differences greater than the standard error. Refer to (Howell, 2012, Chapter 12) for a more thorough introduction to the test.

Model	n	μ	SD	SE
Human Gold-Standard	83	3.9157	0.5291	0.0581
Keyword Baseline	83	2.6675	0.6197	0.0680
Template Model	83	2.4053	0.6329	0.0695

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \mathrm{SD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)},\ \mathrm{SE} = \frac{\mathrm{SD}}{\sqrt{n}}.$$

Table 4.2: Results of the human study. Subjects were asked to rate the productions of two image description systems and a randomly selected gold-standard description (as a point of reference) on a five-point Likert scale. n is the number of subjects in the study, μ is the mean of the human ratings x_i, SD is the standard deviation of the mean and SE is the standard error of the mean.

Model	Mean	Keyword Baseline	Template Model
Human Gold-Standard	3.9157	$1.2482^{\dagger}$	$1.51041^{\dagger}$
Keyword Baseline	2.6675		0.2622
Template Model	2.4053		

† significant at $\alpha = 0.05$, critical difference $= 0.3925$.

Table 4.3: Full matrix of the HSD test.

Template Model	mike and jenny are playing frisbee	jenny is scared of the bear	mike is angry at jenny	mike is kicking the ball	mike and jenny are playing baseball
Keyword Baseline	jenny runs away from mike to see the hot air balloon	mike and jenny are afraid of the bear	blue duck and the dog are watching jenny	the blue duck threw the football at jenny	mike and jenny are playing baseball on a sunny day
Image Similarity Baseline	mike is catching jenny's frisbee	mike puts the pizza on the table	jenny and mike were playing in the sandbox	jenny has kicked the football to mike	jenny is sad
Human Description	jenny throws a frisbee to mike	mike and jenny are scared of the bear	jenny is crying	mike is scoring a touchdown	mike and jenny play baseball

Figure 4.5: Some images are well described by all models.

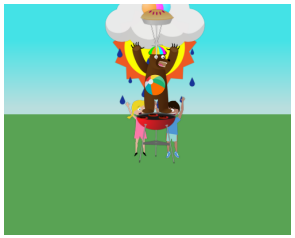
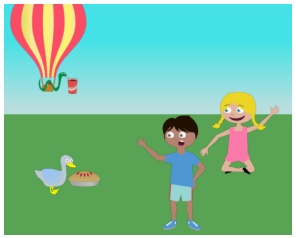
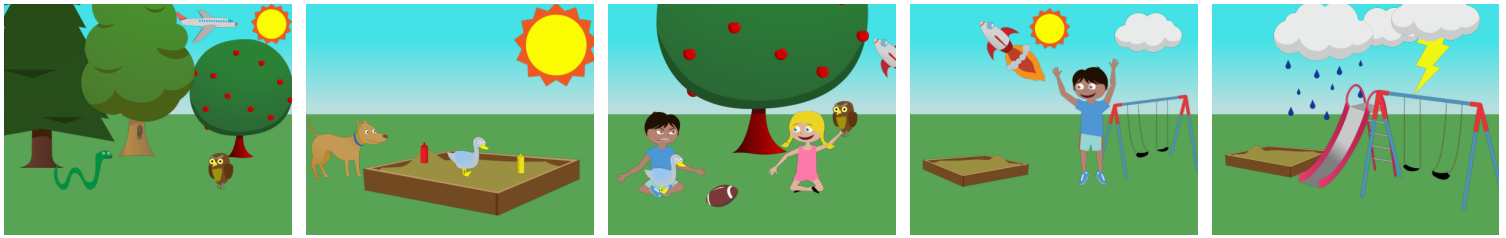
					
Template Model	mike is wearing a hat	mike is holding a pie	mike is holding a pie	mike is holding a kite	mike is holding a hamburger
Keyword Baseline	a mustard bottle is next to mike and jenny	jenny has colorful balloons	jenny and mike are waving to the duck	jenny 's balloons are getting away	mike is sitting next to jenny in the sandbox
Image Similarity Baseline	the ketchup is sitting on the table	the bear is in front of the tree	a pie is in front of the owl	mike is mad at the dog	jenny has a hotball
Human Description	mike and the cat are on the plane	the bear is wearing a propeller beanie hat	the duck is eating a pie	mike is angry at jenny because he wants her kite	the cat is in the sandbox

Figure 4.6: The models handle difficult, strange or noisy images rather well.



Template Model	the owl is in the sky	the sun is in the sandbox	mike is sitting in the sky	jenny is flying in the sandbox	mike is in the sandbox
Keyword Baseline	mike it buying apples from the owl	the dog is wearing the sun glasses	mike and jenny will be very wet	mike and jenny play in the sun	there is a pink slide a few feet from mike and jenny
Image Similarity Baseline	jenny is worried about her dog	the tree has red apples	mike threw all the balls in the air	jenny is standing by the fire	the hot air balloon is in the sky
Human Description	a hungry owl watches the snake	mustard and ketchup are in the sandbox	mike and jenny are near an apple tree	mike is watching the rocket fly away	no one is playing in the sandbox

Figure 4.7: Some images are difficult for all models.

Figure 4.8: Template model frequent phrases — sample images for descriptions starting with *Mike is wearing*.

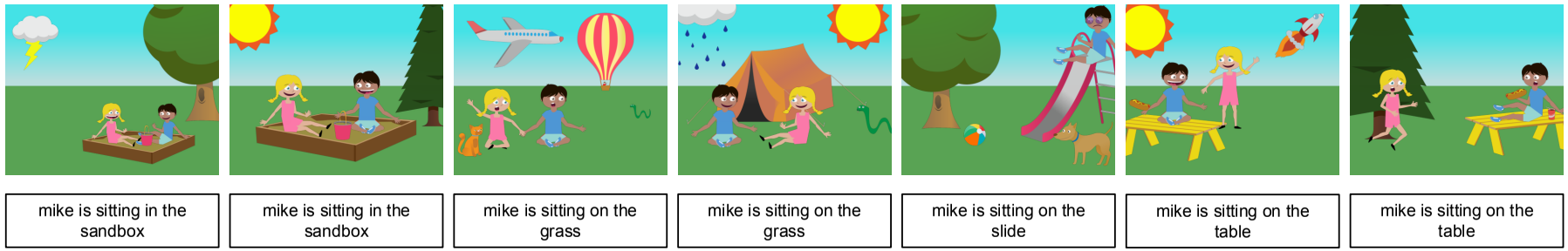


Figure 4.9: Template model frequent phrases — sample images for descriptions starting with *Mike is sitting*.

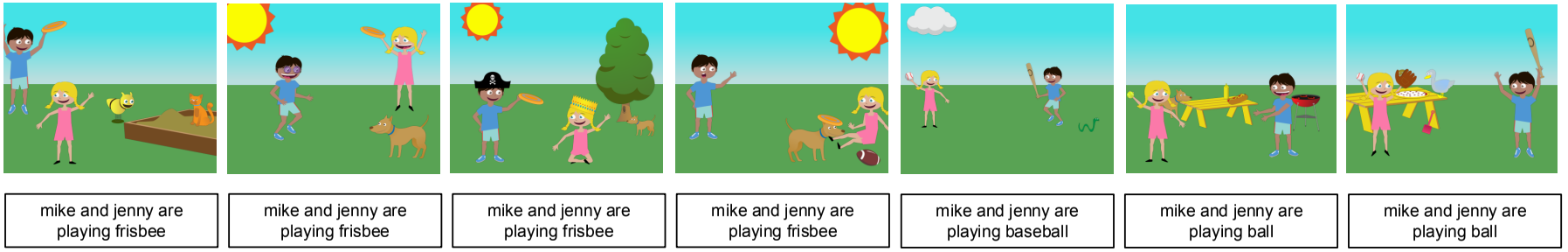


Figure 4.10: Template model frequent phrases — sample images for descriptions starting with *Mike and Jennny*.

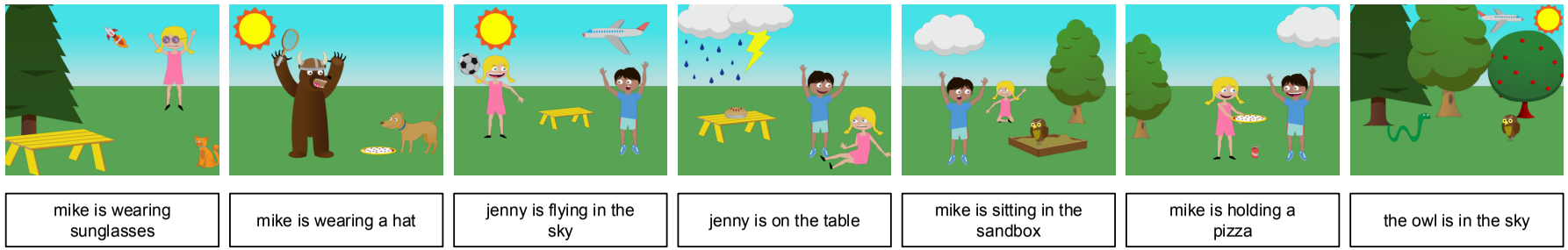


Figure 4.11: Template model systemic errors — prior on objects in certain grammatical functions leads to incorrect descriptions.

Figure 4.12: Template model systemic errors — lack of domain knowledge.

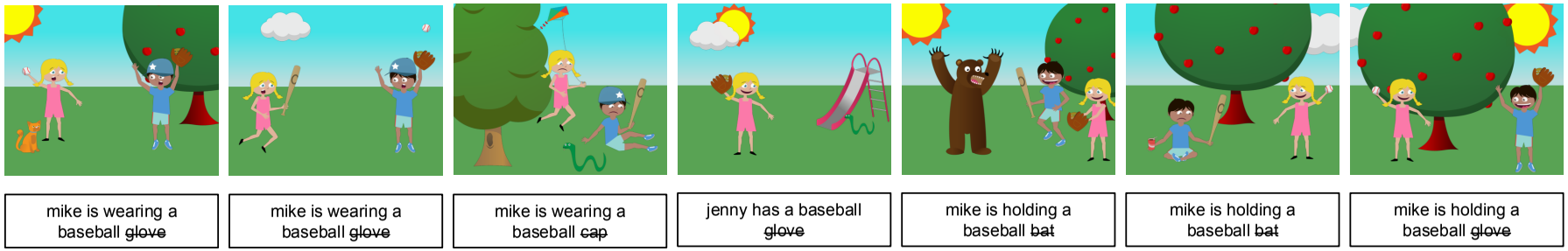


Figure 4.13: Template model systemic errors — templates do not account for collocations.

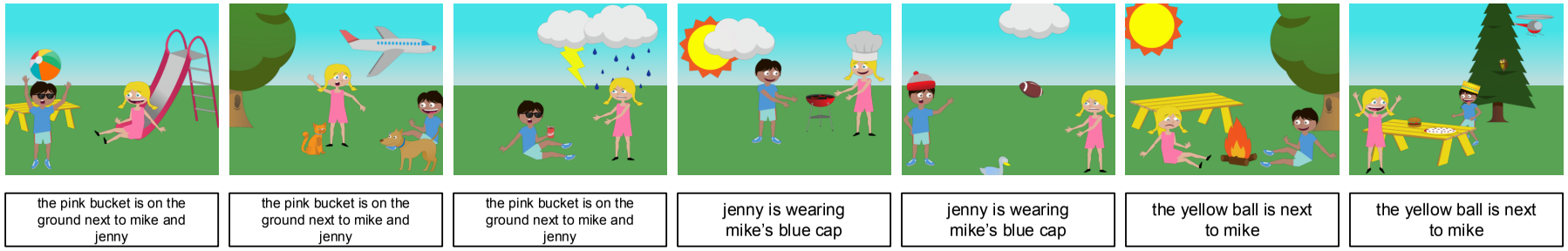


Figure 4.14: Keyword baseline generations — systemic errors due to false associations.

Figure 4.15: Keyword baseline generations — systemic errors due to terms with high IDF scores.

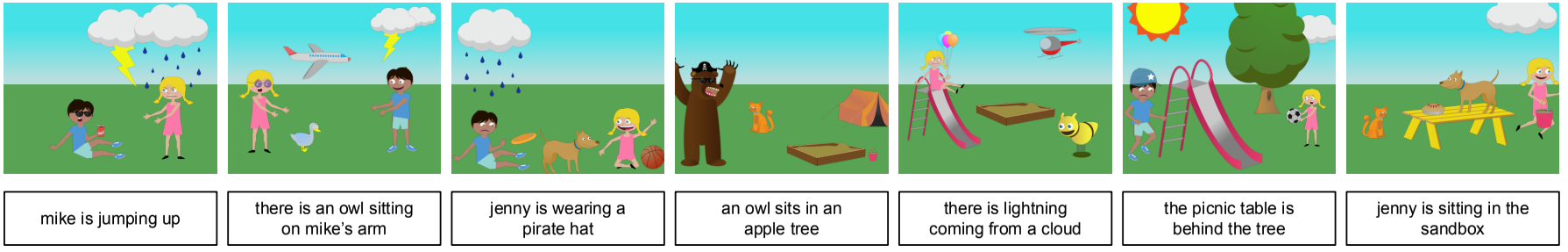


Figure 4.16: Image similarity baseline — the model simply rarely produces good descriptions.

Chapter 5

Conclusions

This report set out to investigate the vision–language interplay by studying the "Abstract Scenes Data-Set" (Zitnick and Parikh, 2013), a collection of semantically similar clip-art images and corresponding human generated descriptions. The simple nature of the images in the data-set allowed us to largely abstract away considerations of object recognition and image segmentation, thus enabling a focus on content planning, content selection and surface realisation.

This final chapter, will review the research contributions of the report (Section 5.1), examine some results that may be of wider interest for the automated image description community at large (Section 5.2) and discuss avenues for future work (Section 5.3).

5.1 Summary of Contributions

Our primary contribution of is a novel, template-based system that generates image descriptions using a data-driven approach (Section 2.3.1). Our model sub-divides the description generation process into three separate steps. First, given an unseen image, a content planner determines the most semantically salient grammatical structures for the purpose of describing that image (Section 2.3.1.1). Then, a content selector determines which terms in our model's vocabulary apply to the image, respecting the grammatical constraints imposed by the previous stage (Section 2.3.1.2). Finally, a surface realiser merges the grammatical structures of the content planner with the content selector's terms, over-generating a bag of description and selecting the best one in terms of grammaticality and adequacy (Section 2.3.1.3).

Content planning and content selection were formulated as generative classification tasks. Surface realisation was formulated as a discriminative classification task. Our model is fully data-driven. All three components of our model were trained on visual and textual features extracted from the "Abstract Scenes Data-Set" (10,020 semantically related clip-art images and 60,396 unrestricted human-generated descriptions, Section 2.1) using an information-theoretic measure (Section 2.2, Section 3.1).

Our system was tuned by optimising all components for BLEU and METEOR scores

(Section 3.2). These metrics were also used to evaluate our model end-to-end. We found that our template-based model significantly outperformed a non-trivial transfer-based baseline system (Section 2.3.2.1, Section 4.1). A human evaluation study performed on Amazon Mechanical Turk, however, did not confirm this finding (Section 4.3), likely because BLEU and METEOR do not correlate well with the human notion of image description quality (Section 4.2.1).

We found that our template-based model consistently produces descriptions of high adequacy and outstanding fluency.

5.2 Wider Considerations

The methodological elaboration and concrete implementation of our template-based model (summarised above) further led to some insights that might be of interest to the automated image description community at large. These findings are sketched below.

Analysing the textual corpus of the "Abstract Scenes Data-Set" in terms of "Stanford typed dependencies" revealed that a small set of grammatical patterns is sufficient to cover a wide variety of image descriptions (Section 2.3.1.1). The "Abstract Scenes Data-Set" is broad in scope, encompassing descriptions of a multitude of actions, scenes, agents, objects, poses and so forth. It follows that our result should transfer to other domains. We thus provide empirical evidence to support the claim of Kulkarni et al. (2011) that descriptive language only uses a limited range of syntactic patterns.

We also gave evidence for the old question of "what is important in an image." Analysing the descriptive language used in the "Abstract Scenes Data-Set" revealed that a limited vocabulary of nouns, prepositions, adjectives, adverbs, and so forth is sufficient to cover most aspects of how humans describe images (Section 3.1.1). Our work therefore corroborates authors such as Yang et al. (2011), Yao et al. (2010), Li et al. (2011) or Kulkarni et al. (2011) who restrict the target vocabulary of their image description generation systems in order to make the language generation task more tractable.

More importantly, the analysis of our human evaluation experiment (Section 4.3) revealed a discrepancy between automated evaluation scores and the human perspective. BLEU and METEOR gave our system excellent scores (in excess of the inter-annotator agreement!)... and yet humans were still able to tell the difference between our system and the gold-standard. BLEU and METEOR did not pick up the kinds of errors that humans are sensitive to. For example, a description where the main agent (and only the main agent) is incorrect will score highly under the automated evaluation metrics since only one unigram is incorrect. However, such a description is deemed mostly useless by humans. This rift informs two concerns. Firstly, it further highlights the types of image-elements that an automated description system really ought to get right (such as the main subject of an image). Secondly, it spotlights the unsatisfactory state of automatic evaluation in the field.

5.3 Future Work

Our template-based image description model could be improved in a number of ways, both from an engineering and from a conceptual point of view.

Low hanging fruit include engineering fixes for our model's three most recurrent and egregious errors (described in Section 4.2.1). For instance:

- Our model has issues with having an overly strong prior on certain words in specific grammatical functions. This could be addressed by tuning the model's content selector locally rather than globally (that is, optimising the parameters of every base-classifier C_i that generates words for a particular grammatical function F_i individually rather than having one set of parameter values for all C and F). Using base-classifiers more sophisticated than Logistic Regression could additionally help with the prior-problem. For instance, Support Vector Machines (Cortes and Vapnik, 1995) are sometimes used to generate text in the automated image description literature (Gupta et al., 2008b) but were not considered by us due to engineering constraints.

- The model's by-design tendency of not handling collocations correctly (e.g. truncating *baseball ~~glove~~*) could be solved in one of two ways. A pragmatic engineering solution could be as simple as expanding the model's vocabulary with collocations (i.e. *baseball-glove* instead of *baseball glove*). A more theoretically sound alternative could handle collocations by modifying the model's content selection component to generate phrases instead of single words (e.g. see Gupta et al. (2012)'s use of unigrams, bigrams and trigrams to generate novel image descriptions).

- The issue with the model's lack of "world knowledge" could be addressed by using a language model trained on a larger corpus during surface realisation. Krishnamoorthy et al. (2013), for instance, use a language model trained on the Google ngram corpus of more than 500 billion words (Michel et al., 2011; Lin et al., 2012). Alternatively, semantic grounding could be incorporated explicitly by following an approach similar to Yang et al. (2011).

From a more conceptual standpoint, there are two main avenues for future research based on our work:

- A number of authors (Kulkarni et al., 2011; Mitchell et al., 2012), including us (Section 3.2.1), have stressed the importance of being able to adapt the structure of a textual description to the image at hand. Our novel two-phase "template-acquisition followed by template-prediction" methodology should be compared with other approaches in the literature. The Tree Substitution Grammars of Mitchell et al. (2012) would be a good starting point for this study.

- We provided ample background for our claim that the findings entailed by our work on the "Abstract Scenes Data-Set" should be transferable to photo-realistic images (Heider and Simmel, 1944; Oatley and Yuill, 1985; Zitnick and Parikh, 2013). In order to validate the viability of the "Abstract Scenes Data-Set" as

a sandbox for semantics-focused studies of the vision–language interplay, our theoretical arguments should be put to the test. This could be done by developing a model for the "Abstract Scenes Data-Set" (or using the model presented by this report) and transferring the resulting methodology to a different corpus of images and descriptions.

Perhaps most importantly, however, we highlighted the need for better automatic evaluation metrics for image description. A number of authors besides us (Section 4.3) have already stressed the inadequacy of using BLEU to evaluate image descriptions due to the high subjectivity and variability inherent in the task (Kulkarni et al., 2011; Gupta et al., 2012). We gave some evidence that a more syntactically forgiving metric such as METEOR (which incorporates paraphrase and synonym information) could work better than BLEU. Additionally, we pointed out that some parts of an image description (e.g. who is the main agent in the image?) are more important to get right than others because humans are not equally sensitive to all types of errors in image descriptions. A more effective approach might involve an automatic evaluation metric based on METEOR that weights components of descriptions according to their grammatical function.

Appendix A

Sample Realisations of Templates

This appendix lists the 20 most frequent grammatical structures ("templates") that are used as the basis for the descriptions generated by the template model of Section 2.3.1 (Table A.1).

Template	Sample sentences
nsubj,aux,root,det,dobj	Jenny is throwing the ball. Jenny is wearing a crown. Mike is petting the cat.
nsubj,aux,root,prep,det,pobj	Mike is sitting on the ground Mike is hiding behind a tree! Jenny is playing with the cat.
det,nsubj,root,prep,det,pobj	The sun is behind a cloud. The bear is by the tent. The shovel is under the table.
nsubj,aux,root,det,nn,dobj	Mike is wearing a ski cap. Mike is holding a tennis racket. Jenny is wearing a pirate hat.
det,nsubj,aux,root,prep,det,pobj	The helicopter was flying in the sky. A cat is sitting by a pizza. The cat is sitting under the tree.
nsubj,aux,root,dobj	Jenny is holding balloons. Mike is wearing glasses. Mike is chasing Jenny.
nsubj,root,det,dobj	Jenny sees the bear. Jenny dropped the hamburger. Jenny has a soda.
nsubj,aux,root,det,amod,dobj	Mike is wearing a silly hat Mike is wearing a blue shirt Mike is making a silly face.
nsubj,root,prep,det,pobj	Mike is on the slide. Mike fell to the ground. Jenny is near a tent.

expl,root,det,nsubj,prep,det,pobj	There is a helicopter in the sky. There is a cloud in the sky. There is a pie on the table.
nsubj,aux,root,prep,pobj	Jenny is waving to Mike. Jenny is sitting by Mike. Jenny is laughing at Mike.
nsubj,cop,advmod,root	Mike is very happy Jenny is quite sad. Mike is very hungry.
nsubj,cc,conj,aux,root,prep,det,pobj	Mike and Jenny are playing under the sun Mike and Jenny are playing with a basketball. Mike and Jenny are running from the bear.
nsubj,cop,root,prep,det,pobj	Mike is angry with a dog. Jenny is terrified of the owl. Jenny is afraid of the snake.
nsubj,root,det,nn,dobj	Jenny kicked the beach ball. Mike has a tennis racquet. Mike has a gold crown.
nsubj,aux,root,advmod,prep,det,pobj	Jenny is sitting next to the tree Mike is sitting next to the table. Mike is sitting alone in the grass.
nsubj,cc,conj,aux,root,dobj	Jenny and mike are playing soccer Mike and Jenny are holding hands. Mike and Jenny are playing baseball.
det,nsubj,aux,root,prep,pobj	The duck is walking toward Jenny. The cat is sitting beside Jenny. An airplane is flying over head.
det,nsubj,root,prep,pobj	The snake crawls towards Mike. A dog is near jenny. The bear is behind Mike.
nsubjpass,auxpass,root,prep,det,pobj	Jenny is scared of the snake. Jenny is scared of the rain. Jenny is frightened by the bear.

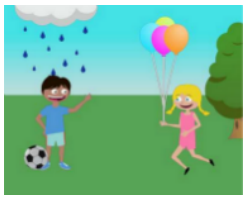
Table A.1: Sample sentences realising the 20 most frequent templates extracted from the “Abstract Scenes Data-Set.”

Appendix B

Instructions for Human Evaluators

In this task you will look at a series of images and image descriptions created by a computer program.

You will be presented with 20 images and 3 descriptions for each image. For each image you will be asked to judge whether the descriptions are relevant to the image. You will do this using a 1-5 rating scale, where 5 is best and 1 is worst. There are no ”correct” answers, so whatever choice seems appropriate to you is a valid response. For example, if you were given the following image and description:



Mike and Jenny are playing ball.

You would probably give the description a high rating (4 or 5) with respect to relevance. As you can see the description captures a relevant aspect of the image. Now, consider the following pair:

Mike is in the sandbox.

Here, the description only marginally relates to the image. Therefore, you would give it a low rating (e.g., 1 or 2) in terms of relevance. Finally, consider the following image and description:

Mike is holding a hamburger.

While the description certainly is relevant to the image, the description maybe doesn't capture the most interesting or exciting aspect of the image. Therefore, you would give it a medium rating (e.g., 3) in terms of relevance.

Appendix C

Sample Model Outputs

This appendix shows images and their automatically generated descriptions. The main purpose of the appendix is to offer further illustration for the arguments proposed in Section 4.2.

Figures C.1–C.2 compare how the template model, keyword baseline and image similarity baseline describe images.

Figures C.3–C.8 show images described by the template model that were described using frequent phrases in the model's vocabulary.

Figure C.9 shows sample images that were grouped together (i.e. deemed similar) by the image similarity baseline's locality sensitive hashing function.

Template Model	mike is sitting in the sandbox	jenny is wearing a silly hat	mike is very sad	mike is wearing a hat	mike is very happy
Keyword Baseline	mike is sitting next to the sandbox	jenny is scared by the bear next to mike	mike and jenny are very sad	jenny is wearing mike 's blue cap	mike and jenny are happy in the park
Image Similarity Baseline	the cat wants to play in the sandbox too	jenny is worried about the snake	mike is close to the fire	jenny dropped her hamburger and she 's sad about it	the dog is standing next to the slide
Human Description	mike and jenny are sitting in the sandbox	jenny is surprised by the snake	it's a rainstorm and jenny runs away to stay dry	mike is wearing a witch hat	jenny and the mike are having fun at the park today

Figure C.1: Sample images that are described adequately by all models.

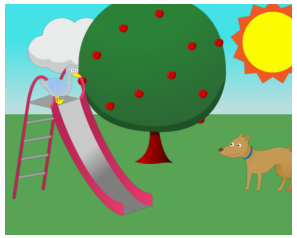
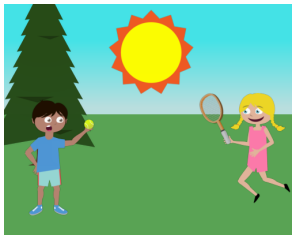
					
Template Model	the sun is in the sky	jenny is flying a kite	mike is holding a tennis ball	jenny is scared of the snake	mike is sitting in the park
Keyword Baseline	the duck is next to the dog	jenny is giving mike balloons and a hot dog	a tennis ball is next to jenny	a green snake is next to mike and jenny	the pink bucket is on the ground next to mike and jenny
Image Similarity Baseline	the apple tree stands in the field	mike is happy because there is pie	mike holds a racket	jenny is jumping because she is excited	mike is frightened by lightning
Human Description	the sun is shining above the brown dog	jenny flies a kite	mike is holding a tennis ball	jenny and mike are afraid of snakes	mike is on the swings

Figure C.2: More example images that are described adequately by all models.

Figure C.3: Template model generations — sample images for descriptions starting with *Mike is holding*.

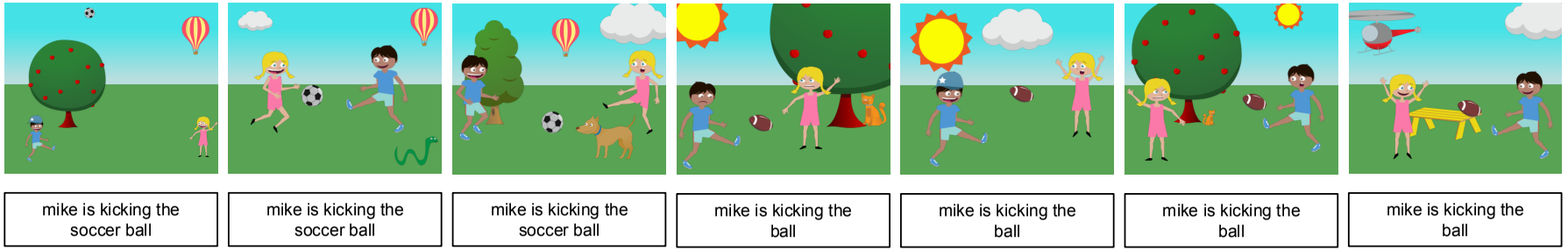


Figure C.4: Template model generations — sample images for descriptions starting with *Mike is kicking*.

Figure C.5: Template model generations — sample images for descriptions starting with *Mike is standing*.

Figure C.6: Template model generations — sample images for descriptions starting with *Mike has a*.

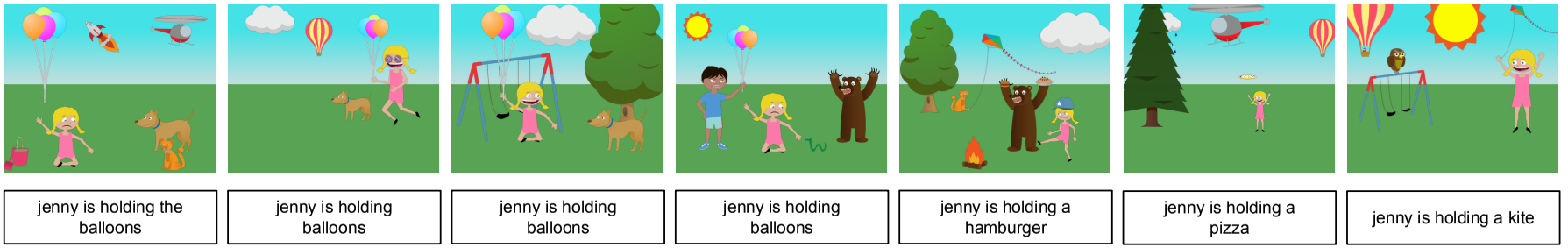


Figure C.7: Template model generations — sample images for descriptions starting with *Jenny is holding*.

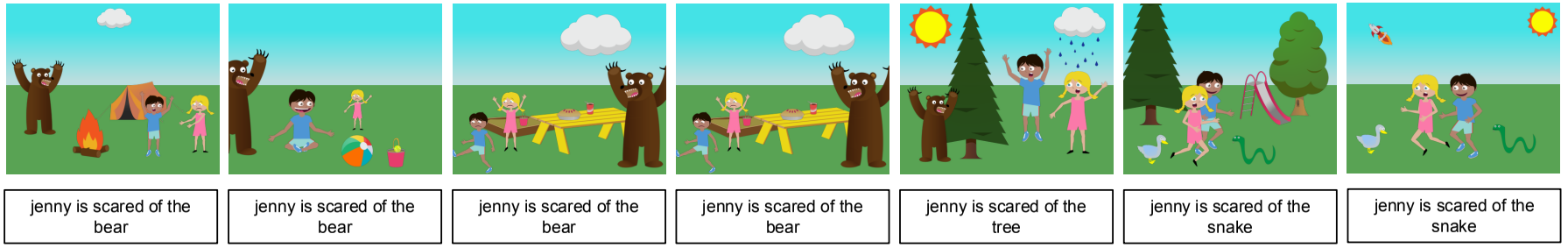


Figure C.8: Template model generations — sample images for descriptions starting with *Jenny is scared*.

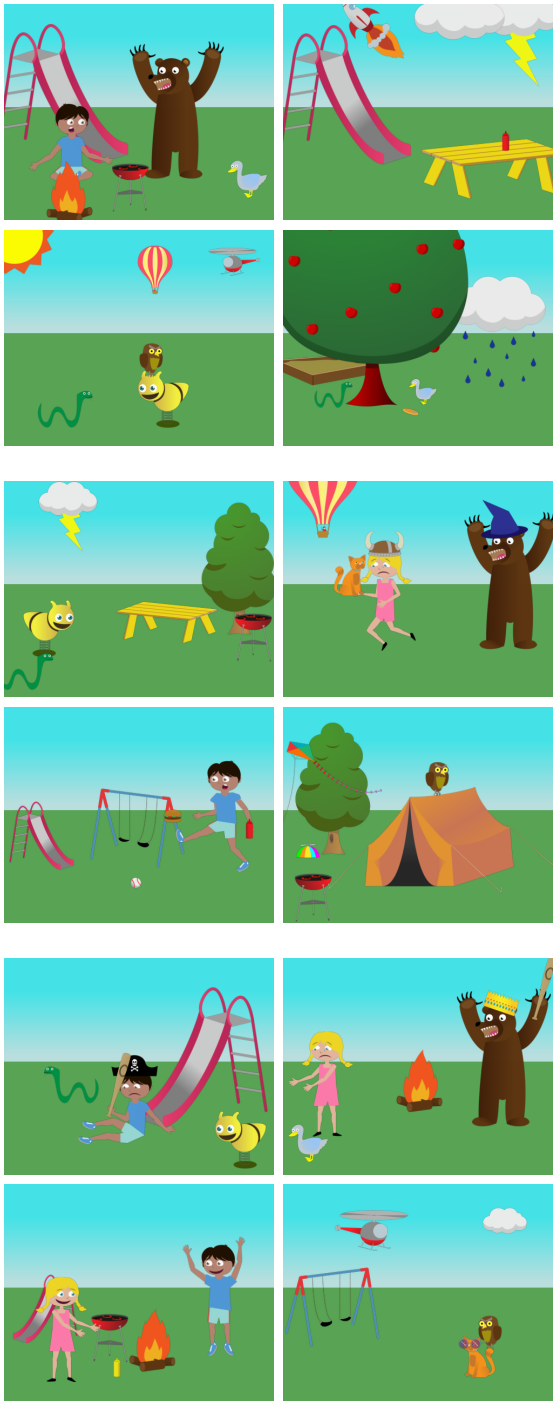


Figure C.9: Sample images taken from three of the locality sensitive hashing buckets used by the image similarity baseline to find similar images.

Bibliography

Ahmet Aker and Robert Gaizauskas. Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1250–1258. Association for Computational Linguistics, 2010.

Linda H Armitage and Peter GB Enser. Analysis of user need in image archives. *Journal of information science*, 23(4):287–299, 1997.

Talis Bachmann. Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology*, 3(1):87–103, 1991.

Kobus Barnard, Pinar Duygulu, and David Forsyth. Clustering art. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–434. IEEE, 2001.

Regina Barzilay and Lillian Lee. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 16–23. Association for Computational Linguistics, 2003.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, pages 1554–1563, 1966.

Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision–ECCV 2010*, pages 663–676. Springer, 2010.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13:27–66, 2012.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised

learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics, 2011.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20 (3):273–297, 1995.

W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.

Doug Cutting. Apache lucene. `http://lucene.apache.org/java/docs/index.html`, 2014. Online; Accessed: 2014-02-30.

Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. `http://nlp.stanford.edu/software/dependencies_manual.pdf`, 2008a.

Marie-Catherine De Marneffe and Christopher D Manning. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics, 2008b.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, 2011.

Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer VisionECCV 2002*, pages 97–112. Springer, 2002.

Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302. Association for Computational Linguistics, 2013.

M Everingham, L Van Gool, C Williams, J Winn, and A Zisserman. The pascal visual object classes challenge 2009. In *2th PASCAL Challenge Workshop*, 2009.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story:

Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010.

Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7(1):10, 2007.

Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

SL Feng, Raghavan Manmatha, and Victor Lavrenko. Multiple bernoulli relevance models for image and video annotation. 2004.

Yansong Feng and Mirella Lapata. Automatic caption generation for news images. 2013.

Leo Ferres, Avi Parush, Shelley Roberts, and Gitte Lindgaard. Helping people with visual impairments gain access to graphical information through natural language: The igraph system. In *Computers Helping People with Special Needs*, pages 1122–1130. Springer, 2006.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 309–316. IEEE, 2009.

Abhinav Gupta and Larry S Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Computer Vision–ECCV 2008*, pages 16–29. Springer, 2008.

Ankush Gupta, Yashaswi Verma, and CV Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012.

Sonal Gupta, Joohyun Kim, Kristen Grauman, and Raymond Mooney. Watch, listen & learn: Co-training on captioned images and videos. In *Machine Learning and Knowledge Discovery in Databases*, pages 457–472. Springer, 2008a.

Sonal Gupta, Joohyun Kim, Kristen Grauman, and Raymond Mooney. Watch, listen & learn: Co-training on captioned images and videos. In *Machine Learning and Knowledge Discovery in Databases*, pages 457–472. Springer, 2008b.

Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243–259, 1944.

David Howell. *Statistical methods for psychology*. Cengage Learning, 2012.

John K Karlof. *Integer programming: theory and practice*. CRC Press, 2005.

Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.

Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, and Sergio Guadarrama. Generating natural-language video descriptions using text-mined knowledge, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, volume 1, page 4, 2012.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating simple image descriptions. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1601–1608. IEEE, 2011.

Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 359–368. Association for Computational Linguistics, 2012.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Irene Langkilde and Kevin Knight. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 704–710. Association for Computational Linguistics, 1998.

Victor Lavrenko, R Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In *NIPS*, volume 1, page 2, 2003.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966.

Stephen C Levinson. Pragmatics (cambridge textbooks in linguistics). 1983.

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

Stan Z Li. *Markov random field modeling in computer vision*. Springer-Verlag New York, Inc., 1995.

Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

Wei-Hao Lin and Alexander Hauptmann. Which thousand words are worth a picture? experiments on video retrieval using a thousand concepts. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 41–44. IEEE, 2006.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174. Association for Computational Linguistics, 2012.

Joseph T. Lizier. Jidt: An information-theoretic toolkit for studying the dynamics of complex systems, 2013.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. Baselines for image annotation. *International Journal of Computer Vision*, 90(1):88–105, 2010.

Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014): 176–182, 2011.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780, 2006.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012.

Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

Hieu V Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Computer Vision–ACCV 2010*, pages 709–720. Springer, 2011.

Keith Oatley and Nicola Yuill. Perception of personal and interpersonal action in a cartoon film. *British Journal of Social Psychology*, 24(2):115–124, 1985.

Aude Oliva and Philippe G Schyns. Diagnostic colors mediate scene recognition. *Cognitive psychology*, 41(2):176–210, 2000.

Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

L Breiman JH Friedman RA Olshen and Charles J Stone. Classification and regression trees. *Wadsworth International Group*, 1984.

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images

using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

Emanuel Parzen et al. On estimation of a probability density function and mode. *Annals of mathematical statistics*, 33(3):1065–1076, 1962.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

Mary C Potter. Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2(5):509, 1976.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.

Irina Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

Murray Rosenblatt et al. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

Andreas Stolcke et al. Srilm—an extensible language modeling toolkit. In *INTERSPEECH*, 2002.

Kristina Toutanova and Christopher D Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics, 2000.

Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3):13, 2008.

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

Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.

Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010.

Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2): 41–75, 2011.

Harry Zhang. The optimality of naive bayes. *A A*, 1(2):3, 2004.

C Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3009–3016. IEEE, 2013.