

# **Stony Brook University**



OFFICIAL COPY

**The official electronic file of this thesis or dissertation is maintained by the University Libraries on behalf of The Graduate School at Stony Brook University.**

**© All Rights Reserved by Author.**

# Composing Image Descriptions in Natural Language

A Dissertation Presented

by

**Polina Kuznetsova**

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

**Doctor of Philosophy**

in

**Computer Science**

Stony Brook University

August 2014

Copyright by  
Polina Kuznetsova  
2014

**Stony Brook University**

The Graduate School

**Polina Kuznetsova**

We, the dissertation committee for the above candidate for the  
Doctor of Philosophy degree, hereby recommend  
acceptance of this dissertation.

**Yejin Choi, Dissertation Advisor**

Assistant Professor, Computer Science Department

**Yevgen Borodin, Chairperson of Defense**

Research Assistant Professor, Computer Science Department

**Paul Fodor, Committee Member**

Research Assistant Professor, Computer Science Department

**Raymond J. Mooney, External Committee Member**

Professor, Computer Science, The University of Texas at Austin

This dissertation is accepted by the Graduate School

Charles Taber

Dean of the Graduate School

Abstract of the Dissertation

**Composing Image Descriptions in Natural Language**

by

**Polina Kuznetsova**

**Doctor of Philosophy**

in

**Computer Science**

Stony Brook University

2014

We study the task of *image description generation*, which can find applications in image search, web accessibility research, story illustration, etc.

Rather than concentrating on precise but robotic descriptions, we aim to generate captions, which are human-like, but which are still relevant to the image content. Human generated text is nontrivial in structure and vocabulary. A purely bottom-up approach, relying only on vision detection vocabulary, would struggle to generate such a description as “A cute squirrel having a feast under a tree”.

To generate descriptions, which are close to human-like in their complexity and richness, we exploit a vast amount of human-written text available on the Internet and use a dataset of images associated with their captions written by users the web-site Flickr. Based on various aspects of the target image, we collect a set of matching images. From the human-written captions of the obtained images we elicit candidate phrases associated with the matching aspects. We selectively glue together extracted phrases into plausible descriptions, using

linguistic patterns and parse tree structure. We tackle this non-trivial task by modeling it as an Integer Linear Programming problem and introducing a novel tree-driven phrase composition framework.

As an optional preprocessing step to the generation process, we introduce the task of *image caption generalization*, the aim of which is to remove extraneous information from image captions written by Flickr users. Evaluation results show that, when using generalized captions as a new source of candidate phrases, we are able to generate descriptions of a better quality in terms of relevance, whilst achieving expressiveness and linguistic sophistication of the resulting output.

## TABLE OF CONTENTS

<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>x</b>
<b>Acknowledgements</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Literature Survey: Image Description Generation in NLP</b>	<b>5</b>
<b>3 Sequence-driven approach to Image Description Generation</b>	<b>13</b>
3.1 Overview . . . . .	13
3.2 Dataset . . . . .	14
3.3 Harvesting Caption Fragments . . . . .	16
3.3.1 Retrieving Noun-phrases (NPs) . . . . .	17
3.3.2 Retrieving Verb-phrases (VPs) . . . . .	18
3.3.3 Retrieving Prepositional-phrases (PPs) . . . . .	19
3.4 ILP for Phrase-based Composition of Image Descriptions . . . . .	21
3.5 Image-level Content Planning . . . . .	23
3.5.1 Variables and Objective Function . . . . .	24
3.5.2 Constraints . . . . .	25
3.5.3 Weight $F_o$ : Object Detection Confidence . . . . .	26
3.5.4 Weight $F_{ot}$ : Ordering and Compatibility . . . . .	26
3.6 Surface Realization – Phrase Composition . . . . .	27
3.6.1 Variables and Objective Function . . . . .	27
3.6.2 Constraints . . . . .	29
3.6.3 Unary Phrase Selection . . . . .	33
3.6.4 Pairwise Phrase Cohesion . . . . .	34
3.6.5 Phrase Score For Beginning/End Of The Sentence . . . . .	35
3.6.6 Cognitive Phrases . . . . .	36
3.7 Evaluation of Sequence-driven Composition Approach . . . . .	37
3.7.1 TestSet . . . . .	37
3.7.2 Baselines, Gold Standard and System Versions . . . . .	37
3.7.3 Automatic Evaluation: . . . . .	38
3.7.4 Human Evaluation I: Multi-Aspect Rating . . . . .	39
3.7.5 Human Evaluation II: Forced Choice . . . . .	39
3.8 Discussion . . . . .	41
<b>4 Dependency-based Sequence-driven Caption Generalization</b>	<b>44</b>
4.1 Extraneous Information . . . . .	44
4.2 Related Work: Sentence Compression . . . . .	46
4.3 Problem Formulation . . . . .	48
4.3.1 Optimization criteria . . . . .	48
4.3.2 DP formulation . . . . .	50

4.4	Enhancing the Task with Dependency Constraints . . . . .	51
4.5	Dynamic Programming + Dependency Constraints + Beam Search . . . . .	54
4.6	Evaluation of Sequence-driven Caption Generalization . . . . .	57
4.6.1	Methods for Compression . . . . .	57
4.6.2	Human-Generalized Captions . . . . .	58
4.6.3	Intrinsic Human Evaluation: Forced Choice . . . . .	58
4.6.4	Extrinsic Evaluation . . . . .	59
4.7	Discussion . . . . .	61
<b>5</b>	<b>CKY-based Tree-driven Caption Generalization</b>	<b>64</b>
5.1	Overview . . . . .	64
5.2	Problem Formulation . . . . .	67
5.2.1	Pruning Case (1): . . . . .	69
5.2.2	Pruning Case (2)/(3): . . . . .	69
5.3	Modelling Compression Criteria . . . . .	70
5.3.1	I. Tree Structure: . . . . .	70
5.3.2	II. Sequence Structure: . . . . .	70
5.3.3	III. Branch Deletion Probabilities: . . . . .	71
5.3.4	IV. Vision Detection (Content Selection): . . . . .	71
5.4	Human Compressed Captions . . . . .	71
5.5	Discussion of the Method . . . . .	72
5.6	Evaluation of Tree-driven Caption Generalization . . . . .	74
5.6.1	Methods for Compression: . . . . .	74
5.6.2	Intrinsic Human Evaluation: Forced Choice . . . . .	74
5.6.3	Extrinsic Evaluation: Image Caption Generation via Phrase-based composition . . . . .	75
5.7	Discussion . . . . .	77
<b>6</b>	<b>Tree-driven approach to Image Description Generation</b>	<b>82</b>
6.1	Overview . . . . .	82
6.2	Harvesting Tree Fragments . . . . .	83
6.3	ILP for Tree-driven Composition of Image Captions . . . . .	84
6.3.1	Variables and Objective Function . . . . .	86
6.3.2	Constraints . . . . .	89
6.3.3	Remarks . . . . .	93
6.4	Evaluation of Tree-driven Composition Approach . . . . .	94
6.4.1	Automatic Evaluation . . . . .	95
6.4.2	Human Evaluation: Forced Choice . . . . .	95
6.5	Discussion . . . . .	97
<b>7</b>	<b>Conclusion</b>	<b>103</b>
7.1	Summary of Contributions . . . . .	103
7.2	Future Research Directions . . . . .	104
7.2.1	Expanding Generation Techniques to Creative Language Generation .	104
7.2.2	Improving Image Descriptions . . . . .	107



7.2.3 Expanding Sentence Compression Algorithm . . . . .	108
<b>A ILP System Variations</b>	<b>109</b>
<b>B Additional Good Examples of Generated Descriptions</b>	<b>110</b>
<b>C Literally not Relevant, but Metaphorically Creative Examples of Generated Descriptions</b>	<b>111</b>
<b>D Examples of Generated Descriptions<sup>1</sup></b>	<b>112</b>
<b>Bibliography</b>	<b>124</b>

---

<sup>1</sup>If images in the figures of the dissertation are marked with numbers, those correspond to indices of images in this Appendix.

## LIST OF FIGURES

1.1	Examples of Human-written Descriptions . . . . .	3
2.1	Previous Work: Examples of Sequential-template Driven Approaches . . . . .	7
2.2	Previous Work: Examples of Tree-template Driven Approaches . . . . .	8
2.3	Previous Work: Usage of Relevant Text . . . . .	10
2.4	Previous work: Mapping Images to Ready Sentences . . . . .	12
3.1	Examples of Object Detections . . . . .	16
3.2	Harvesting Noun Phrases . . . . .	17
3.3	Harvesting Verb Phrases . . . . .	18
3.4	Harvesting Prepositional Phrases from Stuff Matches . . . . .	20
3.5	Harvesting Prepositional Phrases from Scene Matches . . . . .	21
3.6	Images with Multiple Objects . . . . .	24
3.7	Images with Multiple Objects . . . . .	26
3.8	Sentences with the Same Scene Phrase . . . . .	32
3.9	Phrases with the Same Head Word . . . . .	33
3.10	SEQ+LINGRULE & HMM Generated Descriptions, where SEQ+LINGRULE was Preferred . . . . .	40
3.11	Examples with Different Aspects of Problems in the SEQ+LINGRULE Generated Descriptions. . . . .	42
3.12	Examples with Problems in the SEQ+LINGRULE Generated Captions due to Extraneous Information in Image Captions. . . . .	43
4.1	Examples of Human-written Image Captions with Extraneous Information . . . . .	45
4.2	An Example of a Compressed Caption that is More Applicable for Describing New Images . . . . .	45
4.3	Sentence Ngram Probability . . . . .	49
4.4	Sentence Sequence Driven Compression . . . . .	50
4.5	Sentence Sequence Driven Compression with Constraints . . . . .	53
4.6	Sentence Sequence Driven Compression with Constraints by Example . . . . .	54
4.7	Statistics for Words with Future Dependencies . . . . .	55
4.8	Example Image Caption Transfer . . . . .	60
4.9	Good Examples of Generalized Captions . . . . .	61
4.10	Bad Examples of Generalized Captions . . . . .	62
5.1	CKY Parsing Tree . . . . .	65
5.2	CKY Parsing Matrix . . . . .	66
5.3	CKY Compression Tree . . . . .	67
5.4	CKY Compression Matrix . . . . .	68
5.5	Caption Generalization: Good Examples. . . . .	78
5.6	Caption generalization: bad examples . . . . .	79
5.7	Examples of SEQ+LINGRULE Descriptions with Extraneous Information and SEQ+PRUNING Descriptions, for which Extraneous Information was and was not Successfully Removed. . . . .	80

5.8	SEQ+PRUNING Descriptions: Bad Examples. . . . .	81
6.1	Harvesting Phrases for the Target Image Based on Visual Match . . . . .	83
6.2	An Example Scenario of Tree Composition . . . . .	87
6.3	CKY-style Representation of Tree Composition . . . . .	89
6.4	Examples where SEQ+TREE+PRUNING improved captions over SEQ+LINGRULE. . . . .	97
6.5	Examples where SEQ+TREE+PRUNING improved captions over SEQ+PRUNING. . . . .	98
6.6	An example of a description preferred over human gold standard. Image description is improved due to caption generalization. . . . .	99
6.7	SEQ+TREE+PRUNING Descriptions: Examples, where Generalization Helped. . . . .	100
6.8	SEQ+TREE+PRUNING Descriptions: Good Examples. . . . .	101
6.9	Description Generation: Bad Examples. . . . .	102
7.1	Creative and not creative word pairs graph . . . . .	106

## LIST OF TABLES

2.1	Some of the Existing Image Generation Approaches . . . . .	6
3.1	Automatic Evaluation of Sequential ILP . . . . .	39
3.2	Human Evaluation of Sequential ILP: Multi-Aspect Rating . . . . .	39
3.3	Human Evaluation of Sequential ILP: Forced Choice . . . . .	40
4.1	Typed Dependency Constraints for Caption Generalization. . . . .	52
4.2	Intrinsic Human Evaluation of Generalized Captions . . . . .	59
4.3	Image Description Transfer: performance in BLEU and F1 . . . . .	60
5.1	Intrinsic Human Evaluation of Generalized Captions . . . . .	75
5.2	Extrinsic Automatic Evaluation of Generalized Captions . . . . .	76
5.3	Extrinsic Human Evaluation of Generalized Captions . . . . .	76
6.1	Automatic Evaluation of Generated Descriptions . . . . .	95
6.2	Human Evaluation of Generated Descriptions . . . . .	96
A.1	ILP System Variations . . . . .	109

## ACKNOWLEDGEMENTS

I am forever grateful to my advisor, Yejin Choi, for her admirable patience, determination and constructive criticism which guided me through this program. Moreover, her passion for research inspired me for many ideas and made the research process, which has its own frustrating moments, fun and enjoyable.

I would like to extend my greatest thanks to our collaborators on this project, Tamara and Alex Berg and Vicente Ordonez, who continue an amazing research in Computer Vision. Their help was crucial for me as a Natural Language Processing researcher.

My deepest thanks to Professor I.V. Ramakrishnan and Professor Steven Skiena for the good advice when I needed it.

I am very thankful to the students, faculty and staff in the Department of Computer Science for being a great help and an inspiration to me.

I am grateful to my fellow colleagues, Song Feng, Jun S. Kang and Ritwik Banerjee for creating a positive work atmosphere and remaining very good friends.

I would like to offer my special thanks to the dissertation committee members, Raymond J. Mooney, Yevgen Borodin and Paul Fodor for finding time to listen about my research and giving sharp and inspiring comments and suggestions.

## CHAPTER 1

### INTRODUCTION

Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, C can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man?

---

Alan Mathison Turing (1950)

Computing Machinery and Intelligence. *Mind* 49: 433-460.

The connection between visual and linguistic information is one of the fascinating topics occupying researchers minds (e.g., Feng and Lapata (2010e), Krishnamoorthy et al. (2013), Yu and Siskind (2013), Thomason et al. (2014), Yatskar et al. (2013)). Work on the border between Computer Vision and Natural Language Processing has its profound connection with language grounding or, more generally, grounded cognition (Feng and Lapata (2010e), Silberer et al. (2013), Yu and Siskind (2013), Matuszek et al. (2012b), Mooney (2008), Silberer and Lapata (2012), Kant (1998)<sup>1</sup>), making it intriguing from an artificial intelligence (AI) research perspective.

One of the directions in AI research is a behaviour-based one, which argues that having human properties is a necessary condition for a machine to be able to perform human tasks (Arkin (1998), Brooks (1990), Brooks (1985)). From that angle, two modalities, visual and

---

<sup>1</sup>“All our knowledge begins with the senses, proceeds then to the understanding, and ends with reason.” – Immanuel Kant, *Critique of Pure Reason*

linguistic, are essential parts of an intelligent system (Matuszek et al. (2012a), Perzanowski et al. (2001), Hawes et al. (2007)), as an ability to perceive and process visual data, as well as process and produce language, are fundamental qualities of human beings (Chomsky (1968)<sup>2</sup>, Lindberg (1981) ).

We study an example of a multi-modal task, which is very easy for people to perform, yet utterly far from trivial for computers to accomplish: *image description generation* (Yang et al. (2011), Kulkarni et al. (2011), Li et al. (2011), Feng and Lapata (2010a)). This problem is not only interesting from an AI research point of view, but also practically advantageous. It can find its applications in image search<sup>3</sup> (e.g., Kovashka et al. (2012), Parikh and Grauman (2013), Socher et al. (2014), Farhadi et al. (2010)), web accessibility research (e.g., Borodin et al. (2010), Bigham et al. (2008), Bigham et al. (2006)), story illustration (e.g., Joshi et al. (2006), Feng and Lapata (2010c), Aletras and Stevenson (2013)), photo-album organization, etc. (e.g., Pastra et al. (2003)).

Our research is inspired by one of the fundamental AI questions: Can a machine imitate a human being? (Turing (1950), Russell et al. (1996)) With that in mind, we aim to generate descriptions, which are human-like, rather than robotic. Human descriptions are non-trivial in structure and vocabulary. Consider a few examples of human-written image captions taken from the web-site Flickr: “Cute squirrel having a feast under the tree”, “You can see these beautiful hills only in the countryside”, “Spring in a white dress” (Figure 1.1). A pure bottom-up approach relying on the direct vocabulary of Computer Vision Recognition methods would struggle to generate such descriptions. Modifiers and verbs, such as “beautiful” and “having a feast”, as well as poetic expressions, such as “Spring in a white dress”, are problematic for Computer Vision techniques.

Thus, we have decided to exploit a vast amount of human-written text available on the Internet. We use a dataset of images associated with their captions extracted from the

---

<sup>2</sup>“We must postulate an innate structure that is rich enough to account for the disparity between experience and knowledge, one that can account for the construction of the empirically justified generative grammars within the given limitations of time and access to data.” – Noam Chomsky, *Language and Mind*

<sup>3</sup>Flickr, Google, Yahoo!



**Figure 1.1:** Examples of Human-written Descriptions

web-site Flickr (Ordonez et al. (2011)). It allows us to generate descriptions for images with no immediate accompanying text available. Based on various aspects of the target image, such as objects, actions, stuff and scene, we collect a set of matching images. From the human-written captions of the obtained images we elicit phrases associated with the matching aspects. We then glue together these bits of the text into descriptions, that have the kind of complexities and richness that are typically present in people’s casual language use. In order to obtain fluent descriptions, we use language statistics extracted from a vast amount of web data (Brants and Franz., 2006). The task of constructing a sentence from pieces of text is not a trivial one and involves a complex set of operations, such as phrase selection and reordering. We tackle this task by modeling it as a constrained optimization problem, in particular, an Integer Linear Programming (ILP) problem. ILP was successfully used in a number of previous work, such as summarization (Clarke and Lapata (2006), Martins and Smith (2009), Woodsend and Lapata (2010)). We use two versions of our system. The first one is sequence-driven, described in Chapter 3. The second one is a novel parse tree-driven approach and is able to generate grammatically more plausible descriptions (Chapter 6).

Whilst generating human-like descriptions, we also aim to generate descriptions, relevant to the target image. Harvesting bits of text from the ready human-written captions has a drawback in that users of Flickr tend to include a lot of extraneous information into their



descriptions. For example, “during my vacation” and “in Florida” are less likely to be transferable to another image. Thus, we introduce a task of *image caption generalization*, described in Chapters 4 and 5. The goal of this task is to remove extraneous information from the image caption. We model this problem as a sentence compression task.

## CHAPTER 2

### LITERATURE SURVEY: IMAGE DESCRIPTION GENERATION IN NLP

Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers

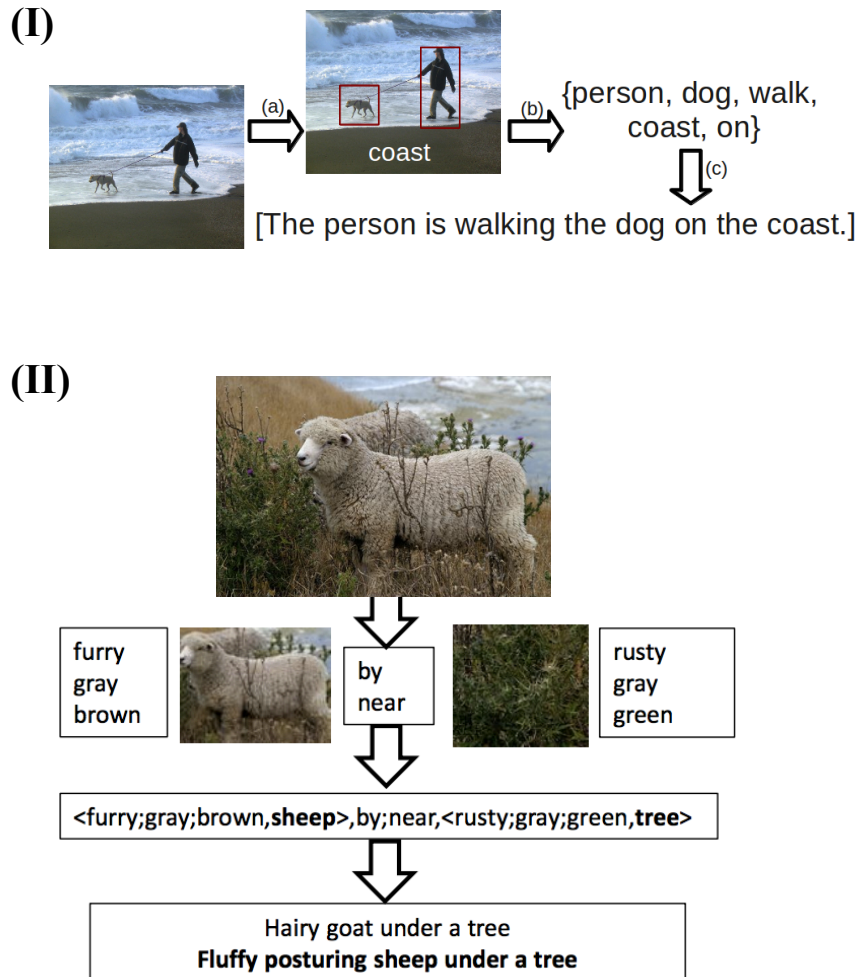
---

D. Stork (HAL's Legacy, 2001) on A. Rosenfeld's vision

There has been a recent spike in efforts to automatically describe visual content in natural language (e.g., Yang et al. (2011), Kulkarni et al. (2011), Li et al. (2011), Krishnamoorthy et al. (2013), Elliott and Keller (2013), Li et al. (2011), Yu and Siskind (2013), Socher et al. (2014)), whether it is images (e.g., Feng and Lapata (2013), Farhadi et al. (2010)) or videos (e.g., Thomason et al. (2014), Guadarrama et al. (2013)) that need to be described. This reflects the understanding that encoding the complexities and subtleties of visual content often requires more expressive language constructs than a set of tags (Leong et al. (2010), Feng and Lapata (2008), Feng and Lapata (2010d), Barnard et al. (2003)). We focus on the image description generation, some existing methods for which are summarized in Table 2.1.

Approach	Tree/Sequence-driven	Template-based	Re-usage of human-written text	Altering human-written text	Meaning space/ image-similarity mapping
Kulkarni et al. (2011)	Seq	+	-	-	-
Yang et al. (2011)	Seq	+	-	-	-
Elliott and Keller (2013)	Seq	+	-	-	-
Yao et al. (2010)	Tree	+	-	-	-
Mitchell et al. (2012)	Tree	+	-	-	-
Feng and Lapata (2013)	n/a	-	+	+	-
Aker and Gaizauskas (2010)	n/a	-	+	+	-
Ordonez et al. (2011)	n/a	-	+	-	+
Mason (2013)	n/a	-	+	+	+
Farhadi et al. (2010)	n/a	-	+	-	+
Socher et al. (2014)	n/a	-	+	-	+
Our – Kuznetsova et al. (2012)	Seq	-	+	+	+
Our – Kuznetsova et al. (2014)	Tree	-	+	+	+

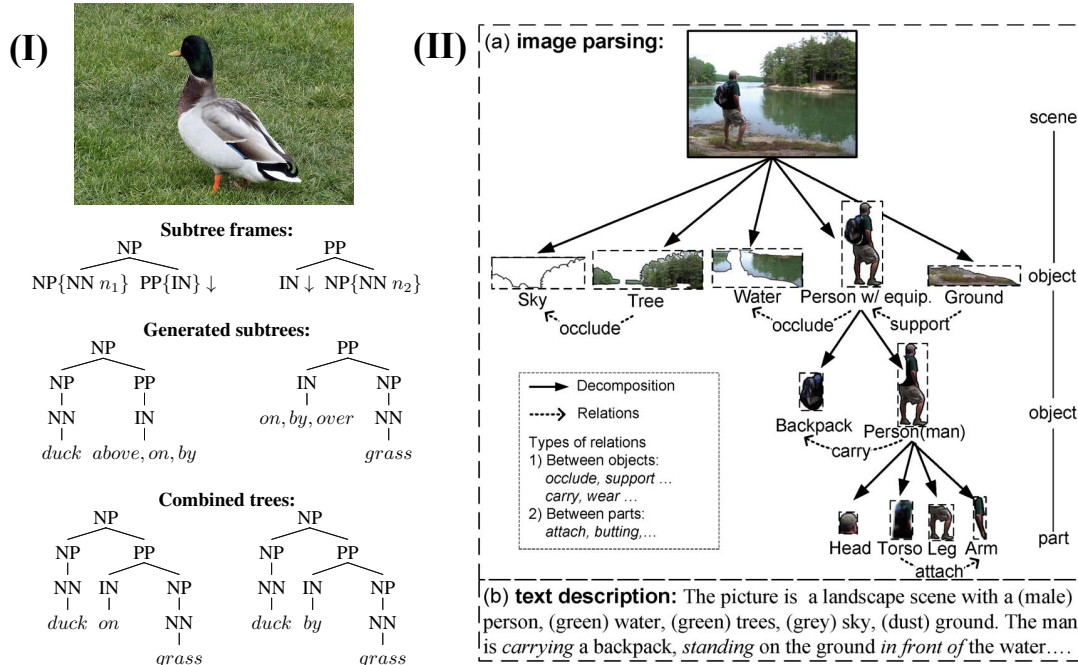
**Table 2.1:** Some of the Existing Image Generation Approaches. Tree/Sequence-driven characteristics are applicable to composition based (mainly template-based) approaches only.



**Figure 2.1:** Previous Work: Examples of Sequential-template Driven Approaches. (I) Yang et al. (2011) and (II) Li et al. (2011).

There have been two main complementary directions explored for automatic image captioning. The first one concentrates on image understanding and precise elaborate captions. The second one, on the other hand, explores rather complex linguistic structures often containing patterns beyond computer vision recognition output. Examples of work from the both directions are given below.

**Image understanding** These approaches focus on describing exactly those items (e.g., objects, attributes) that are detected by vision recognition, which subsequently confines *what* should be described and *how* (e.g., Yao et al. (2010), Kulkarni et al. (2011),



**Figure 2.2:** Previous Work: Examples of Tree-template Driven Approaches. (I) Mitchell et al. (2012) and (II) Yao et al. (2010).

Kojima et al. (2002)).

Some studies focus on flat template-based generation (e.g., Yang et al. (2011), Kulkarni et al. (2011), Li et al. (2011), Elliott and Keller (2013)) and fill templates with text representing vision detections. Figure 2.1 shows examples of such methods.

There are previous approaches which also exploit complex structures like parse trees (Mitchell et al. (2012), Yao et al. (2010)). These approaches however are still limited in the structures they can generate lacking human expressiveness of the captions. Figure 2.2 shows examples of such methods.

Approaches in this direction could be ideal for various practical applications such as image descriptions for the visually impaired. However, it is not clear whether the semantic expressiveness of these approaches can eventually scale up to that of casual but the highly expressive language people naturally use in their online activities. The key technical bottleneck is that the range of describable content (i.e., objects, attributes, actions) is ultimately confined by

the set of items that can be reliably recognized by state-of-the-art vision techniques. Some researchers use an additional information, such as language statistics, to hallucinate descriptive words beyond object and action recognitions (Mitchell et al. (2012), Yang et al. (2011), Elliott and Keller (2013), Li et al. (2011)), whilst still being limited in the sentence structure.


**Enlivening descriptions** Our work contributes to a complementary research avenue, which aims to generate expressive and elaborate descriptions, which are less constrained by visual recognition output or hard-coded templates (e.g., Mason (2013), Feng and Lapata (2013)). Today’s digital world provides researchers with a precious source of readily available human-written text, from which one can distill information needed to generate text similar to human-like in its complexity and sophistication. There has been a lot of work done, in which researchers reuse parts of the text, whether it is through summarization of accompanying text (Feng and Lapata (2010a), Aker and Gaizauskas (2010)) or mapping ready sentences and images to the same meaning space, based on images properties (Ordonez et al. (2011), Mason (2013)) or both, image and linguistic, characteristics (e.g. Farhadi et al. (2010), Socher et al. (2014)).

In these approaches, *the set of what can be described* can be substantially larger than *the set of what can be recognized*, where the former is shaped and defined by the data, rather than by humans. This allows the resulting descriptions to be substantially more expressive, elaborate, and interesting than what would be possible in a purely bottom-up manner.


Feng and Lapata (2010b) presented an approach to generation of descriptions for the images from news articles. Aker and Gaizauskas (2010) and Aker and Gaizauskas (2008) generated caption for geo-tagged images by summarizing multiple documents related to the location depicted in the images. These approaches explore the accompanying text available for an image (Figure 2.3).

Our work is closely related to that one of Ordonez et al. (2011) (Figure 2.4(I)). Their approach is to transfer a human-written caption from other images visually similar to the target image. Approaches similar to Ordonez et al. (2011) are represented in Figure 2.4.

(I)




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



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

(II)



The City of London has St Pauls, but Westminster Abbey is the centrepiece to the City of Westminster. Westminster Abbey should be at the top of any London traveler's list. Westminster Abbey, however, lacks the clear lines of a Rayonnant church,... I loved Westminster Abbey on my trip to London. **Westminster Abbey was rebuilt after 1245 by Henry III's order, and in 1258 the remodeling of the east end of St. Paul's Cathedral began.** He was interred in Westminster Abbey. From 1674 to 1678...

**Figure 2.3:** Previous Work: Usage of Relevant Text. (I) Feng and Lapata (2010b) and (II) Aker and Gaizauskas (2008).

The main idea among all of them is to map images and text to the same meaning space and extract the matching ready descriptions. Some of these approaches additionally modify extracted descriptions to obtain a more relevant result, for example via compression (Mason, 2013).

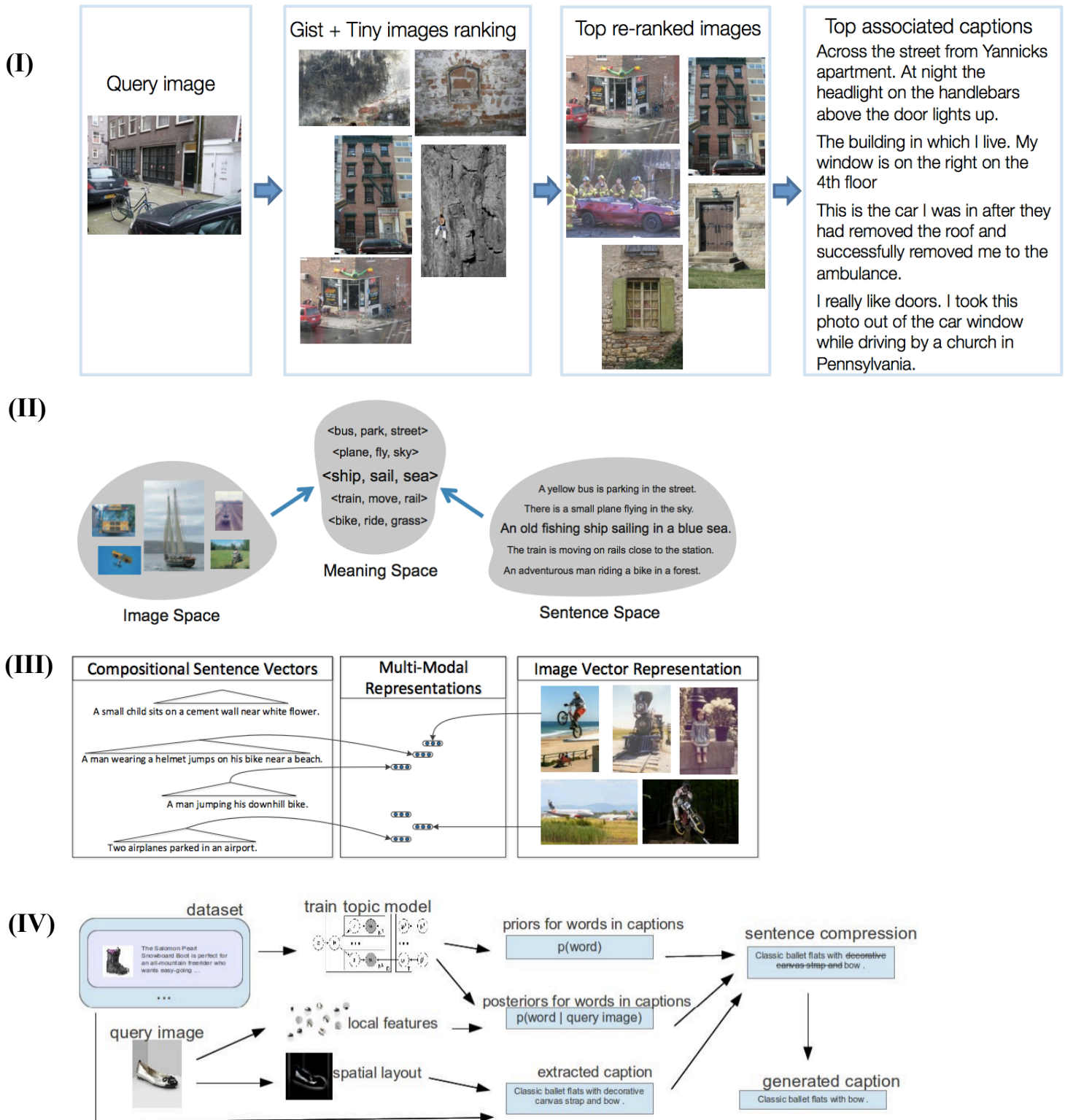
However there is no guarantee that there always exists such a readily available caption which describes all aspects of the query image. We retrieve images similar to the query image in four various aspects: objects, actions, stuff<sup>1</sup> and scene (Kuznetsova et al., 2012). From the captions of retrieved images we extract noun phrases (NP) for objects, verb phrases (VP) for actions and prepositional phrases (PP) for stuff and scene. We then combine those phrases

<sup>1</sup>Stuff is an extended object, usually a mass noun, such as "water" or "grass"

into a description via sequence-driven (Chapter 3) or tree-driven (Chapter 6) approaches.

Methods, which rely on computer vision output usually have to deal with noisy data, such as false object detections. Some existing work deals with noisy vision output via language statistics (e.g. Yang et al. (2011), Kulkarni et al. (2011)). Some study non-visual phrases, which are less likely to be present in the caption (Dodge et al. (2012)). Others explore labels, which are more likely to be used by people when they describe objects in the image (Ordonez et al. (2013)). All these can be very helpful when generating image descriptions. In Chapters 4 and 5 we introduce the *image caption generalization task* (Kuznetsova et al., 2013b) to filter noise in the existing captions. The aim of this task is to make the captions more applicable to the *image description generation task*.





**Figure 2.4:** Previous work: Mapping Images to Ready Sentences. (I) Ordonez et al. (2011), (II) Farhadi et al. (2010), (III) Socher et al. (2014) and (IV) Mason (2013).

CHAPTER 3  
SEQUENCE-DRIVEN APPROACH TO IMAGE DESCRIPTION  
GENERATION

### 3.1 Overview

We propose a holistic data-driven approach that combines and extends the best aspects of the previous approaches – a) using visual recognition to directly predict individual image content elements, and b) using retrieval from existing human-composed descriptions to generate natural, creative, and interesting captions (Kuznetsova et al. (2012), Kuznetsova et al. (2014)). We also lift the restriction of retrieving existing *whole descriptions* (Ordonez et al., 2011) by gathering *visually relevant phrases* which we combine to produce novel and query-image specific descriptions. By judiciously exploiting the correspondence between image content elements and phrases, it is possible to generate natural language descriptions that are substantially richer in content and more linguistically interesting than previous work.

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

Another important thrust of this work is collective *image-level content-planning*, integrating saliency, content relations, and discourse structure based on statistics drawn from a large image-text parallel corpus. This contrasts with previous approaches that generate multiple sentences without considering discourse flow or redundancy (e.g., Li et al. (2011)). For example, for an image showing a flock of birds, generating a large number of sentences stating

the relative position of each bird is probably not useful.

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

Our system consists of two parts. For a query image, we first retrieve candidate descriptive phrases from a large image-caption database using measures of visual similarity (§3.3). We then generate a coherent description from these candidates using ILP formulations for content planning (§3.5) and surface realization (§3.6).

## 3.2 Dataset

We use the SBU Captioned Photo Dataset (Ordonez et al., 2011) for our retrieval database. This dataset contains 1 million images (thus, we also call it 1M image caption corpus) with user associated captions, collected *in the wild* by intelligent filtering of a huge number of Flickr photos. Past work has made use of this dataset to retrieve whole captions for association with a query image (Ordonez et al., 2011). Their method first used global image descriptors to retrieve an initial matched set, and then applied more local estimates of content to re-rank this (relatively small) set (Ordonez et al., 2011). This means that content based matching was relatively constrained by the bottleneck of global descriptors, and local content (e.g. objects) had relatively small effect on performance accuracy.

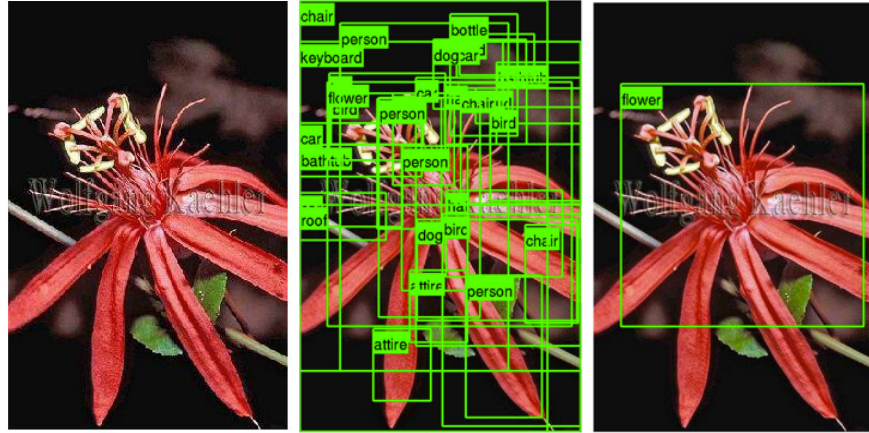
In contrast, we would like to directly access similar local image content (e.g. visually similar objects or similar object-background relationships). Therefore, we perform a very large amount of careful image processing on the entire million image database to extract useful

and accurate estimates of local content. We also want to retrieve bits of caption rather than being constrained to whole caption matching, so we parse each caption into its constituent phrases. This enables us to obtain a pool of meaningful content and caption entities on which to perform retrieval. Processing consists of 4 different types: caption processing into phrases, object detection, rough image parsing to obtain background elements, and scene classification.

**Caption Parsing:** Since our objective is to transfer individual phrases/constituents from a database caption to a query image we use the state of the art Berkeley PCFG parser (Petrov et al., 2006; Petrov and Klein, 2007) to obtain a hierarchical parse tree for each caption. From this tree we can recover individually transferable phrases – e.g. noun-phrases (NPs), verb-phrases, and prepositional-phrases (PPs).

**Object detection:** The first kind of image content we extract is object detections. Here care must be taken because running tens or hundreds of object detectors on an image produces extremely noisy results (e.g. Fig 3.1 centre). Instead, we propose a way to impose intelligent priors on image content – by only running detectors for objects (or their synonyms and hyponyms, e.g. Chihuahua for dog) mentioned in the caption associated with a database image. This produces *much cleaner results* (e.g. Fig 3.1 right). As our detectors we use standard state of the art deformable part-based models (Felzenszwalb et al., ) for 89 common object categories, including: the original 20 objects from Pascal (Everingham et al., 2010), 49 objects from Object Bank (Li et al., 2010), and 20 from Im2Text (Ordonez et al., 2011). Making use of results from our entire million photo database we obtain a large pool of (up to 20k) highly confident object detections for each object category.

**Image parsing:** We use coarse image parsing to estimate background elements in each database image. Six possible background (stuff) categories are considered: sky, water, grass, road, tree, and building. For this we use the stuff detectors from Im2Text (Ordonez et al., 2011) which use color, texton, HoG (Dalal and Triggs, 2005) and Geometric Context (Hoiem et al., 2005) as input features to an SVM classifier that scores all regions in the image using



Ecuador, amazon basin, near coca,  
rain forest, passion fruit flower

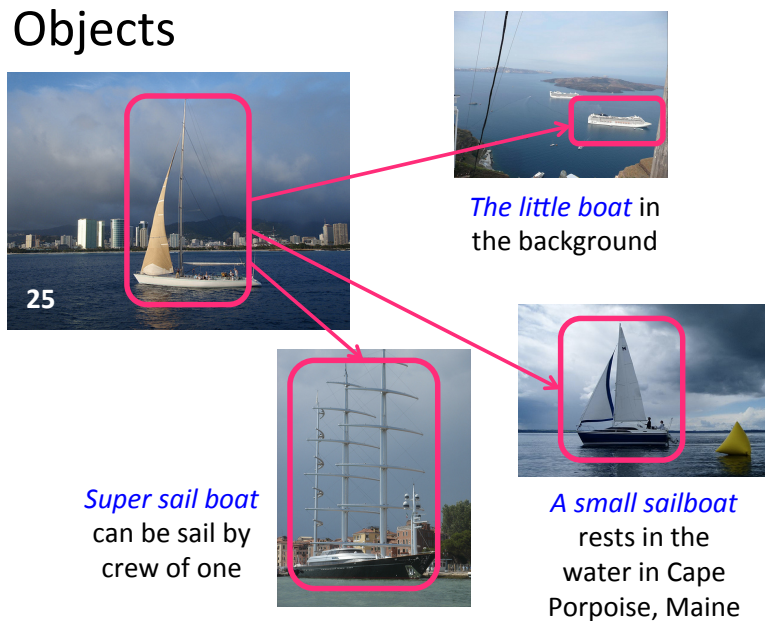
**Figure 3.1:** Examples of Object Detections

a sliding window. These detectors are run on all images in the database, creating a large pool of background elements for retrieval.

**Scene Classification:** We obtain scene descriptors for each image by computing scene classification scores for 26 common scene categories, using the features, classification method and training data from the SUN dataset (Xiao et al., 2010). Because we use a descriptor composed of a number of scene categories, it is useful for capturing and matching overall global scene appearance for a wide range of scene types. Scene descriptors are computed on approximately 700,000 images from the database to obtain a large pool of scene descriptors from which to retrieve matches.

### 3.3 Harvesting Caption Fragments

Overall, for a query image, we would like to visually retrieve relevant phrases of several types: noun-phrases (NPs), verb-phrases (VPs), and prepositional-phrases (PPs). Both local similarity (objects and background elements) and global similarity (overall scene) will be used for retrieval. Several different kinds of visual features will be used as descriptors for measuring visual similarity:



**Figure 3.2:** Harvesting Noun Phrases

**Color:** A color histogram in LAB color space.

**Texture:** A bag-of-words histogram of vector quantized responses to a bank of filters at different scales and orientations (Leung and J., 1999a).

**SIFT Shape:** A bag-of-words histogram of vector quantized SIFT descriptors (Lowe, 2004a) computed in a dense grid.

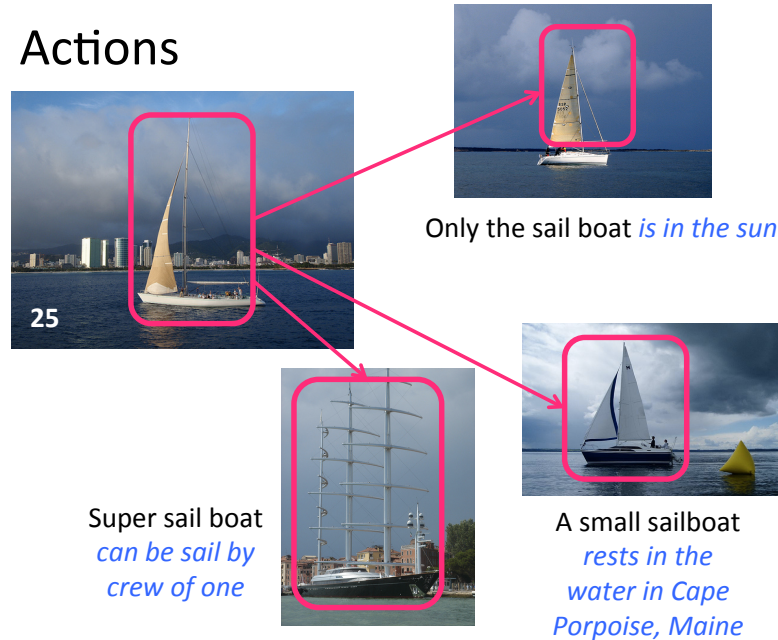
**HoG Shape:** A bag-of-words histogram of vector quantized HoG (Dalal and Triggs, 2005) descriptors computed densely.

**Scene:** A vector of classifier scores for 26 common scene categories (described in Sec 3.2).

The first 4 features are computed locally within an (object or background element) region of interest. The last feature is computed globally on an entire image.

### 3.3.1 Retrieving Noun-phrases (NPs)

For each proposed object detection in a query image, we retrieve a set of relevant noun-phrases from the database. For example, if a “boat” is detected in a query image, then we



**Figure 3.3:** Harvesting Verb Phrases

retrieve NPs from database image captions with visually similar “boat” detections. This process is illustrated in Fig 3.2, where a query image “boat” detection is matched to visually similar “boat” detections (and thus to their referring NPs) from the data base. Note that matches can include phrases referring to synonyms or holonyms of a detection category (e.g. “yacht”, “ship”, etc., for the boat category). Visual similarity for NPs is computed as a combination of color, textron, SIFT, and HoG similarity with equal weights for each feature. Other feature weightings were evaluated on a held out evaluation set, but although some features were found to be better for a few categories (e.g. color for fruit), overall there was no clear advantage to using unequal weights. We find that this usually produces visually similar and conceptually relevant NPs for a query object.

### 3.3.2 Retrieving Verb-phrases (VPs)

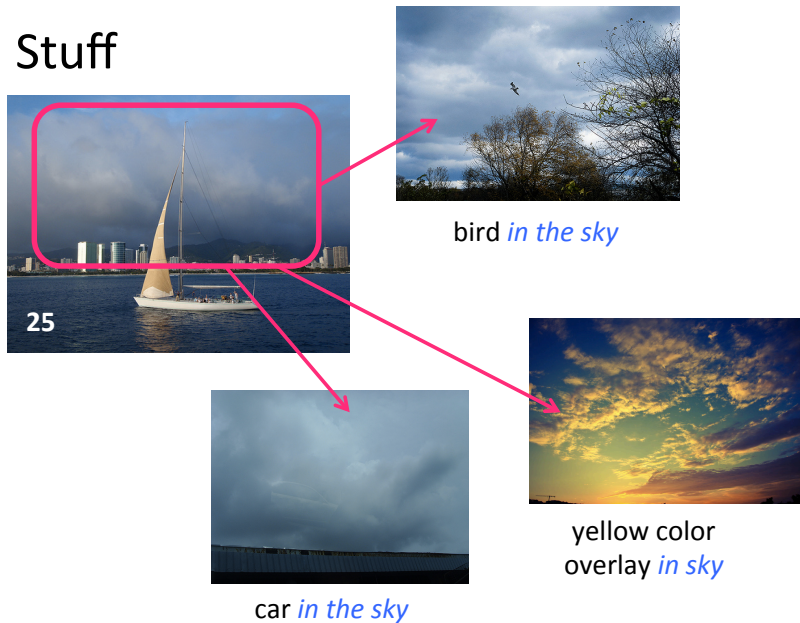
In a similar manner as the previous section, for each proposed object detection in a query image, we retrieve a set of relevant verb-phrases from the database. Here we associate VPs in

database captions to object detections in their corresponding database images. To associate VPs with the appropriate detections, we only annotate a detection with a VP if the sentence also contains an NP where the head word is the detection category (or a synonym/hyponym). Consider an example from Fig 3.3. The top right database caption reads “Only the sail boat is in the sun.”. The Berkeley parser parses this into an NP – “Only the sail boat” – and a VP – “is in the sun.” Therefore we associate this VP with the “boat” detection in the top right database image for matching and retrieval. Our measure of visual similarity is based on an equally weighted linear combination of cosine similarities of color, textron, SIFT and HoG features. As demonstrated in Fig 3.3, this measure often captures similarity in pose between the query and matched objects. Notice, that we do not crop verb phrase into a shorter one, since we don’t know which information we should remove. Sometime a verb phrase contains a PP, which would overlap with a stuff PP, described in Section 3.3.3 (“in the water”, “in the grass”). However, we cannot make a general rule to remove all PPs from VP, as we can distort examples, such as “is in the sun” (Figure 3.3), when no stuff related to “sun” was detected. Another examples include but not limited by “attracted to colorful flowers”, “posing for a photo”, “staring at me”, “munching on a fish in Sua”, etc. The problem of extraneous information is tackled separately and described in Chapters 4 and 5.

### 3.3.3 Retrieving Prepositional-phrases (PPs)

We retrieve two kinds of prepositional-phrases: 1) prepositional phrases referring to the relationship between and object and background/stuff elements, and 2) prepositional phrases referring to the overall setting or scene. **Image parsing-based PPs:** For each proposed object detection and background element detection in a query image, we retrieve relevant PPs according to visual and spatial relationship similarity (illustrated in Fig 3.4 for a query “sky” detection). Here visual similarity between a background query region and background database regions is computed based on color, textron, and SIFT co-sine similarity. Spatial relationship similarity is computed based on the similarity in geometric configuration between





**Figure 3.4:** Harvesting Prepositional Phrases from Stuff Matches

the query object-background pair and object-background pairs observed in the database (where the object in the database pairs need not be the same object as the query). Geometric similarity is measured as a combination of: 1) Vector between the centre of the object detection and the centre of mass of the stuff region, 2) Size of the intersection between the image region and the object detection, and 3) Absolute vertical position of the object in the image, all normalized appropriately by image size. Here visual similarity and geometric similarity measures are given equal weights and produce visually appealing results (Fig 3.4).

**Scene-based PPs:** For a query image, we also retrieve PPs based on our global image scene descriptors by retrieving PPs from database captions with the most similar scene descriptor vectors (illustrated in Fig 3.5). Of all of our phrase matches, these are probably the least reliable since unconstrained prepositional phrases within a sentence could refer to a wide variety of things. However, we still sometimes retrieve useful information about a query image in the matched phrases, corresponding to places (e.g. “Fourviere”, “Paris”), or general scenes (e.g. “across the street”, “in front of our beach house”, “in the ocean”). Using these PPs also encourages our compositions to sound more human because they provide a



**Figure 3.5:** Harvesting Prepositional Phrases from Scene Matches

backstory or context to what could otherwise be a rather boring caption (e.g. “an old man walking in the old town”).

### 3.4 ILP for Phrase-based Composition of Image

#### Descriptions

We described image caption generation task in section 2 and outline of our approach (phrase-based composition of captions) in the beginning of section 3. We retrieve four types of phrases from images similar to a target image: NP for objects, VP for actions, PP for stuff and scene.

For each image we use 10 phrases per phrase type. The goal is to select one phrase of each type and combine selected phrases into a plausible description, which is relevant to image content.

There are images with multiple objects in them. For those we generate a sentence for each object. We also use ILP to select which objects we want to describe and in which order. We

do this step of generation because images can contain many objects and describing each one of them would result in overloaded descriptions.

Although not directly focused on image description generation, some previous work in the realm of summarization shares the similar problem of content planning and surface realization. There are subtle, but important differences however. First, sentence compression is hardly the goal of image description generation, as human-written descriptions are not necessarily succinct.<sup>1</sup> Second, unlike summarization, we are not given with a set of coherent text snippet to begin with, and the level of noise coming from the visual recognition errors is much higher than that of starting with clean text. As a result, choosing an additional phrase in the image description is much riskier than it is in summarization.

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

Sentence fusion has been studied mostly for multi-document summarization (e.g., Barzilay and McKeown (2005)), where redundancy across multiple sentences serves as a guideline for syntactic and semantic validity of generation. In contrast, in our work, we do not have natural redundancy to rely upon, therefore demanding a composition algorithm that collectively models both the tree and the sequence structure of generation.

---

<sup>1</sup>On a related note, the notion of saliency also differs in that human-written captions often digress on details that might be tangential to the visible content of the image. E.g., “This is a dress *my mom made.*”, where the picture does not show a woman making the dress.

## Overview of ILP Formulation

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

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

The ILP formulation of Section 3.5 addresses T1 & T2, i.e., content-planning, and the ILP of Section 3.6 addresses T3 & T4, i.e., surface realization.<sup>2</sup>

### 3.5 Image-level Content Planning

First we describe image-level content planning, i.e., *abstract generation*. For each image we will produce a set of objects corresponding to future sentences. Each object is of a particular type ('person', 'bird', 'car', etc.) and defines the subject for a sentence. The goals of the image-level content planning are to (1) select a subset of the objects based on saliency and semantically compatibility, and (2) order the selected objects based on their content relations.

Figure 3.6 shows examples of descriptions generated for an image without content planning. Description **(a)** contains a sentence for each object, while description **(b)** is a result of content planning. The final description mentions a table and chairs. We can see that table comes first in the description as by order statistics most people would mention table first and then

---

<sup>2</sup>It is possible to create one conjoined ILP formulation to address all four operations T1—T4 at once. For computational and implementation efficiency however, we opt for the two-step approach.



**(a)** Leather chairs surrounded by cookbooks in my building. A high chair in the building. The chair missing her brother in my building. The table frightened vance in the back.  
**(b)** The table frightened vance in the back. Leather chairs surrounded by cookbooks in my building.

**Figure 3.6:** Images with multiple objects without **(a)** and with **(b)** content planning.

chairs. The final description also does not contain descriptions of all chair. Indeed a person would not describe a single chair one by one.

### 3.5.1 Variables and Objective Function

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

$$\alpha_{ok} = \begin{cases} 1, & \text{if object } o \text{ is selected} \\ & \text{for position } k \\ 0, & \text{otherwise} \end{cases} \quad (3.1)$$

where  $k = [0, D)$  ( $k$  is an integer) encodes the position (order) of the selected objects,  $D$  is total number of detected objects and  $o$  indexes one of the objects. In addition, we define a set of variables indicating specific pairs of adjacent objects  $o$  and  $t$ :

$$\alpha_{otk} = \begin{cases} 1, & \text{if } \alpha_{ok} = \alpha_{t(k+1)} = 1 \\ 0, & \text{otherwise} \end{cases} \quad (3.2)$$

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

$$F = \sum_o F_o \cdot \sum_{k=0}^{D-1} \alpha_{ok} + \sum_{ot} F_{ot} \cdot \sum_{k=0}^{S-2} \alpha_{otk} \quad (3.3)$$

where  $F_o$  quantifies the salience/confidence of the object  $o$ , and  $F_{ot}$  quantifies the semantic compatibility<sup>3</sup> between the objects  $o$  and  $t$ . These coefficients (weights) will be described

<sup>3</sup>we use a negative value of the score, similar to surface realization explained further in Section 3.6

in Section 3.5.3 and Section 3.5.4. We use IBM CPLEX (ILOG, Inc, 2006) to optimize this objective function subject to the constraints introduced next in Section 3.5.2.

## 3.5.2 Constraints

We consider *consistency* and *discourse* constraints.

**Consistency Constraints:** We enforce consistency between indicator variables for individual objects (Eq. 3.1) and consecutive objects (Eq. 3.2) so that  $\alpha_{otk} = 1$  iff  $\alpha_{ok} = 1$  and  $\alpha_{t(k+1)} = 1$ :

$$\forall_{otk}, \alpha_{otk} \leq \alpha_{ok} \tag{3.4}$$

$$\alpha_{otk} \leq \alpha_{t(k+1)} \tag{3.5}$$

$$\alpha_{otk} + (1 - \alpha_{ok}) + (1 - \alpha_{t(k+1)}) \geq 1 \tag{3.6}$$

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

$$\sum_o \alpha_{o1} = 1 \tag{3.7}$$

To enforce contiguous positions be selected:

$$\forall k = 2, \dots, S - 1, \sum_o \alpha_{o(k+1)} \leq \sum_o \alpha_{ok} \tag{3.8}$$

### Discourse constraints:

Sometimes content planning is not enough. It still leaves many objects. For example frequency for flower appearing after flower is 1925 (maximum and mean frequencies are 9598 and 62.4 respectively). When we mention objects of the same category many times, the description looks machine-like. Figure 3.7 shows a picture with many flowers. If we describe each detected flower, we get description that is too long (description **(a)**).



**(a)** A blue flower found in a conservation zone in a local cemetery at Gurre Hegn. A pink flower growing together in the center of a big green bush in Tucson Arizona over green grass.  
 Very small blue flower found in a conservation zone in a local cemetery over green grass. A pink flower found in a conservation zone in a local cemetery over green grass.  
**(b)** A blue flower found in a conservation zone in a local cemetery at Gurre Hegn. A pink flower growing together in the center of a big green bush in Tucson Arizona over green grass.

**Figure 3.7:** Images with multiple objects. Description **(b)** shows a caption generated with “at most 2 objects of the same category” constraints

To avoid spurious descriptions, we allow at most two objects of the same category<sup>4</sup> (description **(b)** in Figure 3.7), where  $c_o$  is the category for an object  $o$ :

$$\forall c \in objCategories, \quad \sum_{o \in \{i: c_i=c\}} \sum_{k=1}^S \alpha_{ok} \leq 2 \quad (3.9)$$

If we apply this constraint we get description **(b)** at figure 3.7. The description looks much better. However there are still ways to improve it. For example we can transform multiple objects of the same category into plural form. We leave it as future work.

### 3.5.3 Weight $F_o$ : Object Detection Confidence

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

### 3.5.4 Weight $F_{ot}$ : Ordering and Compatibility

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

<sup>4</sup>Object categories correspond to object detectors described in Section 3.2

**Object Ordering Statistics:** People have biases on the order of topic or content flow. We measure these biases by collecting statistics on ordering for object naming from the 1 million image descriptions in the SBU Captioned Dataset (Ordonez et al., 2011). Let  $f_{\text{ord}}(w_1, w_2)$  be the number of times  $w_1$  appeared before  $w_2$ . For instance,  $f_{\text{ord}}(\text{window}, \text{house}) = 2895$  and  $f_{\text{ord}}(\text{house}, \text{window}) = 1250$ , suggesting that people are more likely to mention a window before mentioning a house/building<sup>5</sup>. We use these ordering statistics to enhance content flow. We define score for the order of objects using Z-score for normalization as follows:

$$F_{ot} = \frac{f_{\text{ord}}(c_o, c_t) - \text{mean}(f_{\text{ord}})}{\text{std\_dev}(f_{\text{ord}})} \quad (3.10)$$

We then transform  $F_{ot}$  to be in the range  $[0,1]$  making sure that its value smaller for better choice.

## 3.6 Surface Realization – Phrase Composition

Recall that for each image, the computer vision system identifies phrases from descriptions of images that are similar in a variety of aspects. The result is a set of phrases representing four different types of information : object-NPs, action-VPs, region/stuff-PPs, and scene-PPs. From this assortment of phrases, we aim to select a subset and glue them together to compose a complete sentence that is linguistically plausible and semantically truthful to the content of the image.

### 3.6.1 Variables and Objective Function

From content planning (Section 3.5), we have selected tuples  $\{\alpha_{ok}\}$ , for which we now apply surface realization. Assume ILP selected  $D'$  objects, we number them from 0 to  $D' - 1$ . We begin by introducing the variables and objective function of our ILP formulation. Each variable is indexed by a selected object  $o^6$ , corresponding to a separate sentence. This allows

---

<sup>5</sup>We take into account synonyms.

<sup>6</sup>we display object index  $o$  as a superscript in the equations



us to use a single objective function for all sentences in the description. We could run a separate ILP instance for each sentence, but having a single ILP instance allows us to introduce discourse constraints, such as a single scene phrase per description.

**Variables for Sequence Structure:** Variables  $\alpha$  encode phrase selection and ordering:

$$\alpha_{ik}^o = 1 \quad \text{iff} \quad \text{phrase } i \in P \text{ is selected} \quad (3.11)$$

for position  $k \in [0, N)$

Where  $k \in [0, N)$  is one of the  $N=4$  positions in a sentence<sup>7</sup>. Additionally, we define variables for each pair of adjacent phrases to capture sequence cohesion:

$$\alpha_{ijk}^o = 1 \quad \text{iff} \quad \alpha_{ik}^o = \alpha_{j(k+1)}^o = 1 \quad (3.12)$$

We model tree composition as maximization of the following objective function<sup>8</sup>:

$$F = \sum_o \left( \sum_i F_i^o \times \sum_{k=0}^{N-1} \alpha_{ik}^o \right) + \sum_{ij} F_{ij}^o \times \sum_{k=0}^{N-2} \alpha_{ijk}^o \quad (3.13)$$

This objective is comprised of three types of weights (confidence scores):  $F_i^o, F_{ij}^o$  (All weights are normalized using z-score).  $F_i^o$  represents the phrase selection score based on visual similarity, described in Section 3.3.  $F_{ij}^o$  quantifies the sequence cohesion across phrase boundaries. For this, we use  $n$ -gram scores ( $n \in [2, 5]$ ) between adjacent phrases computed using the Google Web 1-T corpus (Brants and Franz., 2006). Finally,  $F_r$  quantifies PCFG rule scores (log probabilities) estimated from the 1M image caption corpus (Ordonez et al., 2011) parsed using Stanford parser (Klein and Manning, 2003).

<sup>7</sup>The number of positions is equal to the number of phrase types, since we select *at most* one from each type.

<sup>8</sup>Note that we indicate object index  $o$  as a superscript in both, scores and variables

One can view  $F_i^o$  as a *content selection* score, while  $F_{ij}^o$  corresponds to *linguistic fluency* scores capturing sequence and tree structure respectively. If we set positive values for all of these weights, the optimization function would be biased toward verbose production, since selecting an additional phrase will increase the objective function. To control for verbosity, we set scores corresponding to linguistic fluency, i.e.,  $F_{ij}^o$  uses negative values (smaller absolute values for higher fluency), to balance dynamics between content selection and linguistic fluency.

Negative scores encourage ILP to variate number of variable assigned to 1. This allows us to avoid overloaded descriptions akin to [The little boat] [rests in the water in Cape Porpoise, Maine] [in the sky] [in front of our beach house] and generate simpler description, such as [The little boat] [ in front of our beach house], where not all four types of phrases are selected.

We optionally prepend the first sentence in a generated description with a *cognitive phrase*.<sup>9</sup> These are generic constructs that are often used to start a description about an image, for instance, “This is an image of...”. We treat these phrases as an additional type, but omit corresponding variables and constraints for brevity.

### 3.6.2 Constraints

**Soundness Constraints:** We need constraints to enforce consistency between different types of variables (Equations 3.11 and 3.12). Constraints for a product of two variables have been discussed by Clarke and Lapata (2008). We add the following constraints.

$$\forall i, j \in P, k \in [0, N),$$

---

<sup>9</sup>We collect most frequent 200 phrases of length 1-7 that start a caption from the SBU Captioned Photo Collection.

$$\forall_{ijk}, \alpha_{ijk} \leq \alpha_{ik} \quad (3.14)$$

$$\alpha_{ijk} \leq \alpha_{j(k+1)}$$

$$\alpha_{ijk} + (1 - \alpha_{ik}) + (1 - \alpha_{j(k+1)}) \geq 1$$

**Sequence Congruence Constraints:** To generate informative descriptions for sequence driven ILP, we choose to include at least two phrases for each sentence:

$$\forall s, \sum_{ij} \alpha_{i0}^o = 1 \quad (3.15)$$

$$\forall s, \sum_{ij} \alpha_{i1}^o = 1 \quad (3.16)$$

For a sentence we allow variable-length generation, i.e., up to  $N$  phrases can be selected in the final output. When  $l$  phrases are selected, we require the first contiguous  $l$  slots to be filled:

$$\forall k = 2, \dots, N - 2, \sum_i \alpha_{i(k+1)}^o \leq \sum_i \alpha_{ik}^o \quad (3.17)$$

Note that the constraints given by Eq. 3.17 together with the initial conditions (Eq. 3.15 and 3.16) will also enforce that at most one phrase can be placed in any position.

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

$$\forall o, \sum_{i \in P^{NP}} \sum_{k=0}^{N-1} \alpha_{ik}^o = 1 \quad (3.18)$$

Where,  $P^{NP}$  is a set of  $NP$  phrases.

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

$$\forall o, T, \sum_{i \in P^T} \sum_{k=0}^{N-1} \alpha_{ik}^o \leq 1 \quad (3.19)$$

Where,  $P^T$  is a set of phrases of type  $T$  (NP,VP,PP for stuff or PP for scene).

To ensure that we generate grammatically correct sentences we disallow plural (singular) form of a noun be chosen together with singular (plural) form of a verb: To enforce plural/singular agreements between NP and VP: We enforce plural/singular agreements between NP and VP and correct minor grammatical errors (e.g., gender, determiner agreement, etc.) through simple post-processing – descriptions omitted for brevity.

$$\begin{aligned} \forall o, \quad & \sum_{i \in NP_{singular}} \sum_{k=0}^{N-1} \alpha_{ik}^o + \\ & + \sum_{i \in VP_{plural}} \sum_{k=0}^{N-1} \alpha_{ik}^o \leq 1 \end{aligned} \quad (3.20)$$

$$\begin{aligned} \forall o, \quad & \sum_{i \in NP_{plural}} \sum_{k=0}^{N-1} \alpha_{ik}^o + \\ & + \sum_{i \in VP_{singular}} \sum_{k=0}^{N-1} \alpha_{ik}^o \leq 1 \end{aligned} \quad (3.21)$$

Finally, we restrict verb phrases in a sentence after a cognitive phrase to be in gerund form or in past tense. Denote the set of verb phrases that cannot be used in a sentence after a cognitive phrase as  $VP_{nocogn}$ , then,

$$\sum_{i \in VP_{nocogn}} \sum_{k=0}^{N-1} \alpha_{ik}^o + \sum_c \alpha_c^{cogn} \leq 1 \quad (3.22)$$

**Discourse constraints:** When we are choosing a prepositional phrase for a scene most likely it will be the same for all objects in an image. As shown at figure 3.8, description (a) has the same scene phrase “in yellow pine forest” for both sentences.

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

$$\forall i \in PP_{scene}, \quad \sum_{o,k} \alpha_{ik}^o \leq 1 \quad (3.23)$$



**(a)** Many blue flowers in yellow pine forest.  
The white water lily in yellow pine forest.

**(b)** Many blue flowers growing in Susan.  
The white water lily looked so pure and fresh  
bathing in the rain.

**Figure 3.8:** Sentences with the Same Scene Phrase

In this case we get description (b) at figure 3.8 as a result.

Another problem that we can face is when phrases with the same head words are chosen. Figure 3.9 contains an example of this situation. As we see from description (a) of the top image scene PP (in the tree) and stuff PP (in a tree) have the same head word “tree”. Similarly for the bottom image two phrases with the same head “water” are chosen.

We add constraints that prevent the inclusion of more than one phrase with identical head words:

$\forall o$  and  $i, j$  with the same heads,

$$\sum_{k=1}^N \alpha_{ik}^o + \sum_{k=1}^N \alpha_{jk}^o \leq 1 \quad (3.24)$$

Then for the example at figure 3.9 we get description (b) as a result, which for both images sounds more human-like.

Additionally, we disallow a verb-phrase at the beginning of a sentence:

$$\forall o, \quad \sum_{i \in P^{VP}} \alpha_{i0}^o = 0 \quad (3.25)$$

Where,  $P^{VP}$  is a set of  $VP$  phrases.



**(a)** The rose pots my mom has out around our house **in the tree in a tree**. The flower was in the trees.

**(b)** The rose pots my mom has out around our house **in the tree**. The flower was in the trees.



**(a)** Yellow bird seen near the port in Aruba **in the water in water**. An interesting looking bird munching on a fish in Sua in water.

**(b)** Yellow bird seen near the port in Aruba **at the beach by the water**. An interesting looking bird munching on a fish in Sua by the water.

Figure 3.9: Phrases with the Same Head Word

### 3.6.3 Unary Phrase Selection

Let  $M_i^o$  be the confidence score for phrase  $\alpha_i^o$  given by the image-phrase matching algorithm. To make the scores across different phrase types comparable, we normalize them using Z-score:

$$F_i^o = \text{norm}'(M_i^o) = (M_i^o - \text{mean}_{T_i}) / \text{dev}_{T_i} \quad (3.26)$$

Where  $T_i$  is the type of phrase  $i$ . We then transform the values into the range of  $[0,1]$  to make it comparable with the range  $[0, 1]$  of  $F_{oij}$ .<sup>10</sup>

$$F_{oj} = \text{norm}(M_{oi}) = (\text{norm}'(M_{oi}) - 3) / 6 \quad (3.27)$$

<sup>10</sup>This works because 99% of resulting Z-scores  $\in [-3, 3]$ .

### 3.6.4 Pairwise Phrase Cohesion

In this section, we describe the pairwise phrase cohesion score  $F_{ij}^o$  defined for each  $\alpha_{ij}^o$  in the objective function (Eq. 3.13). Via  $F_{ij}^o$ , we aim to quantify the degree of syntactic and semantic cohesion across two phrases  $\alpha_i^o$  and  $\alpha_j^o$ . Note that we subtract this cohesion score from the objective function. This trick helps the ILP solver to generate sentences with varying number of phrases, rather than always selecting the maximum number of phrases allowed.

**Pointwise Mutual Information:** PMI is used to compute the score, normalized to ensure weights are in range  $[0,1]$ : , defined as:

$$PMI(ngr) = \log \frac{f(ngr)}{\prod_{w \in ngr} f(w)} \quad (3.28)$$

Where  $f(ngr)$  is ngram frequency and  $f(w)$  is unigram frequency for a word  $w$  belonging to ngram  $ngr$ .

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

$$F_{ij}^{o, NGRAM} = 1 - \frac{\sum_{l \in L_{ij}^o} NPMI(l)}{size(L_{ij}^o)} \quad (3.29)$$

$$NPMI(ngr) = \frac{PMI(ngr) - PMI_{min}}{PMI_{max} - PMI_{min}} \quad (3.30)$$

Where NPMI is the normalized point-wise mutual information.<sup>11</sup> Notice that by taking average of ngrams pmi we ensure that  $F_{oij}^{o, NGRAM}$  is in the range  $[0,1]$ . By subtracting the fraction from 1 we ensure that the score is minimum for the best cohesion. We need to do that because we subtract the score from objective function.

<sup>11</sup>We include the n-gram cohesion for the sentence boundaries as well, by approximating statistics for sentence boundaries with punctuation marks in the Google Web 1-T data.

**Co-occurrence Cohesion Score:** To capture long-distance cohesion, we introduce a co-occurrence-based score, which measures order-preserved co-occurrence statistics between the head words  $h_i^o$  and  $h_j^o$ <sup>12</sup>. Let  $f_\Sigma(h_i^o, h_j^o)$  be the sum frequency of all n-grams of length between 2 and 5 that start with  $h_i^o$ , end with  $h_j^o$  and contain a preposition  $prep(j)$  of the phrase  $j$ . We take  $prep(spq)$  from the phrase  $spq$  if it is a prepositional phrase. Then the co-occurrence cohesion is computed as:

$$F_{ij}^{o,CO} = \frac{\max(f_\Sigma) - f_\Sigma(h_i^o, h_j^o)}{\max(f_\Sigma) - \min(f_\Sigma)} \quad (3.31)$$

Note that we subtract the score from objective function, therefore by defining  $F_{ij}^{o,CO}$  this way we make sure the score is larger for worse ordering choice.

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

$$F_{ij}^o = \frac{w^{NGRAM} \cdot F_{ij}^{o,NGRAM} + w^{CO} \cdot F_{ij}^{o,CO}}{w^{NGRAM} + w^{CO}} \quad (3.32)$$

where  $w^{NGRAM}$  and  $w^{CO}$  can be tuned via grid search, and  $F_{ij}^{o,NGRAM}$  and  $F_{ij}^{o,CO}$  are normalized  $\in [0, 1]$  for comparability. Notice that  $F_{ij}^o$  is in the range  $[0, 1]$  as well.

### 3.6.5 Phrase Score For Beginning/End Of The Sentence

We collect statistics from Google Web 1-T dataset for each word to be at the beginning and at the end of sentence. We then use it to compute the score as the following:

$$F_i^{o,BEGIN} = \frac{f_{begin}^o(i) - \min(f_{begin})}{\max(f_{begin}) - \min(f_{begin})} \quad (3.33)$$

---

<sup>12</sup>For simplicity, we use the last word of a phrase as the head word, except VPs where we take the main verb.



$$F_i^{o,END} = \frac{f_{end}^o(i) - \min(f_{end})}{\max(f_{end}) - \min(f_{end})} \quad (3.34)$$

Where  $f_{begin}^o(i)$  is a frequency of the first word of phrase  $i$  being at the beginning of the sentence and  $f_{end}^o(i)$  is a frequency of the last word of phrase  $i$  being at the end of the sentence

### 3.6.6 Cognitive Phrases

For our problem we define cognitive phrases as generic constructs that are often used by people to start a description about an image, for instance: “This is an image of...”, “I like the way the...”, “In this picture we can see a...”. We collected a set of 193 cognitive phrases from the 1 million image descriptions in the SBU Captioned Photo Dataset (Ordonez et al., 2011) by mining the most common phrases of different word lengths from these image descriptions and filtering out the phrases containing references to specific object names. We automatically select one of these cognitive phrases to start our descriptions by incorporating them into our optimization formulation.

The score for cognitive phrases is defined as N-gram Cohesion Score between cognitive phrase and the phrase adjacent to it.

## 3.7 Evaluation of Sequence-driven Composition

### Approach

#### 3.7.1 TestSet

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

#### 3.7.2 Baselines, Gold Standard and System Versions

We compare our ILP approaches with a few non-trivial baselines:

- **HMM**: an HMM approach (comparable to Yang et al. (2011)), which takes as input the same set of candidate phrases described in Section 3.3, but for decoding, we fix the ordering of phrases as [ NP – VP – Region PP – Scene PP] and find the best combination of phrases using the Viterbi algorithm. We use the same rich set of pairwise phrase cohesion scores (Sections 3.6.4) used for the ILP formulation, producing a strong baseline<sup>13</sup>.
- **HMM+COGN**: HMM enhanced with cognitive phrases.
- **RETRIEVAL**: a RETRIEVAL based description method (Ordonez et al., 2011), that searches the large parallel corpus of images and captions, and transfers a caption from a visually similar database image to the query. This again is a very strong baseline, as

---

<sup>13</sup>Including other long-distance scores in HMM decoding would make the problem NP-hard and require more sophisticated decoding, e.g. ILP.

it exploits the vast amount of image-caption data, and produces a description high in linguistic quality (since the captions were written by human annotators).

We call our system SEQ+LINGRULE to reflect sequence-driven phrase composition enhanced with linguistically motivated constraints. We experiment with 2 versions of our ILP system: SEQ.v.1+LINGRULE and SEQ.v.2+LINGRULE. They differ only slightly, mainly in weights assigned to each term of the objective function (equation 3.13 and Section 3.6.4). The reason for having two system versions is purely experimental. Later in our work we abandon the first version and use only the second one. Note, that HMM baseline is using the same scores, thus, it also has two versions, compatible to SEQ+LINGRULE versions. For experiments, we use only the first version of HMM as the results are very close to each other.

As a GOLD STANDARD we use descriptions written by Flickr users - HUMAN.

### 3.7.3 Automatic Evaluation:

Automatically quantifying the quality of machine generated sentences is known to be difficult. BLEU score (Papineni et al., 2002), despite its simplicity and limitations, has been one of the common choices for automatic evaluation of image descriptions (Farhadi et al., 2010; Kulkarni et al., 2011; Li et al., 2011; Ordonez et al., 2011), as it correlates reasonably well with human evaluation (Belz and Reiter, 2006). While recent studies showed that Meteor (Denkowski and Lavie, 2011) has a higher correlation (Elliott and Keller, 2014).

We use the NIST implementation of BLEU score<sup>14</sup>. For Meteor score we weight Precision and Recall equally.

Table 3.1 shows the the BLEU@1 (single reference) and Meteor score against the original caption of 1000 images. We can see that SEQ+LINGRULE performs better than HMM. Furthermore, the second version of SEQ+LINGRULE improves the scores over the first version. Surprisingly, *cognitive phrases* improved the scores for SEQ+LINGRULE, which could be due to their presence in the HUMAN captions.

---

<sup>14</sup><ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a-20091001.tar.gz>

Method	Bleu	Meteor		
		P	R	M
HMM.V.1	0.1136	0.098	0.181	0.083
HMM.V.1+COGN	0.1061	0.093	0.195	0.083
SEQ.V.1+LINGRULE+COGN	0.1450	0.125	0.154	0.088
SEQ.V.2+LINGRULE	<b>0.1607</b>	0.133	0.149	0.091
SEQ.V.2+LINGRULE+COGN	0.1518	0.130	0.170	<b>0.095</b>

**Table 3.1:** Automatic Evaluation of Sequential ILP

	Grammar	Cognitive	Relevance
HMM+COGN	3.40( $\sigma=.82$ )	3.40( $\sigma=.88$ )	2.25( $\sigma=1.37$ )
SEQ+LINGRULE+COGN	3.56( $\sigma=.90$ )	3.60( $\sigma=.98$ )	2.37( $\sigma=1.49$ )
HUMAN	4.36( $\sigma=.79$ )	4.77( $\sigma=.66$ )	3.86( $\sigma=1.60$ )

**Table 3.2:** Human Evaluation of Sequential ILP: Multi-Aspect Rating ( $\sigma$  is a standard deviation)

### 3.7.4 Human Evaluation I: Multi-Aspect Rating

Neither BLEU nor METEOR directly measures grammatical correctness over long distances and may not correspond perfectly to human judgements. Therefore, we complement the automatic evaluation with Amazon Mechanical Turk (AMT) evaluation (Snow et al., 2008). Table 3.2 presents rating in the 1–5 scale (5: perfect, 4: almost perfect, 3: 70~80% good, 2: 50~70% good, 1: totally bad) in three different aspects: *grammar*, *cognitive correctness*,<sup>15</sup> and *relevance*. We find that SEQ+LINGRULE improves over HMM in all aspects, however, the relevance score is noticeably worse than scores of two other criteria. It turns out human raters are generally more critical against the relevance aspect, as can be seen in the ratings given to the original human generated captions.

### 3.7.5 Human Evaluation II: Forced Choice

We ask AMT users to choose a better caption between two choices<sup>16</sup>. We do this rating with and without showing the images, as summarized in Table 3.3. When images are shown,

<sup>15</sup>E.g., “A desk on top of a cat” is grammatically correct, but cognitively absurd.

<sup>16</sup>We present two captions in a randomized order.

Method-1	Method-2	w/Images	Method-1 preferred (%)
SEQ.v.1+LINGRULE	HMM.v.1	-	67
SEQ.v.1+LINGRULE+COGN	HMM.v.1+COGN	-	66
SEQ.v.1+LINGRULE	HMM.v.1	+	53
SEQ.v.1+LINGRULE+COGN	HMM.v.1+COGN	+	55
SEQ.v.1+LINGRULE+COGN	RETRIEVAL	+	72
SEQ.v.1+LINGRULE+COGN	HUMAN	+	16
SEQ.v.2+LINGRULE+COGN	SEQ.v.1+LINGRULE+COGN	+	54

**Table 3.3:** Human Evaluation of Sequential ILP: posed as a binary question “*which of the two options is better?*”. According to Pearson’s  $\chi^2$  test, all results are statistically significant.



**HMM.v.1+Cogn:** These was taken on the flowers growing in a rock garden in the field in two sorts. This little flower sprouted up in defiance in the field in two sorts. A full open flower sprouted up in defiance in the field in gardens. Bright yellow flowers growing in a rock garden in the field in gardens.

**Seq.v.1+LingRule+Cogn:** [This is a photo of this little flower sprouted up in defiance against grass. Bright yellow flowers growing in a rock garden at Volcan Mombacho.](#)

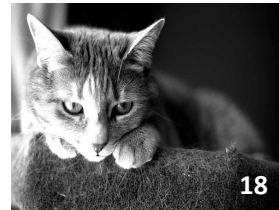
**Human:** Yellow flower in my field.



**HMM.v.1+Cogn:** Love the castle known for being the home of Hamlet in the Shakespeare play among the green and the sky across the water.

**Seq.v.1+LingRule+Cogn:** [Love the castle across the water among the green and the sky.](#)

**Human:** We had to cross this bridge in order to see the Leaning Tower of Pisa. This area was really peaceful.



**HMM.v.1+Cogn:** Found kitty resting in her bed and looking out the window from the baptismal pool.

**Seq.v.1+LingRule+Cogn:** [This is what happens when a cat resting in her bed and looking out the window.](#)

**Human:** Cat in the cat tree - Black and White



**HMM.v.1+Cogn:** This is the first cellar door left back bedroom in center and clothes dryer to the right to the building in the house. This HUGE screen hanging on the wall outside a burned down building in the house. My truck parked on first avenue in the east village by the glass buildings in the house.

**Seq.v.1+LingRule+Cogn:** [Found trucks parked on first avenue in the east village.](#)

**Human:** Flat bed Chisholms truck on display at the vintage vehicle rally at Astley Green Colliery near Leigh Lancs.

**Figure 3.10:** SEQ+LINGRULE & HMM Generated Descriptions, where SEQ+LINGRULE (blue underlined font) was Preferred

raters evaluate content relevance as well as linguistic quality of the captions. Without images, raters evaluate only linguistic quality.







We found that raters generally prefer SEQ+LINGRULE generated captions over HMM generated ones, twice as much (67.2% SEQ+LINGRULE V.S. 32.8% HMM), if images are not

presented. However the difference is less pronounced when images are shown. There could be two possible reasons. The first is that when images are shown, the Turkers do not try as hard to tell apart the subtle difference between the two imperfect captions. The second is that the relative content relevance of ILP generated captions is negating the superiority in linguistic quality. We explore this question using multi-aspect rating, described below. Examples of images, for which descriptions generated by SEQ+LINGRULE are better than those generated by HMM, are shown in Figure 3.10.

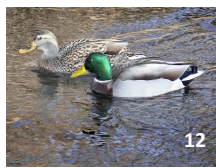
Note that SEQ+LINGRULE generated captions are exceedingly (71.8 %) preferred over the RETRIEVAL baseline (Ordonez et al., 2011), despite the generated captions tendency to be more prone to grammatical and cognitive errors than retrieved ones. This indicates that the generated captions must have substantially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

### 3.8 Discussion

Figure 3.11 some bad examples, generated by our system. One of the problems is due to noisy Computer Vision object detections. For example, a “turtle” was detected as a “shoulder bag”. This particular type of a problem is a hard one and outside of the scope of our research, which is NLP oriented. We, however, deal with a related problem, examples of which are given in Figure 3.12. Descriptions shown in the figure contain a lot of visually unverifiable information, which we call extraneous. For example, “in Venezuela”. We deal with this problem in the next two Chapters 4 and 5. Resolving cognitive problem requires taking into account deeper semantic relation between phrases, which can be a heavier load for the system. Whilst it would be interesting to resolve this problem, we decide to leave it as a future work. We would like to deal with another problem, namely, grammar and address it in Chapter 6.

Relevance problems	 6	<p><b>Seq.v.1+LingRule+Cogn:</b> This is what happens when a train lost against the mountain background at Horseshoe Curve in April 1986 under a rock.</p> <p><b>Seq.v.2+LingRule+Cogn:</b> This is a train lost against the mountain background at Horseshoe Curve in April 1986 under a rock.</p> <p><b>Human:</b> Found this guy right outside my sleeping bag waking up in OKC.</p>
	 7	<p><b>Seq.v.1+LingRule+Cogn:</b> This is a shoulder bag with a blended rainbow effect.</p> <p><b>Seq.v.2+LingRule+Cogn:</b> This is a shoulder bag featuring many vintage buttons in different materials with a blended rainbow effect.</p> <p><b>Human:</b> A softshell turtle spotted during a walk in Ft. Worth, TX USA.</p>
Grammar problems	 8	<p><b>Seq.v.1+LingRule+Cogn:</b> This is a shot of the road fixed to the table the cup.</p> <p><b>Seq.v.2+LingRule+Cogn:</b> One of the road fixed to the table the cup.</p> <p><b>Human:</b> Cup by Corning, plate marked Sterling vitrified china, East Liverpool, OH, G-3.</p>
	 9	<p><b>Seq.v.1+LingRule+Cogn:</b> Of apples the butterfly feeding in Judy flower garden by a tree.</p> <p><b>Seq.v.2+LingRule+Cogn:</b> Of apples the butterfly feeding in Judy flower garden by a tree.</p> <p><b>Human:</b> At a butterfly house somewhere in North Wales.</p>
Cognitive problems	 10	<p><b>Seq.v.1+LingRule+Cogn:</b> I like the way the clouds walking in the poppy field under cloudy sky.</p> <p><b>Seq.v.2+LingRule+Cogn:</b> One of the clouds walking in the poppy field under a cloudy sky.</p> <p><b>Human:</b> A bike in a field dreams of unconvencionnal places.</p>
		<p><b>Seq.v.1+LingRule+Cogn:</b> Here you can see a cross by the frog in the sky.</p> <p><b>Seq.v.2+LingRule+Cogn:</b> This is a cross standing by the frog.</p> <p><b>Human:</b> Kites were flying all around the Washington monument during the cherry blossom festival. It was beautiful.</p>

**Figure 3.11:** Examples with Different Aspects of Problems in the SEQ+LINGRULE Generated Descriptions.



**Seq.v.1+LingRule+Cogn:** Photo of a mandarin duck having a fight over a chip a boy had thrown into the water of the pond.

**Seq.v.2+LingRule+Cogn:** At De Wolfe Point state park in water arrived weeks ago and were sitting around on the snow until the pond thawed these ducks.

**Human:** Maybe the most common bird in the neighborhood, not just the most common water fowl in the neighborhood! Ralston Creek Trail, 12-16-09.



**Seq.v.1+LingRule+Cogn:** This is the view from the clock tower converted into an office along Pacific Coast Highway in Seal Beach CA Approximately 16200 PCH from Mauthausen.

**Seq.v.2+LingRule+Cogn:** View from the top of the clock tower converted into an office along Pacific Coast Highway in Seal Beach CA Approximately 16200 PCH from Mauthausen.

**Human:** Clock tower in downtown.



**Seq.v.1+LingRule+Cogn:** Here you can see the butterflies attracted to the colorful flowers in Hope Gardens.

**Seq.v.2+LingRule+Cogn:** I liked the way into their life attracted to the colorful flowers in Hope Gardens the butterflies.

**Human:** A butterfly on a flower near the Hammocks Clubhouse in Bald Head Island, North Carolina.



**Seq.v.1+LingRule+Cogn:** The flower in a field near Flagstaff dancing with the wind by the road side. The flowers in a field buds under the microscope.

**Seq.v.2+LingRule+Cogn:** Found this flower taken in Madrid March 2006 near Flagstaff. A native flower found in Venezuela.

**Human:** Yellow flower near Morava river.

**Figure 3.12:** Examples with Problems in the SEQ+LINGRULE Generated Captions due to Extraneous Information (red underlined font) in Image Captions.



CHAPTER 4  
DEPENDENCY-BASED SEQUENCE-DRIVEN CAPTION  
GENERALIZATION

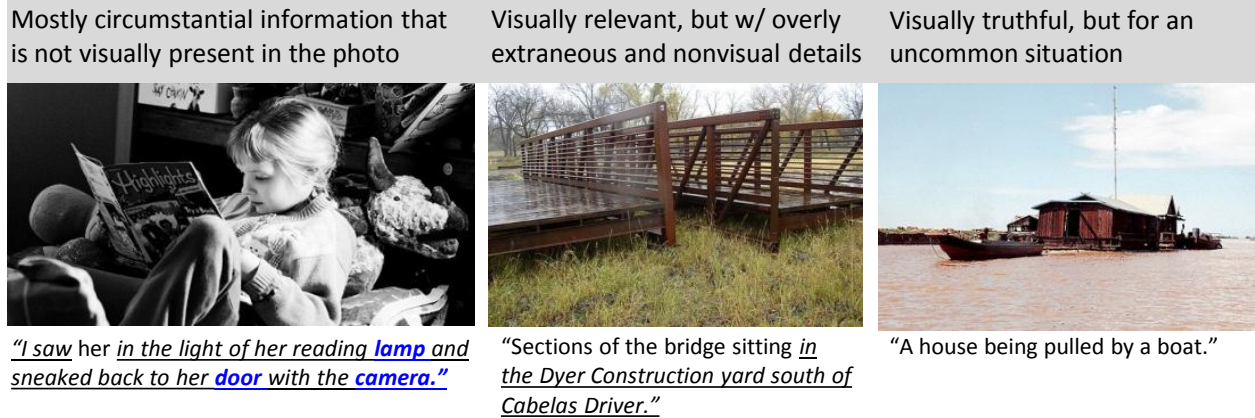
## 4.1 Extraneous Information

In Chapter 3 Section 3.8 we saw some examples of generated descriptions not quite relevant to the image content (Figure 3.12). In other words, some information, present in the descriptions, is not visually verifiable. For example, “in Hope Gardens” or “having a fight over a chip boy had thrown into the water of the pond”. This information is mainly circumstantial or overly extraneous with non-visual details. This happened because we exploit human-written captions to compose a new image description. Users of Flickr, from which we obtain ready human-written captions, tend to include extraneous information into the captions. As been noted by recent studies (e.g., Mason and Charniak (2013), Kuznetsova et al. (2013b), Jamieson et al. (2010), Dodge et al. (2012)), naturally existing image captions often include contextual information that does not directly describe visual content, which ultimately hinders their usefulness for describing other images.

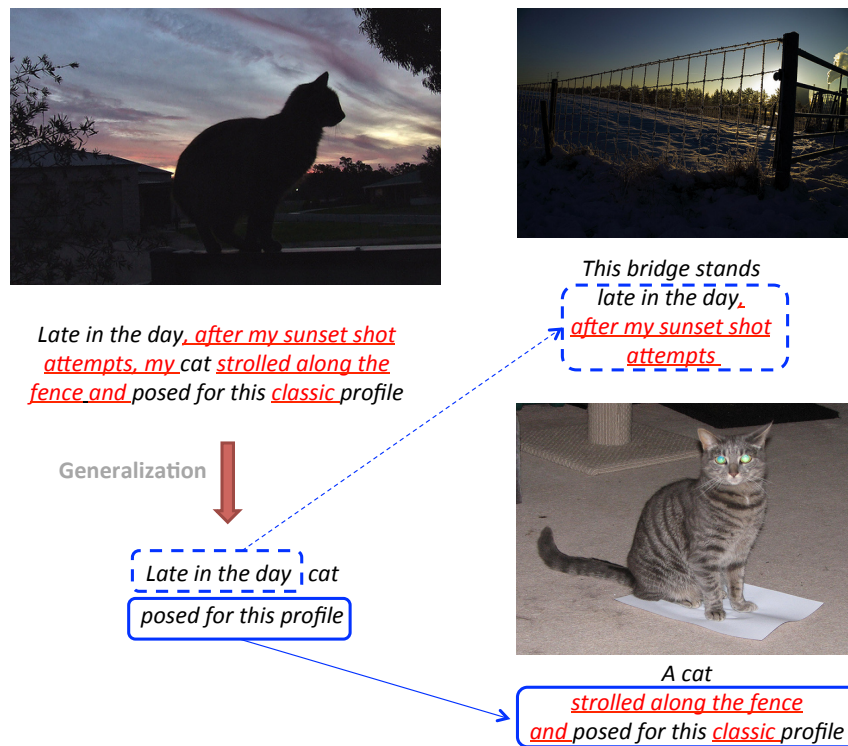
Consider the first image and its caption shown in Figure 4.1. The original caption was written by the person who uploaded, and presumably took, the photo. It contains specifics such as “*in the light of her reading lamp and sneaked back to her door with the camera.*” that may have been relevant to the circumstance in which the picture was taken, but objects such as “lamp”, “door”, “camera” are not visually present anywhere in the photo.

The second image shows a similar but somewhat different issue. Its caption “*Sections of the bridge sitting in the Dyer Construction yard south of Cabelas Driver.*” concerns visually present objects such as “*bridge*” and “*yard*”, but some of the details such as “*Dyer Construction*” and “*south of Cabelas Driver*” are overly specific and not visually detectable.

In the case of third image, the caption “*A house being pulled by a boat*” is the most pertinent to the visual content of the image, but such a caption is unlikely to be applicable



**Figure 4.1:** Examples of Human-written Image Captions with Extraneous Information



**Figure 4.2:** Compressed captions (on the left) are more applicable for describing new images (on the right).

to many other similar photos as is, because such a scenery of a boat pulling a floating house is not a common one. More general, therefore more useful bits of text would be "a boat and a house" or "a boat is parked next to the house".

Content misalignment between images and text, such as shown in Figure 4.1 makes it

difficult to draw reliable mappings between visual content and text.

Because we are generating a description from the parts of such noisy captions, we are facing a problem with composed image descriptions. The problem is that extraneous information present in human-written captions is being transferred to the new description. Consider an example given at Figure 4.2 in the context of image caption generation. “Late in the day, after my sunset shot attempts, my cat strolled along the fence and posed for this classic profile”. From the first glance this caption perfectly describes the given image. However, if we are to compose a description for another image from parts of this description, we are more likely to make mistakes like “This bridge stands late in the day, after my sunset shot attempts”. Furthermore, we cannot be sure that every profile is classic, as in the image to the right cat’s profile is not perfect. Ideally we want a description for the image in the left of the figure to be “Late in the day cat posed for this profile”, which is simple and visually verifiable.

In order to generate human-like plausible descriptions, it would be hard to avoid the usage of human-written text. Thus we need to obtain much cleaner human-written captions. One way to solve the problem of extraneous information is to ask people to write more informative captions with minimum visually unverifiable information. This would be time and resource consuming<sup>1</sup>. Thus, we aim to automatically clean up the existing captions. In this Chapter we introduce a new task, which aims to do so, *image caption generalization*.

## 4.2 Related Work: Sentence Compression

We cast caption generalization as sentence compression. Latter was previously explored in a substantial number of research work. The common intent among most of those approaches is to generate informative, yet concise sentences from the original input. This can be achieved by removing auxiliary and redundant parts of the sentence or even rephrasing and reordering some sentence regions. The resulting compressions should retain the most

---

<sup>1</sup>Recall that our dataset consists of 1M images

important information present in the original sentence, additionally preserving fluency and grammaticality. Much work has considered deletion-only edits like ours (e.g. Knight and Marcu (2000), Turner and Charniak (2005), Cohn and Lapata (2007), Filippova and Altun (2013)), while recent ones explore more complex edits, such as substitutions, insertions and reordering (e.g. Cohn and Lapata (2008)). The latter generally requires a larger training corpus. While approaches, modelled with a more complex set of operations, can produce a larger variety of compressions, we allow only deletions. This makes our model easier for initial formulation. There are many modifications, which are potentially possible to make for our model to incorporate other operation. We leave it outside the scope of this work, as we do not aim to claim and prove any scalability of proposed compression algorithm at this stage.

A lot of work look at importance, grammaticality and compression rate as the main characteristics of a good compression (e.g. McDonald (2006), Cohn and Lapata (2008)). Importance and grammaticality can be explored via model structure itself, for instance, via dynamic programming approaches (McDonald (2006)). Compression criteria can also be encoded via a more general model with specific to the task parameters, for instance, as a global inference framework (e.g. Clarke and Lapata (2008)), or as a tree-transformation systems (e.g. Cohn and Lapata (2009), Turner and Charniak (2005), Knight and Marcu (2000), Galley and McKeown (2007), Woodsend and Lapata (2011)).

We consider two approaches to sentence compression. First one is based on dependency parse (Kuznetsova et al. (2013b)) and described in this Chapter. Second one is driven by a PCFG tree structure (Kuznetsova et al. (2014)) and described in the next Chapter 5.

In this Chapter we cast the generalization task as *visually-guided* sentence compression with lightweight revisions and formulate an optimization problem that aims to maximize the mixture of content selection and local linguistic fluency while satisfying a collection of constraints driven from dependency parse trees. Dependency-based constraints guide the generalized caption to be grammatically valid (e.g., keeping articles in place, preventing

dangling modifiers) while semantically compatible with respect to the given pair of an image and text (e.g., preserving predicate-argument relations).

## 4.3 Problem Formulation

We define a sentence  $X$  as an ordered set of words  $x_0x_1\dots x_{m-1}$ . Summarization task in this case is to find a subset  $Y = x_{i_1}\dots x_{i_c} = y_0, \dots, y_c$ , where  $\{i_1, \dots, i_c\} \subset \{0, \dots, m_s\}$ , such as an objective function  $F(Y)$  is maximized.

We will start with optimization criteria (Section 4.3.1), then we will formulate DP algorithm for the problem without constraints (Section 4.3.2). Further we will describe typed dependencies, which serve as a basis for soft and hard constraints in our task (Section 4.4). We will show that problem enhanced with constraints is much harder than the one without any constraints. Finally, we will describe DP and Beam Search to improve performance of the task (Section 4.5).

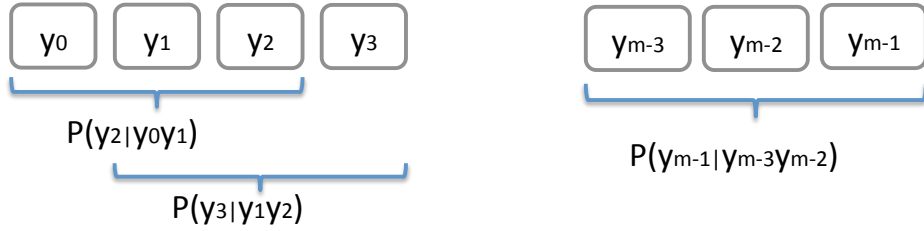
### 4.3.1 Optimization criteria

**Linguistics Fluency** We start with linguistics function, i.e. ngram probability of a sentence defined in eq. 4.1. We denote ngram size as  $n$ . In our experiments we used ngram of size 3. Computation of the function is shown at figure 4.3.

$$P(Y) = \prod_{i=1}^c P(y_i | y_{i-1} \dots y_{i-n+1}) \quad (4.1)$$

We experiment with two different ngram statistics, one extracted from the Google Web 1T corpus (Brants and Franz., 2006), and the other computed from the 1M image-caption corpus (Ordonez et al., 2011)

**Corpus Statistics** We consider three ways to incorporate corpus statistics into content selection:



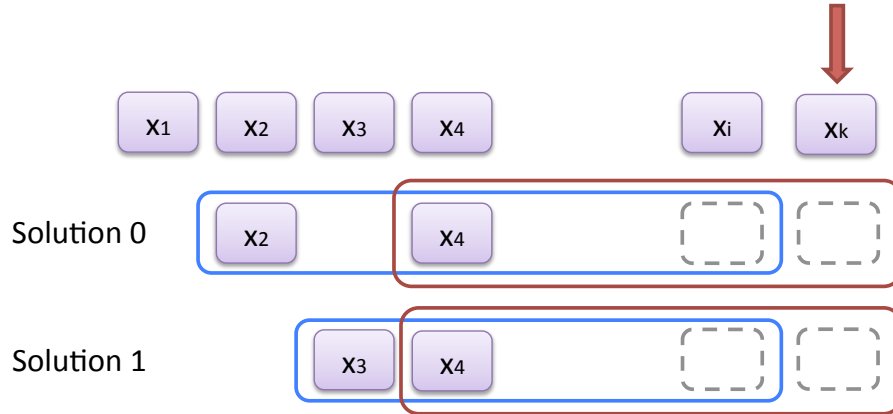
**Figure 4.3:** Sentence Ngram Probability

- $tf(x_i)$ : We hypothesize that words occurring frequently in the image caption corpus might correspond to more visually descriptive words. Hence, we consider  $\phi(x_i, v) = tf(x_i)$ , as total term frequency of  $x_i$  within the corpus. We set  $tf(x_i) = 0$  for all function words.
- $idf(x_i), tf.idf(x_i)$ : We also consider the use of  $idf$  and  $tf.idf$ , as content selection based on  $tf$  might favour overly generic, less informative caption words.

Additionally, we assign a very low content selection score ( $-\infty$ ) for proper nouns and numbers and assign a very high score (larger than maximum  $idf$  or visual score) for top 2k words in our corpus.

**Visual Relevance** Our task is to remove information not relevant to image content. Thus we use visual information to score our summarized sentence. The particular **computer vision system** used here consists of 7404 visual classifiers trained to recognize leaf level synsets from WordNet (Fellbaum, 1998). Each classifier is trained using labelled images from the ImageNet dataset (Deng et al., 2009) – an image database currently consisting of over 14 million hand labelled images organized according to the WordNet hierarchy in which each node of the hierarchy contains on average over 500 images.<sup>2</sup>

<sup>2</sup>Image similarity is represented using a Spatial Pyramid Match Kernel (SPM) (Lazebnik et al., 2006) with Locality-constrained Linear Coding (LLC) (Wang et al., 2010) on shape based SIFT features (Lowe, 2004b). Models are linear SVMs followed by a sigmoid to produce probability for each node. Code for this was provided by the authors of (Deng et al., 2012).



**Figure 4.4:** Sentence Sequence Driven Compression

### 4.3.2 DP formulation

If we are to approach this problem using brute force strategy we would have to try all possible summaries, i.e.  $2^m$  solutions. We can find the optimal summary,  $\hat{Y}$ , using dynamic programming (McDonald, 2006) described in this Section.

We start with describing a problem with only one optimization criteria: linguistic fluency, as its scope spreads beyond a single word. Later we will show how to enhance our task with the rest of the optimization criteria.

One way to approach this problem is to move from left to right along the sentence, computing sub-solutions at each step. At Figure 4.4 we can see an iteration of such an algorithm, where language model uses ngrams of size  $n$ . We need to find the solution with maximum ngram probability. The algorithm computes sub-solutions ending at an ngram of size  $n - 1$ . For example if  $n = 3$ , we compute sub-solutions ending at  $x_i x_k$  (Figure 4.4). We have to consider all combination of  $n - 1 = 2$  words preceding  $x_i x_k$ . This scenario is reflected in the equation 4.2, where we compute a sub-solution  $S[i_1 \dots i_{n-1}]$  ending at ngram  $x_{i_1} \dots x_{i_{n-1}}$  of size  $n - 1$ . The sub-solution depends on the previous sub-solutions  $S[j_1, \dots, j_{n-1}]$  ending at ngram of size  $n - 1$  as well. This recursive definition finds the most probable sequence ending at each word.

$$\begin{aligned}
S[i_1 \dots i_{n-1}] = & \arg \max_{j_1, j_2, \dots, j_{n-1} < 1 \dots (i_1-1)} [P(S[j_1, \dots, j_{n-1}]) \\
& \times P(x_{j_1} \dots x_{j_{n-1}} x_{i_1}) \\
& \dots \\
& \times P(x_{j_{n-1}} x_{i_1} \dots x_{i_{n-1}})]
\end{aligned} \tag{4.2}$$

The algorithm formulated in equation 4.2. To compute a single value defined in the equation we need to consider  $O(m^{n-1})$  possible previous sub-solutions  $j_1, \dots, j_{n-1}$ <sup>3</sup>. There are  $O(m^{n-1})$  such values  $S[i_1 \dots i_{n-1}]$ . Thus, to compute all the values we need  $O(m^{2n-2})$  computations.

## 4.4 Enhancing the Task with Dependency Constraints

N-gram based language models alone would not be able to generate syntactically and semantically correct compression. Therefore, we incorporate hard constraints driven from typed dependencies (de Marnee and Manning (2008)). Table 4.1 defines the list of dependencies used as dependency constraints. The column labelled “Direction” specifies the direction in which each constraint is applied. For example,  $dep(I \leftarrow J)$ , denotes that “ $I$ ” must be included in the summary whenever “ $J$ ” is included in the summary. Similarly,  $dep(I \longleftrightarrow J)$  denotes that “ $I$ ” and “ $J$ ” must be either included together or eliminated together.

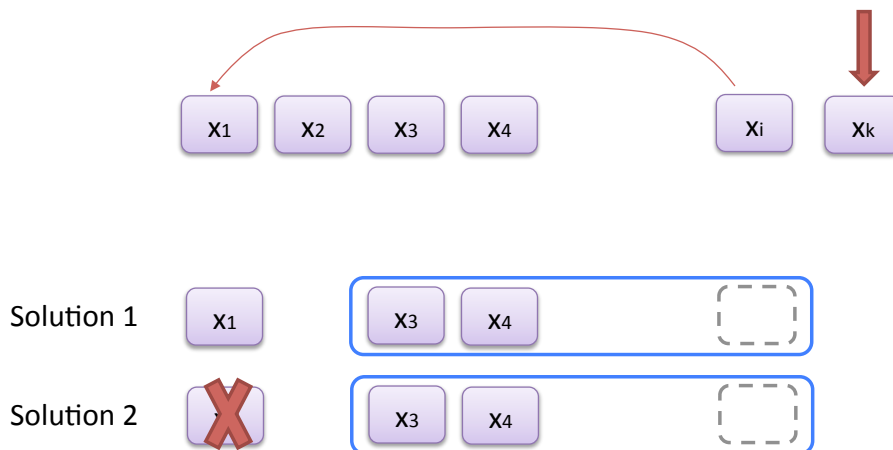
---

<sup>3</sup>We have  $n - 1$  positions to fill, each position can take one of potentially  $m$  values. This, however, is a rough upper bound



Name	Direction	Example	
		Sentence	Dependency
amod	←	A wooden chair in the living room	amod(chair← wooden)
advmod	←	This train car is parked permanently...	advmod(parked-5← permanently-6)
aux	↔	This crazy dog was jumping...	aux(jumping↔ was)
prep	←	A view from the balcony	prep(view← from)
det	↔	A cozy street cafe...	det(cafe↔ A)
dobj	↔	An inquisitive cow surveys the road...	dobj(surveys↔ road)
expl	↔	There are holes in the roof...	expl(are↔ There)
iobj	↔	...rock gives the water the color	iobj(gives↔ water)
neg	↔	Not a cloud in the sky...	neg(cloud↔ Not)
pobj	↔	This branch was on the ground...	pobj(on↔ ground)
prt	↔	...looking down at a building	prt(looking↔ down)
xcomp	→	The wind seems to talk...	xcomp(seems→ talk)
xsubj	→	The wind seems to talk...	xsubj(talk→ wind)
acomp(↔), agent(←), attr(↔), auxpass(↔), cc(↔), ccomp(→), complm(←), cop(↔), csubj(↔), csubjpass(↔), infmod(↔), mark(↔), mwe(↔), nn(←), npadvmod(←), nsubj(↔), nsubjpass(↔), num(←), number(↔), parataxis(←), partmod(←), pcomp(↔), poss(↔), possessive(↔), preconj(←), predet(←), purpcl(←), quantmod(←), rmod(←), ref(←), rel(↔), tmod(←), advcl(←)			

**Table 4.1:** Typed Dependency Constraints for Caption Generalization.

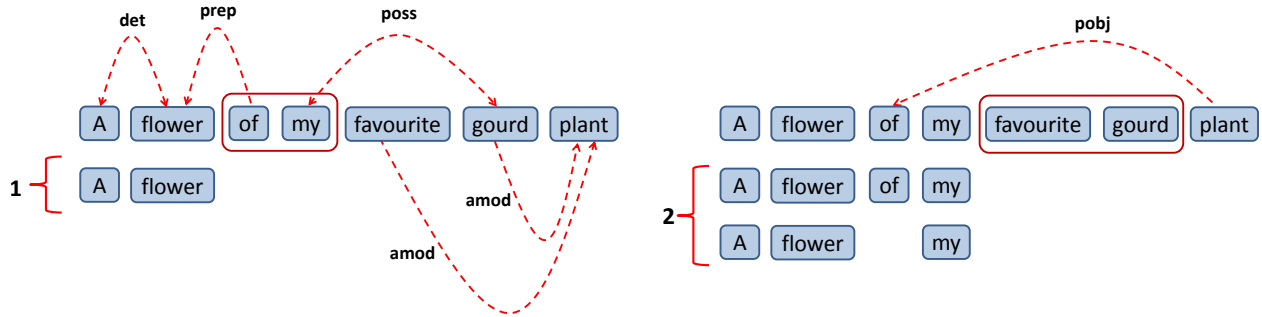


**Figure 4.5:** Sentence sequence driven compression with constraints (if systems keeps  $x_i$ , it must keep  $x_1$ ). Solution 2 does not satisfy the constraints.

Note that for some of the dependencies, the constraints are on the conservative side. For example, the bidirectional constraint  $\text{dobj}(\longleftrightarrow)$  may not be necessary for verbs that do not require direct objects. A generally more ideal approach could be to learn the constraints conditioning on the actual lexical items. However, given that the end goal of our approach is to produce better captions to serve as a parallel image-text corpus for other end applications, and that learning-based constraints are not likely to be perfect (especially that we do not have in-domain training data), we needed to take a more conservative strategy.

Enhancing the task with constraints makes summarization problem harder. Consider an algorithm iteration shown in Figure 4.5. Arrow from  $x_i$  to  $x_1$  shows that there is a constraint: if include  $x_i$  then include  $x_1$ . For computing ngram probability we however need only 2 preceding to  $x_i$  words and  $x_1$  is not necessarily among them. Thus at each step, we have to keep sub-optimal solutions, containing  $x_1$  and not containing  $x_1$ . That is to say, if we are currently at the iteration, corresponding to  $x_i$ , we have to keep a few sub-solutions, which vary by the words included before  $x_i$ . In this case the running time of the algorithm grows exponentially as we eventually will have to try all  $2^m$  solutions.

One way to reduce running time is to keep only the solutions which we need. For example



**Figure 4.6:** Sentence Sequence Driven Compression with Constraints by Example

in Figure 4.5, if there are no other dependencies involving words preceding  $x_k$ , we need only to keep 2 sub-solutions ending at  $x_3x_4$ : the one that contains  $w_1$  and the one that does not.

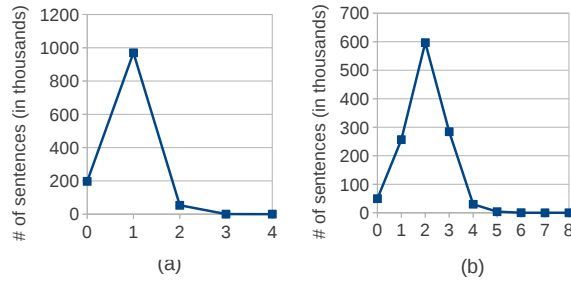
We describe this technique in Section 4.5, motivated by additional examples.

## 4.5 Dynamic Programming + Dependency

### Constraints + Beam Search

**Hard Constraints** We show it is possible to efficiently find the optimal solution subject to the hard constraints in Section 2.4, because we can compute the exact beam size necessary to find the optimal solution during decoding. If the needed beam size is not larger than some limit, it is possible to efficiently attain the global optimum solution. Note that we adjust the beam size dynamically at each location in the given sentence to avoid keeping an unnecessarily long history.

Assume that we are computing probabilities for sub-sequences ending at  $x_{i_1} \dots x_{i_k}$ . We need to consider only the most probable subsequence (satisfying the constraints) as well as all sub-sequences containing words dependent on any word following  $x_{i_k}$ . Figure 4.6 shows an example of such a situation using a trigram model. For example, if we are considering all sequences ending at “of my”, we need to iterate over all most probable sequences ending at each preceding word, in this case ending at “A” or “flower”. At this point there are no words among those whose inclusion into the final summary depends on any future word.



**Figure 4.7:** (a) Number of sentences for each average number of words with future dependencies (b) Number of sentences for each maximum number of words with future dependencies

Thus the number of solutions we need to store is 1. On the other hand if we are to evaluate all sequences ending at “favourite gourd” we need to take into account dependency between words “of” and “plant”, which forces us to include “of” whenever “plant” is included. In this case even if the most probable preceding sequence ending at “my” did not contain “of” we have to consider a intermediate sequence with “of” included. In this case we need to keep 2 intermediate solutions.

The needed beam size at each step depends on how many words have dependency constraints involving any word following the current one. The beam size is at most  $2^p$ , where  $p$  is the maximum number of words dependent on any future words at each step. In practice the number sentences containing many words with future dependencies is very low as shown by the statistics in Figure 4.7. We can see that for most of the sentences there is only 1 word on average or 2 words maximum for which we need to keep intermediate solutions. Thus for most sentences we do not need to keep more than 4 intermediate solutions at each step of the algorithm.

**Soft Constraints** We can introduce soft dependency constraints in order to allow a room for parser mistakes or seek for a compromise between ngram probability and grammatical correctness. In particular, instead of disallowing word combinations which do not satisfy particular constraints, we can add a penalty to the objective function. This way we are creating a tension between grammatical correctness of the generalized caption

and its ngram probability. We manually choose some of the initial dependency constraints (Table 4.1) to be hard constraints and the rest of them to be soft constraints.

## 4.6 Evaluation of Sequence-driven Caption

### Generalization

We evaluate intrinsic and extrinsic usefulness of the generalized captions. For extrinsic evaluation we apply image transfer task, not our composition based approach. This is done for simplicity as our second approach to caption generalization, described in Chapter 5, outperforms dependency-based approach, described in this sections. We evaluate phrase composition approach with a better caption generalization in Section 5.6 of Chapter 5.

We denote sequence-driven compression of this Chapter as SEQC.

#### 4.6.1 Methods for Compression

- ORIG: original uncompressed captions
- HUMAN: compressed by humans (See Section 4.6.2)
- SEQC-SALIENCY: Dependency based compression, described in this Chapter, with linguistic fluency + saliency-based content selection + dependency constraints (Kuznetsova et al., 2013b)
- SEQC-VISUAL: Dependency based compression (Kuznetsova et al., 2013b) with linguistic fluency + visually-guided content selection + dependency constraints
- $X$  w/o CONSTR: method  $x$  without dependency constraints
- SEQC-NGRAM-ONLY: Dependency based compression (Kuznetsova et al., 2013b) with linguistic fluency only

Recall, the we also experiment with two different ngram statistics, one extracted from the Google Web 1T corpus (Brants and Franz., 2006), and the other computed from the 1M image-caption corpus (Ordonez et al., 2011). Unless otherwise stated, by default, we use image caption corpus statistics.

## 4.6.2 Human-Generalized Captions

To give a notion of upper bound, we ask AMT users (turkers) to write generalized captions for 500 images, then evaluate against two tasks below. Note that for the first evaluation, the retrieved images are still based on computer vision system, and only the part that involves textual rewriting is done by humans.

## 4.6.3 Intrinsic Human Evaluation: Forced Choice

AMT<sup>4</sup> users are provided with an image and two captions (produced by different methods) and are asked to select a better one, i.e., the most relevant and plausible caption that contains the least extraneous information. Results are shown in Table 4.2. We observe that SEQC-VISUAL (full model with visually guided content selection) performs the best, being selected over SEQC-SALIENCY (content-selection without visual information) in 72.48% cases, and *even over the original image caption in 81.75% cases.*

This forced-selection experiment between SEQC-VISUAL and ORIG demonstrates the degree of noise prevalent in the image captions in the wild. Of course, if compared against human-compressed captions, the automatic captions are preferred much less frequently – in 19% of the cases. In those 19% cases when automatic captions are preferred over human-compressed ones, it is sometimes that humans did not fully remove information that is not visually present or verifiable, and other times humans overly compressed. To verify the utility of dependency-based constraints, we also compare two variations of SEQC-VISUAL, with and without dependency-based constraints. As expected, the algorithm with constraints is preferred in the majority of cases.

---

<sup>4</sup>Recall, that we use Amazon Mechanical Turk for all human evaluation settings

Method-1	Method-2	Method-1 preferred (%)
SEQC-SALIENCY	ORIG	76
SEQC-VISUAL	ORIG	82
SEQC-VISUAL	SEQC-SALIENCY	72
SEQC-VISUAL	SEQC-VISUAL w/o CONSTR	84
SEQC-VISUAL	SEQC-NGRAM-ONLY	90
SEQC-VISUAL	HUMAN	19

**Table 4.2:** Intrinsic Human Evaluation of Generalized Captions: posed as a binary question “*which of the two options is better?*” with respect to *Relevance*. We show images for each question. According to Pearson’s  $\chi^2$  test, all results are statistically significant.

#### 4.6.4 Extrinsic Evaluation

To verify the usefulness of generalized image captions, we demonstrate its applicability in relation to common end-user applications, such as image caption transfer task (Ordonez et al., 2011). We use a test set consisting of 1000 images and their associated captions randomly selected from the captioned image database (Ordonez et al., 2011). We apply automatic evaluation method (BLEU) to assess differences in performance using our newly generalized captions versus the original owner provided captions. To control visual parameters, such a visual synonyms (e.g., “cat” and “kitten”, “boat” and “yacht”, etc.), we experiment with our own implementation of BLEU score. We use two BLEU settings, one with strict matching and one with semantic matching. The first one counts word matches only if the words are the same<sup>5</sup>. The second one takes into account visual synonyms and WordNet similarity between the words (Fellbaum, 1998), i.e. if for a word in the candidate sentence no matches are found in the reference (gold standard), a WordNet similarity score is added to the number of matches instead of 1.

In this configuration, we evaluate the usefulness of our new image-text parallel corpus for automatic generation of image descriptions. Here the task is to produce, for a query image, a relevant natural language description (a visually descriptive caption).

For brevity we show the results only for DP with hard dependency constraints as DP with

---

<sup>5</sup>BLEU is very similar to the precision measure, except it uses a few enhancements. It depends on the counts of word matches between the candidate and the reference (gold standard)





Figure 4.8: Example Image Caption Transfer

Method	LM Corpus	strict matching				semantic matching			
		BLEU	P	R	F	BLEU	P	R	F
ORIG	N/A	0.063	0.064	<b>0.139</b>	<b>0.080</b>	0.215	0.220	<b>0.508</b>	0.276
SALIENCY	Image	0.060	0.074	0.077	0.068	0.302	0.411	0.399	0.356
VISUAL	Image	0.060	<b>0.075</b>	0.075	0.068	<b>0.305</b>	<b>0.422</b>	0.397	<b>0.360</b>
SALIENCY	Google	0.064	0.070	0.101	0.074	0.286	0.337	0.459	0.340
VISUAL	Google	<b>0.065</b>	0.071	0.098	0.075	0.296	0.354	0.457	0.350

Table 4.3: Image Description Transfer: performance in BLEU and F1 with *strict* & *semantic* matching. P, R and F stand for Precision, Recall and F1 score respectively

soft constraints did not perform better.

We take a non-parametric approach to image description generation – the global matching based approach proposed in (Ordonez et al., 2011) which demonstrated the power of collecting a large image-caption paired dataset for the challenging task of image description. The outline of the approach is shown in (Figure 4.8). A query image is captioned by finding the most similar image within a large paired image-caption database and then simply transferring the caption associated with the closest database image to the query image.<sup>6</sup>

Results are shown in Table 4.3, demonstrating that our newly generalized captions (rows 2-3) produce better results than using the original database (row 1).

<sup>6</sup>Image similarity is computed using two global (whole) image descriptors. The first is the gist feature (Oliva and Torralba, 2001), an image descriptor related to perceptual characteristics of scenes – naturalness, roughness, openness, etc. The second descriptor is also a global image descriptor, computed by resizing the image into a “tiny image” (Torralba et al., 2008), which is effective in matching the structure and overall color of images. To find visually relevant images, we compute the similarity of the query image to images in the whole dataset using an unweighted sum of gist similarity and tiny image similarity.



**Orig:** Huge wall of glass at the Conference Centre in Yohohama Japan.

**SeqC-Visual:** Huge wall of glass.



**Orig:** Dix Stadium opened in 1969 and is the home football field for the Kent State Golden Flashes in Kent, Oh.

**SeqC-Visual:** Stadium opened and is the home football field.



**Orig:** Vancouver, British Columbia, Canada. A bridge which crosses over the World. Incredible men work through the forest.

**SeqC-Visual:** A bridge which crosses. Men work through the forest.



**Orig:** Yawning yellow dog on a Ganges River ghat in Varanasi, India.

**SeqC-Visual:** Yawning yellow dog.



**Orig:** A small little elephant fellow alone in a big world on the main road toi Skukuza in the Kruger Park.

**SeqC-Visual:** A little elephant fellow alone in a world on the main road.

**Figure 4.9:** Good (shown in blue font, underlined) of Generalized Captions

## 4.7 Discussion

We provide a few good and bad examples of the captions generalized by our algorithm in Figures 4.9 and 4.10. Good examples show that generalization is a very promising direction. Our approach was able to remove extraneous information, such as “at the Conference Centre in Yohohama Japan”, “in 1969”, etc.

However, there are still bad examples, where our approach failed (Figure 4.10). For





<i>Generalization failure</i>		<p><b>Orig:</b> Our house and car in Washington.  <b>SeqC-Visual:</b> <u>Our</u> house and car.</p>
		<p><b>Orig:</b> The world's most powerful lighthouse sitting beside the house with the world's thickest curtains.  <b>SeqC-Visual:</b> <u>Sitting beside</u> the house</p>
		<p><b>Orig:</b> World famous tower bridge in London  <b>SeqC-Visual:</b> Tower bridge <u>in London.</u></p>
<i>Semantically odd</i>		<p><b>Orig:</b> This dog is running near an old barn. An owl used to live in there.  <b>SeqC-Visual:</b> <u>This dog is running near an old barn . An owl used to live.</u></p>
		<p><b>Orig:</b> Sophia at Jeremy's desk in his office.  <b>SeqC-Visual:</b> <u>Sophia in his office.</u></p>
<i>Low compression ratio</i>		<p><b>Orig:</b> Colourful bird at our campsite. When others went swimming I followed birds around with my camera.  <b>SeqC-Visual:</b> <u>Colourful bird . When others went swimming I followed birds around with my camera.</u></p>

Figure 4.10: Bad (shown in red underlined font) of Generalized Captions

example, it does not always remove the Proper Names, such as “London” or introduces semantic issues: “This dog is running near an old barn. An owl used to live.”. Latter problem is mainly due to the fact that some of the descriptions are already so noisy, that it is hard to improve them. Additionally, we rely on vision scores, when generalizing captions. Computer Vision techniques are noisy by themselves.

In the next Chapter we experiment with a PCFG tree driven approach as opposed to dependency-based and find that this method improves the generalized captions. For the next approach we additionally use the statistics collected from human-generalized captions. This statistics helps us to learn which pieces of text we need to delete, such as proper names.

For the next method we also address grammaticality of the compressed sentence in a more principal way rather than hand-coded constraints exploited in this Chapter.

## CKY-BASED TREE-DRIVEN CAPTION GENERALIZATION

## 5.1 Overview

Recall, that task of *image caption generalization* introduced in previous Chapter 4 (Kuznetsova et al., 2013b) had a goal to produce a cleaner image caption dataset by removing extraneous information from the image captions.

At the core of the image caption generalization task is sentence compression. In this Chapter we cast this task as tree compression with lightweight CKY parsing, in conjunction with several other considerations such as visually guided content selection and leaf-level ngram cohesion scores (Kuznetsova et al., 2014).

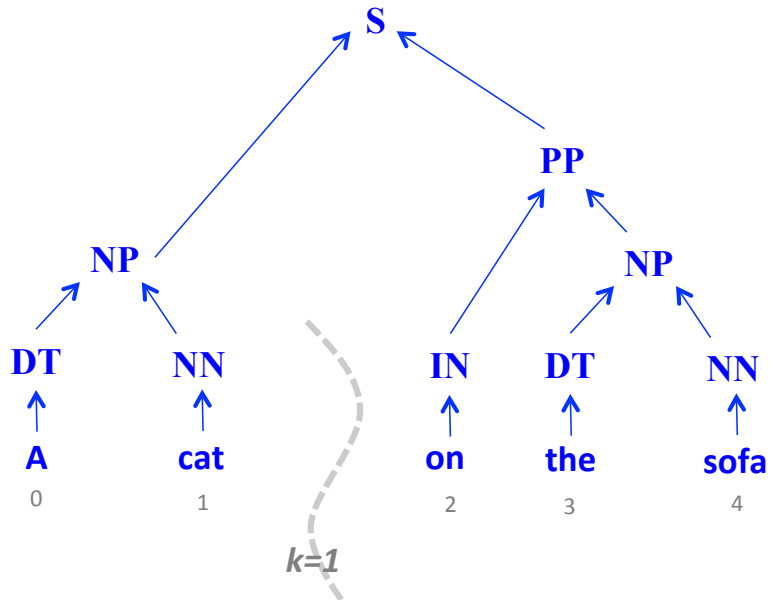
At a high-level, the compression operation resembles bottom-up CKY-parsing, but in addition to parsing, we also consider deletion of parts of the trees. When deleting parts of the original tree, we might need to re-parse the remainder of the tree. Note that we consider re-parsing only with respect to the original parse tree produced by a state-of-the-art parser, hence it is only a *light-weight* parsing.<sup>1</sup>

The approach has some connections to shift-reduce-drop idea of Knight and Marcu (2000), who adopted parsing technique to sentence compression. Our motivation is to tune model specific to caption generalization task as opposed to solving general compression problem from global perspective. The specific designed model helps us to control grammaticality of the caption in a more principal way as opposed to the method described in Chapter 4, which is based on hand-coded constraints.

We consider both tree- and string-based scores directly in the objective function, along with content-selection scores, without involving a feature-vector discriminative classifier internally, and find the plausible solution using dynamic programming. However we stress out that we

---

<sup>1</sup>Integrating full parsing into the original sentence would be a straightforward extension conceptually, but may not be an empirically better choice when parsing for compression is based on vanilla unlexicalized parsing.



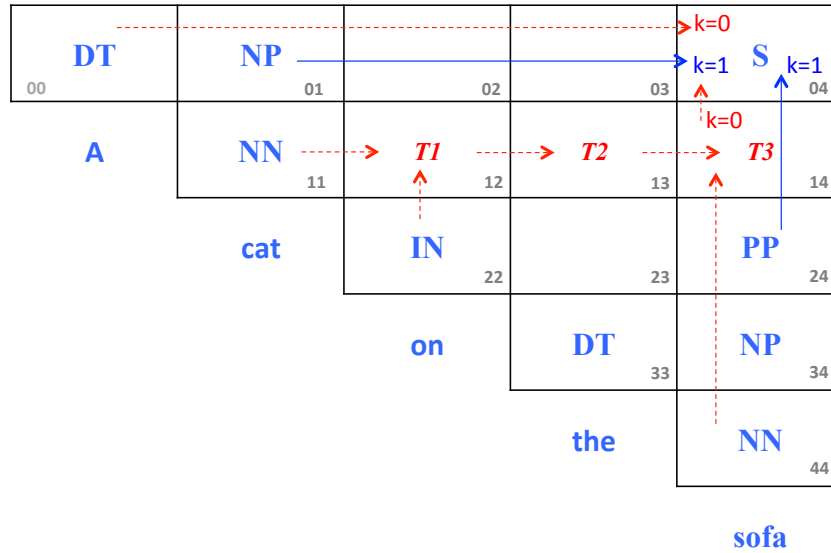
**Figure 5.1:** CKY Parsing Tree.

do not find global optimum for ngram and tree aspects of the objective, as we consider string-based scores in the scope of two branches<sup>2</sup>. In addition, our method performs a light-weight parsing on the fly based on PCFG rules.

Here as well as in Chapter 4 we do not consider any operations beyond deletion. In future, however, we can explore word reordering.

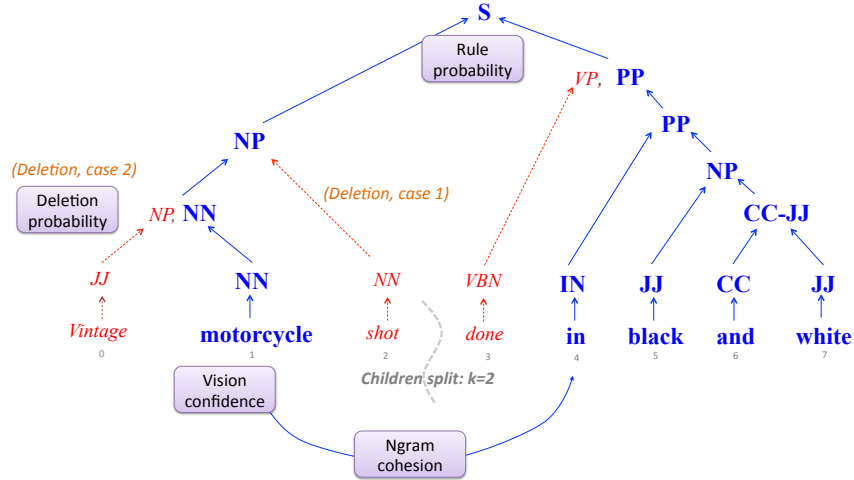
**CKY-parsing** CKY-parsing (Cocke (1969), Kasami (1965), Younger (1967)) is a Dynamic Programming based approach, which reused parses from smaller sub-sequences of the target string. Given string  $X$ , the algorithm finds sub-solutions for all sub-strings  $X[i, j]$ . Parses of larger sub-strings depend on the parses of smaller ones. For each sub-string  $X[i, j]$  algorithm tries all possible splits into children  $X[i, k]$  and  $X[k, j]$ , where two latter problems are already solved. Given a set of PCFG rules, algorithm determined the best rule to apply among all the splits. For example, in Figure 5.1, for the final tree, CKY algorithm split string “A cat on the sofa” into two tree branches “A cat” and “on the sofa”. Parsing task for each

<sup>2</sup>We also do not store solutions corresponding to each compressed subsequence of a branch, we rely only on the tags of the compressed branches



**Figure 5.2:** CKY Parsing Matrix. Both the chosen rules (blue bold font and blue solid arrows) and not chosen rules (red italic smaller font and red dashed lines) are shown.

of the branches was performed separately. As most of Dynamic Programming approaches, CKY algorithm is easier to view as a matrix (Figure 5.2), where each cell  $ij$  corresponds to a subsequence  $X[i, j]$ . Solutions are found by moving diagonally. At each cell, algorithms tries possible splits into children  $k \in [0, i)$  and chose the best PCFG rule to apply, given a set of possible tags found for the cells  $X[i, k]$  and  $X[k, j]$ . For example, in Figure 5.2, cell 04 corresponds to the whole string  $X$  and solution with split  $k = 1$  and rule  $S \rightarrow NP PP$  is found to be the best.



**Figure 5.3:** CKY compression. Both the chosen rules and phrases (blue bold font and blue solid arrows) and not chosen rules and phrases (red italic smaller font and red dashed lines) are shown.

## 5.2 Problem Formulation

Figure 5.3 shows an example compression, and Figure 5.4 shows the corresponding CKY matrix.

Input to the algorithm is a sentence, represented as a vector  $\mathbf{x} = x_0 \dots x_{n-1} = x[0 : n - 1]$ , and its PCFG parse  $\pi(\mathbf{x})$  obtained from the Stanford parser. For simplicity of notation, we assume that both the parse tree and the word sequence are encoded in  $\mathbf{x}$ . Then, the compression can be formalized as:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} \prod_i \phi_i(\mathbf{x}, \mathbf{y}) \quad (5.1)$$

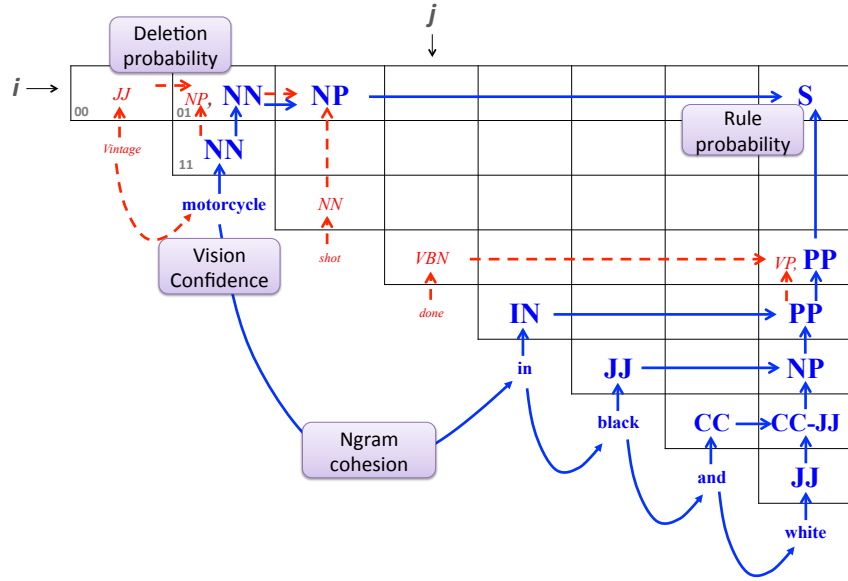
Where each  $\phi_i$  is a potential function, corresponding to a criteria of the desired compression:

$$\phi_i(\mathbf{x}, \mathbf{y}) = \exp(\theta_i \cdot f_i(\mathbf{x}, \mathbf{y})) \quad (5.2)$$

Where  $\theta_i$  is the weight for a particular criteria (described in Section 5.3), whose scoring function is  $f_i$ .

We solve the decoding problem (Equation 5.1) using dynamic programming. For this, we need to solve the compression sub-problems for sequences  $x[i : j]$ , which can be viewed as





**Figure 5.4:** CKY compression. Both the chosen rules and phrases (blue bold font and blue solid arrows) and not chosen rules and phrases (red italic smaller font and red dashed lines) are shown.

branches  $\hat{y}[i, j]$  of the final tree  $\hat{y}[0 : n - 1]$ . For example, in Figure 5.3, the final solution is  $\hat{y}[0 : 7]$ , while a sub-solution of  $x[4 : 7]$  corresponds to a tree branch  $PP$ . Notice that sub-solution  $\hat{y}[3 : 7]$  represents the same branch as  $\hat{y}[4 : 7]$  due to word deletion. Some computed branches, e.g.,  $\hat{y}[1 : 4]$  in Figure 5.3, get dropped from the final compressed tree.

We define a matrix of scores  $\Phi[i, j, h]$  (Equation 5.3), where  $h$  is one of the non-terminal symbols being considered for a cell indexed by  $i, j$ , i.e. a candidate for the root symbol of a branch  $\hat{y}[i : j]$ , and  $R_h = \{r \in R : r = h \rightarrow pq \vee r = h \rightarrow p\}$ . Eventually, when all values  $\Phi[i, j, h]$  are computed, we take

$$\hat{h} = \arg \max_h \Phi[0, n - 1, h]$$

and backtrack to reconstruct the final compression (the exact solution to equation 5.1).

$$\Phi[i, j, h] = \max_{\substack{k \in [i, j] \\ r \in R_h}} \left\{ \begin{array}{l} (1) \quad \Phi[i, k, p] + \Phi[k + 1, j, q] \\ \quad \quad \quad + \Delta\phi[r, ij] \\ (2) \quad \Phi[i, k, p] + \Delta\phi[r, ij] \\ (3) \quad \Phi[k + 1, j, p] + \Delta\phi[r, ij] \end{array} \right. \quad (5.3)$$

The three cases ((1) – (3)) above equation correspond to the following tree pruning cases:

### 5.2.1 Pruning Case (1):

None of the children of the current node is deleted. Index  $k$  determines a split point for child branches of a sub-tree  $ij$ . For example, in Figures 5.3 and 5.4, the PCFG rule  $PP \rightarrow IN PP$ , corresponding to sequence “*in black and white*”, is retained and The split point for children of the sub-tree is  $k = 4$ . Another situation that can be encountered is tree re-parsing, i.e. choosing an alternative rule from original.

### 5.2.2 Pruning Case (2)/(3):

Deletion of the left/right child respectively. There are two types of deletion, as illustrated in Figures 5.3 and 5.4. The first corresponds to deletion of a child node. For example, the second child  $NN$  of rule  $NP \rightarrow NP NN$  is deleted, which yields deletion of “*shot*”.

The second type is a special case of propagating a node to a higher-level of the tree. For example, in Figure 5.4, this situation occurs when deleting  $JJ$  “*Vintage*”, which causes the propagation of  $NN$  from cell 11 to cell 01. For this purpose, we expand the set of rules  $R$  with additional special rules of the form  $h \rightarrow h$ , e.g.,  $NN \rightarrow NN$ , which allows propagation of tree nodes to higher levels of the compressed tree.<sup>3</sup>

<sup>3</sup>We assign probabilities of these special propagation rules to 1 so that they will not affect the final parse tree score. Turner and Charniak (2005) handled propagation cases similarly.

## 5.3 Modelling Compression Criteria

The  $\Delta\phi$  term<sup>4</sup> in Equation 5.3 denotes the sum of log of potential functions for each criteria  $i$ :

$$\Delta\phi[r, ij] = \sum_i \theta \cdot \Delta f(r, ij) \quad (5.4)$$

Note that  $\Delta\phi$  depends on the current rule  $r$  under consideration, along with the historical information before the current step  $ij$ , such as the original rule  $r_{ij}$ , which can be re-parsed, and ngrams on the border between left and right child branches of rule  $r_{ij}$ . We use the following four criteria  $f_i$  in our model, which are also demonstrated in Figures 5.3 and 5.4.

### 5.3.1 I. Tree Structure:

We capture PCFG rule probabilities estimated from the corpus as:

$$\Delta f_{pcfg} = \log P_{pcfg}(r) \quad (5.5)$$

### 5.3.2 II. Sequence Structure:

We incorporate ngram cohesion scores only across the border between two branches of a sub-tree.

$$\Delta f_{ngr} = \log P_{ngr}(\mathbf{y}_{ij}^1, \mathbf{y}_{ij}^2) \quad (5.6)$$

---

<sup>4</sup>We use  $\Delta$  to distinguish the potential value for the whole sentence from the gain of the potential during a single step of the algorithm.

### 5.3.3 III. Branch Deletion Probabilities:

We compute probabilities of deletion for children as:

$$\Delta f_{del} = \log P(r_t|r_{ij}) = \log \frac{\text{count}(r_t, r_{ij})}{\text{count}(r_{ij})} \quad (5.7)$$

Where  $\text{count}(r_t, r_{ij})$  is the frequency in which  $r_{ij}$  is transformed to  $r_t$  by deletion of one of the children. We estimate this probability from a training corpus, described in Section ??.

$\text{count}(r_{ij})$  is the count of  $r_{ij}$  in uncompressed sentences.

### 5.3.4 IV. Vision Detection (Content Selection):

We want to keep words referring to actual objects in the image. Thus, we use  $V(x_j)$ , detection score, as our confidence of an object corresponding to word  $x_j$ . Detection score is obtained using the hedging technique from (Deng et al., 2012).

$$\Delta f_{vis} = \begin{cases} V(x_j) & \text{if } i = j \\ 0 & \text{o/w} \end{cases} \quad (5.8)$$

For all probabilities values we use log. This is also implicitly reflected in equation 5.3 through summation of scores as opposed to multiplication.

Note that some test instances include rules that we have not observed during training. We default to the original caption in those cases. The weights  $\theta_i$  are set using a tuning dataset, e.g., we control over-compression by setting the weight for  $f_{del}$  (described below) to a small value relative to the other weights.

## 5.4 Human Compressed Captions

Although we model image caption generalization as sentence compression, in practical applications we may want the outputs of these two tasks to be different. For example, there may

be differences in what should be deleted, (named entities in newswire summary could be important content to keep, while they may be extraneous for image caption generalization). To learn the syntactic patterns for caption generalization, we collect a small set of example compressed captions (380 in total) using AMT<sup>5</sup>. For each image, we asked 3 turkers to first list all the visible objects in the given image and then to write a compressed caption by removing parts of the caption not visually verifiable from the image content. We then align the original and compressed captions to measure rule deletion probabilities, excluding those pairs with misalignment, similar to Knight and Marcu (2000). Note that we remove this dataset from the 1M caption corpus when we extract phrases for description generation.

## 5.5 Discussion of the Method

Our method can be further improved by learning its parameters through machine learning techniques. Additionally, as for (Knight and Marcu, 2000), with slight modifications our method can produce a set of summaries, instead of a single summary for other systems to choose from. For, example if we return the whole CKY-matrix, another system, can choose between summaries of various length and top tags. We would have to add another dimension in CKY matrix, corresponding to local compression length, similar to modification in Dynamic Programming algorithm, described by McDonald (2006). Also another approach, which is base on noisy-channel model, presented in Knight and Marcu (2000) performed better then their shift-reduce-drop summarization strategy. We did not perform any comparison and did not try other methods except the one explored in Chapter 5. However, in this Chapter, on the other hand, a method based on parsing together with summarization in parallel performed well enough for our task. Plus, this model potentially allows more flexibility on the output parse tree. This way we strive to take into account cases of generalization, where original compression tree is more substantially modified rather then by simple

---

<sup>5</sup>again recall, Amazon Mechanical Turk

deletion of branches. This makes generalization task different from simple compression<sup>6</sup>. Further improvement on the output for generalization task can involve allowance of other operation besides deletions, for instance, substitutions or even insertions.

It would be interesting to compare how our formulation would compare with above mentioned alternatives for sentence compression, but such investigation is clearly beyond the scope of this work, hence we leave it as future direction.

---

<sup>6</sup>here we mean simplified sentence compression as opposed to summarization in general, as summarization can involve very complex tree modifications

## 5.6 Evaluation of Tree-driven Caption Generalization

Here we aim to evaluate usefulness of our second approach to *image caption generalization*. In Chapter 4 Section 4.6 we saw that the task is promising and produces much cleaner captions as opposed to the original ones. We, however, did not probe generalized captions in the phrase composition based description generation task. Here we show that our second compression method is intrinsically better than the one described in Chapter 4 and show extrinsic advantages of the new generalization technique by applying generalized captions in image description generation task, described in Chapter 3.

### 5.6.1 Methods for Compression:

- SEQC-VISUAL (Kuznetsova et al., 2013b): Method, described in Chapter 4. Inference for the objective function operates over the sequence structure. Although optimization is subject to constraints derived from dependency parse, parsing is not an explicit part of the inference structure. (Dependency based compression (Kuznetsova et al., 2013b) with linguistic fluency + visually-guided content selection + dependency constraints). We use language statistics, collected from image caption corpus.
- TREEPRUNING Method, described in this Chapter, PCFG parse tree-driven compression.
- HUMAN Human compressed captions (around 100) separate from Section 5.4.

### 5.6.2 Intrinsic Human Evaluation: Forced Choice

AMT users (turkers) are provided with an image and two captions (produced by different methods) and are asked to select a better one, i.e., the most relevant and plausible caption that contains the least extraneous information. Results are shown in Table 5.1. The agreement among turkers is a frequent concern. Therefore, we vary the set of dependable users based on their Cohen’s kappa score ( $\kappa$ ) against other users. It turns out, filtering users based

on  $\kappa$  does not make a big difference in determining the winning method.

We observe that TREEPRUNING was selected over SEQC-VISUAL (content-selection without visual information) in 65-66% cases.

Method-1	Method-2	Method-1 preferred (%)		
		all turkers	turkers w/ $\kappa > 0.55$	turkers w/ $\kappa > 0.6$
TREEPRUNING	SEQC-VISUAL	65	65	66
TREEPRUNING	HUMAN	20	-	-

**Table 5.1:** Intrinsic Human Evaluation of Generalized Captions: posed as a binary question “*which of the two options is better?*” with respect to *Relevance*. We show images for each question. According to Pearson’s  $\chi^2$  test, all results are statistically significant.

### 5.6.3 Extrinsic Evaluation: Image Caption

#### Generation via Phrase-based composition

Here we apply generalized captions to image description generation task, described in Chapter 3. We use the 1M captioned image corpus of Ordonez et al. (2011). Out of 1M captions we select 1K test images and generate image description for them using the rest of the images for phrase extraction.

We experiment with the following approaches<sup>7</sup>:

#### Proposed Approaches:

- SEQ.V.2+LINGRULE+COGN: an approach, described in Chapter 3. It uses linguistically motivated constraints and cognitive phrases.
- SEQ: Our sequence-driven composition approach as described in Chapter 3, but it does not use some of the linguistically motivated constraints and exploits a slightly different ngram statistics(see Table A.1).
- SEQ+PRUNING: SEQ applied to TREEPRUNING.

<sup>7</sup>For the more principal description of approaches see Table A.1



Method	Bleu	Meteor		
		P	R	M
SEQ.v.2+LINGRULE	0.1518	0.130	0.170	0.095
SEQ	0.1375	0.117	<b>0.184</b>	0.094
SEQ+PRUNING	<b>0.1772</b>	0.153	0.156	<b>0.101</b>

**Table 5.2:** Extrinsic Automatic Evaluation of Generalized Captions

Method-1	Method-2	Method-1 preferred (%)		
		all turkers	turkers w/ $\kappa > 0.55$	turkers w/ $\kappa > 0.6$
SEQ+PRUNING	SEQ	58	58	57

**Table 5.3:** Extrinsic Human Evaluation of Generalized Captions: posed as a binary question “*which of the two options is better?*” with respect to *Relevance*. We show images for each question. According to Pearson’s  $\chi^2$  test, all results are statistically significant.

For SEQ we do not apply some of the linguistic rules, because our goal in this Section is to show usefulness of generalized captions, rather than concentrating on grammaticality of the generated descriptions. LINGRULE were mainly used to encourage natural phrase ordering in consistence with English grammar. Furthermore, Section 3.8 demonstrates that even with those linguistic rules/constraints, system does make mistakes. In Chapter 6 we propose an approach to resolve these mistakes in a more principled way. However, this is outside the scope of the current Chapter and Section.

**Automatic Evaluation** We perform automatic evaluation using two measures widely used in machine translation: BLEU (Papineni et al., 2002)<sup>8</sup> and METEOR (Denkowski and Lavie, 2011).<sup>9</sup> We remove all punctuation and convert captions to lower case. We use 1K test images from the captioned image corpus,<sup>10</sup> and assume the original captions as the gold standard captions to compare against. The results in Table 5.2 show that generalizing captions using tree compression (+PRUNING) improve the BLEU score significantly, while improving METEOR only moderately (due to improvement on precision with decrease in recall.)

<sup>8</sup>We use the NIST implementation: <ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a-20091001.tar.gz>

<sup>9</sup>With equal weight between precision and recall in Table 5.2.

<sup>10</sup>Except for those for which image URLs are broken, or CPLEX did not return a solution.

**Human Evaluation: Forced Choice** Neither BLEU nor METEOR directly measures grammatical correctness over long distances and may not correspond perfectly to human judgements. Therefore, we supplement automatic evaluation with human evaluation. We present two options generated from two competing systems, and ask turkers to choose the one that is better with respect to *relevance*. Results are shown in Table 5.3 with 3 turker ratings per image. We filter out turkers based on a control question. We then compute the selection rate (%) of preferring method-1 over method-2. The agreement among turkers is a frequent concern. Therefore, we vary the set of dependable users based on their Cohen’s kappa score ( $\kappa$ ) against other users. It turns out, filtering users based on  $\kappa$  does not make a big difference in determining the winning method. Still, for extrinsic evaluation SEQ+PRUNING won over SEQ showing that image caption generalization helps with description generation.

We perform a more thorough evaluation in the next Chapter 6, which also addresses grammar problems.

## 5.7 Discussion

Figures 5.5 and 5.6 shows good examples of generalized captions. Among good examples we can see improvements over SEQC-VISUAL approach, for example, “Tower bridge” is an improvement over “Tower bridge in London”. However, bad examples show that we still have a room for improvement. One of the problems is SALIENCE. For example, if an image depicts both a cat and a chair and object detectors produced most confident results for a “chair”, compressed caption would most likely drop a cat. A person, on the other hand, would describe a cat, leaving a chair as a secondary object or not mentioning it at all.

Nevertheless, in Figure 5.7 we can see improvements in the image description generation task Figure 5.8 shows problematic SEQ+PRUNING examples. The problem is mainly description grammaticality, which is addressed in the next Chapter 6.



**Orig:** Note the pillows, they match the chair that goes with it, plus the table in the picture is included.

**SeqC-Visual:** The table in the picture.

**TreePruning:** The chair with the table in the picture.



**Orig:** Our house and car in Washington.

**SeqC-Visual:** Our house and car.

**TreePruning:** House and car.



**Orig:** Only in wintertime we see these birds here in the river.

**SeqC-Visual:** See these birds in the river.

**TreePruning:** These birds in the river.



**Orig:** The world's most powerful lighthouse sitting beside the house with the world's thickest curtains.

**SeqC-Visual:** Sitting beside the house

**TreePruning:** Powerful lighthouse beside the house with the curtains.



**Orig:** The cat of my frind, timy, playing in a bag.

**SeqC-Visual:** The cat.

**TreePruning:** The cat in a bag.



**Orig:** World famous tower bridge in London

**SeqC-Visual:** Tower bridge in London.

**TreePruning:** Tower bridge.

**Figure 5.5:** Caption generalization: good examples (blue underlined font).

**Grammar mistakes**



**Orig:** There's something about having 5 trucks parked in front of my house that makes me feel all important-like.

**SeqC-Visual:** Front of my house.

**TreePruning:** Trucks in front my house.

**Relevance problem**



**Orig:** Orange cloud on street light - near Lanakila Street (phone camera).

**SeqC-Visual:** Orange street

**TreePruning:** Phone camera.

**Saliency problem**



**Orig:** Portrait of a cat sitting under a blue wooden chair.

**SeqC-Visual:** Portrait of a cat sitting under a blue chair.

**TreePruning:** Wooden chair.

**Change of meaning**



**Orig:** Black and white graffiti and the skyline of residential building in Shanghai.

**SeqC-Visual:** Black building.

**TreePruning:** Black and white building.

**Semantical dissonance**

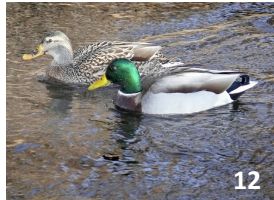


**Orig:** Luna in the back seat looking out the window.

**SeqC-Visual:** In the back seat

**TreePruning:** Back seat looking out the window.

Figure 5.6: Caption generalization: bad examples (red underlined font).



**Seq.v.1+LingRule+Cogn:** Photo of a mandarin duck having a fight over a chip a boy had thrown into the water of the pond.

**Seq.v.2+LingRule+Cogn:** At De Wolfe Point state park in water arrived weeks ago and were sitting around on the snow until the pond thawed these ducks.

**Seq+Pruning:** The duck was having a feast of the pond in golden water.

**Human:** Maybe the most common bird in the neighborhood, not just the most common water fowl in the neighborhood! Ralston Creek Trail, 12-16-09.

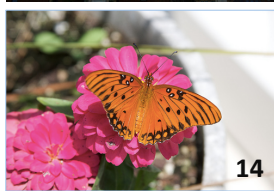


**Seq.v.1+LingRule+Cogn:** This is the view from the clock tower converted into an office along Pacific Coast Highway in Seal Beach CA Approximately 16200 PCH from Mauthausen.

**Seq.v.2+LingRule+Cogn:** View from the top of the clock tower converted into an office along Pacific Coast Highway in Seal Beach CA Approximately 16200 PCH from Mauthausen.

**Seq+Pruning:** Tower in the town.

**Human:** Clock tower in downtown



**Seq.v.1+LingRule+Cogn:** Here you can see the butterflies attracted to the colorful flowers in Hope Gardens.

**Seq.v.2+LingRule+Cogn:** I liked the way into their life attracted to the colorful flowers in Hope Gardens the butterflies.

**Seq+Pruning:** The butterflies are attracted to the colorful flowers to the car.

**Human:** A butterfly on a flower near the Hammocks Clubhouse in Bald Head Island, North Carolina.



**Seq.v.1+LingRule+Cogn:** The flower in a field near Flagstaff dancing with the wind by the road side. The flowers in a field buds under the microscope.

**Seq.v.2+LingRule+Cogn:** Found this flower taken in Madrid March 2006 near Flagstaff. A native flower found in Venezuela.

**Seq+Pruning:** Beautiful flower in the field.

**Human:** Yellow flower near Morava river.

**Figure 5.7:** Examples of SEQ+LINGRULE Descriptions with Extraneous Information and SEQ+PRUNING Descriptions, for which Extraneous Information was (blue font, underlined with a dashed line) and was not (red font underlined) Successfully Removed.



**Seq+Pruning:** Have a unique look in this area across the street from the current building of the cow hand side of the photo a streetlight.

**Human:** Greenpeace guy in the green lamp.



**Seq+Pruning:** Flying above our boat a hawk in the sky.

**Human:** A black headed gull flies low over a lake in Cannon Hill Park, Birmingham, England.



**Seq+Pruning:** Blooming blue flowers in the grass.

**Human:** Blue flower in the shadow.



**Seq+Pruning:** Swimming pool in the summer the duck.

**Human:** A male mallard enjoying his reflection in the water.

Figure 5.8: SEQ+PRUNING Descriptions: Bad Examples (red underlined font).

## TREE-DRIVEN APPROACH TO IMAGE DESCRIPTION GENERATION

## 6.1 Overview

Some of the bad examples of generated descriptions, shown in Chapters 3.7 and 5.6 (Figures 3.11 and 5.8 respectively) introduced grammar problems. Those are due to the fact that we use sequence-driven approach and do not take into account any parse tree structure. Some linguistically motivated constraints and sentence boundary statistics resolved simple problems, such as sentence beginning and end and singular/plural correspondence. However, in order to resolve grammar problems in a more principal way we need to take into account long-distance grammar relations. Latter can be achieved by a tree-driven approach.

Recall, that the high-level idea of our system is to harvest useful bits of text (from this point we will view them as tree fragments) from existing image descriptions using detected visual content similarity, and then to compose a new description by selectively combining these extracted (and optionally pruned via caption generalization) tree fragments. This overall idea of *composition based on extracted phrases* was described in Chapter 3 (Kuznetsova et al., 2012), however, we improve this approach by introducing a parse tree structure into the ILP formulation.

We propose a novel stochastic *tree composition* algorithm based on extracted tree fragments that integrates both tree structure and sequence cohesion into structural inference. Our algorithm permits a substantially higher level of linguistic expressiveness, flexibility, and creativity than those based on rules or templates (e.g., Kulkarni et al. (2011), Yang et al. (2011), Mitchell et al. (2012)), while also addressing long-distance grammatical relations in a more principled way than the approach described in Chapter 3. Our system is driven by both phrase cohesion (Kuznetsova et al. (2012)) and tree structure (Kuznetsova et al. (2014)).



**Figure 6.1:** Harvesting phrases (as tree fragments) for the target image based on (partial) visual match.

## 6.2 Harvesting Tree Fragments

We retrieve tree fragments from captions of matching images in the same way as it was done in Section 3.3. More concretely, as illustrated in Figure 6.1, for a query image visual detection, we extract four types of phrases (as tree fragments). First, from those images with detected objects of the same category, we extract relevant noun phrases. We use color, texture (Leung and J., 1999b), and shape (Dalal and Triggs, 2005; Lowe, 2004b) based features sampled in a spatial pyramid to retrieve visually similar detections. Second, we also extract verb phrases for which the corresponding noun phrase takes the subject role. Visual detection of verbs is still a challenging open problem. We improve on current results by exploiting the semantic relations between noun and verb phrases. Third, from those images with “*stuff*” detections, e.g. “*water*”, or “*sky*” (typically mass nouns), we extract prepositional phrases based on similarity of both visual appearance and relative spatial relationships between detected objects and “*stuff*”. Finally, we use global “*scene*” similarity (L2 distance between classification score vectors (Xiao et al., 2010)) to extract prepositional phrases referring to the overall scene, e.g., “*at the conference*”, “*in the market*”.

We extract phrases for each object detected in a query image and generate one sentence for each object. All sentences are then combined together to produce the final description. Optionally, we apply image caption generalization (via compression) (Section 4.1) to all captions in the corpus prior to the phrase extraction and composition. We use the captioned image corpus of Ordonez et al. (2011) for the phrase extraction and caption compression.



## 6.3 ILP for Tree-driven Composition of Image

### Captions

We model tree composition as constrained optimization. The input to the algorithm is the set of harvested phrases (i.e., tree fragments), as illustrated in Section 6.1. Let  $P = \{p_0, \dots, p_{L-1}\}$  be the set of *all* phrases across four phrase types (objects, actions, stuff and scene). We assume a mapping function  $f : [0, L) \rightarrow T$ , where  $T$  is the set of phrase *types*, so that the phrase type of  $p_i$  is  $f(i)$ . In addition, let  $R$  be the set of PCFG production rules and  $S$  be the set of non-terminal symbols of PCFG. The goal is to find a sequence of phrases  $G$ ,  $|G| \leq |T| = N = 4$ , drawn from  $P$ . I.e., the goal of the algorithm is to select a subset of these phrases (at most one phrase from each phrase type) and reorder them while considering both the parse structure and n-gram cohesion across different phrasal boundaries.

Figure 6.2 shows a simplified example of a composed sentence with its parse structure. For brevity, the figure shows only one phrase for each phrase type, but in actuality there would be a set of candidate phrases for each type. Figure 6.3 shows the CKY-style<sup>1</sup> representation of the internal mechanics of constrained optimization for the example composition shown in Figure 6.2. Each cell  $ij$  of CKY matrix corresponds to  $G_{ij}$ , which is a subsequence of  $G$ , starting at position  $i$  and ending at position  $j$ . If a cell in CKY matrix is labelled with non-terminal symbol  $s$ , it means that a corresponding tree of  $G_{ij}$  has  $s$  as its root.

Although we visualize the operation using CKY-style representation in Figure 6.3, note that composition requires more complex combinatorial decisions than CKY parsing due to two additional considerations: (1) *selecting* a subset of candidate phrases, and (2) *re-ordering* the selected phrases (hence NP-hard). Therefore, we encode our problem using Integer Linear Programming (ILP) (e.g., Roth and Yih (2004), Clarke and Lapata (2008)) and use Cplex (ILOG, Inc, 2006) solver.

We extract phrases for each object detected in a query image and generate one sentence

---

<sup>1</sup>Recall, that CKY parsing overview was given in Section 5.1

for each object. All sentences are then combined together to produce the final description. We also apply image-level content planning, described in Section 3.5.

Similarly to approach, proposed in Chapter 3, we use the captioned image corpus of Ordonez et al. (2011) for the phrase extraction and caption compression.

Figure 6.2 shows a simplified example of a composed sentence with its parse structure. For brevity, the figure shows only one phrase for each phrase type, but in actuality there would be a set of candidate phrases for each type. Figure 6.3 shows the CKY-style representation of the internal mechanics of constrained optimization for the example composition shown in Figure 6.2. Each cell  $ij$  of CKY matrix corresponds to  $G_{ij}$ , which is a subsequence of  $G$ , starting at position  $i$  and ending at position  $j$ . If a cell in CKY matrix is labelled with non-terminal symbol  $s$ , it means that a corresponding tree of  $G_{ij}$  has  $s$  as its root.

Although we visualize the operation using CKY-style representation in Figure 6.3, note that composition requires more complex combinatorial decisions than CKY parsing due to two additional considerations: (1) *selecting* a subset of candidate phrases, and (2) *re-ordering* the selected phrases (hence NP-hard). Therefore, we encode our problem using Integer Linear Programming (ILP) (e.g., Roth and Yih (2004), Clarke and Lapata (2008)) and use Cplex (ILOG, Inc, 2006) solver. In order to reduce running time, we set Cplex parameters to return sub-optimal solution within 30 sec of running time for each description. This allows us to obtain fluent enough descriptions within a reasonable running time.

The process is represented by assignment of a particular tag to a matrix cell. The chosen tag must be a head of a rule, for example cell 01 in Figure 6.3 is being assigned the tag  $NP$ , corresponding to rule  $NP \rightarrow NP PP$ . This rule connects leafs “A cow” and “in the countryside”. The problem is to find tag assignment for each cell of the matrix, given that some cells can be empty. Each cell represents a branch of the tree corresponding to a sub-string of the description. For example, cell 01 correspond to a noun phrase “A cow in the countryside”. If cell is empty, it means that the sub-string cannot be represented as a complete single-rooted tree. For instance, cell 12, representing a sub-string “in the

countryside was staring at me”, corresponds to a couple of branches rather than a single tree. We use technique similar to the one used in CKY parsing approach.

We have two main versions of the algorithm: sequence-based and tree-based. The latter subsumes the former one. They differ in the variables and constraints used as well as a number of options added to the algorithm (Table A.1)

### 6.3.1 Variables and Objective Function

As in Chapter 3, each variable is indexed by a selected object  $o$  (superscript).

**Variables for Sequence Structure:** Variables  $\alpha$  encode phrase selection and ordering:

$$\alpha_{ik}^o = 1 \quad \text{iff} \quad \text{phrase } i \in P \text{ is selected} \quad (6.1)$$

for position  $k \in [0, N)$

Where  $k \in [0, N)$  is one of the  $N=4$  positions in a sentence. We also define variables for each pair of adjacent phrases:

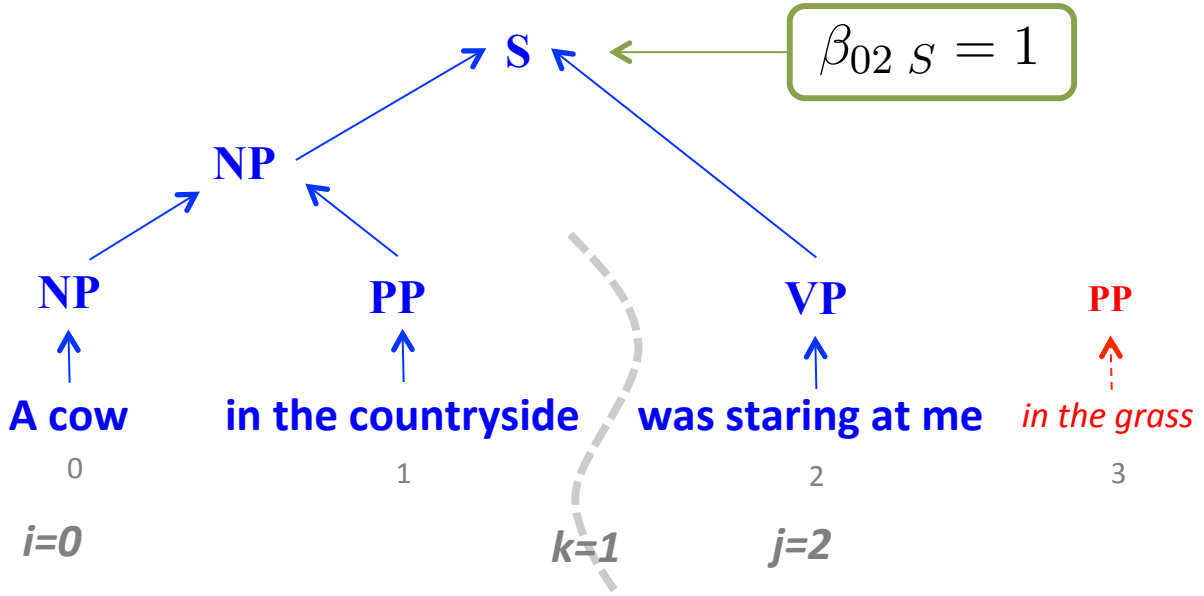
$$\alpha_{ijk}^o = 1 \quad \text{iff} \quad \alpha_{ik}^o = \alpha_{j(k+1)}^o = 1 \quad (6.2)$$

**Variables for Tree Structure:** Variables  $\beta$  encode the parse structure:

$$\beta_{ijs}^o = 1 \quad \text{iff} \quad \text{the phrase sequence } G_{ij} \quad (6.3)$$

maps to the nonterminal symbol  $s \in S$

Where  $i \in [0, N)$  and  $j \in [i, N)$  index rows and columns of the CKY-style matrix in Figure 6.3. A corresponding example tree is shown in Figure 6.2, where the phrase sequence  $G_{02}$  corresponds to the cell labelled with  $S$ . We also define variables to indicate selected PCFG rules in the resulting parse:



**Figure 6.2:** An example scenario of tree composition. Only the first three phrases are chosen for the composition.

$$\begin{aligned} \beta_{ijk_r}^o = 1 \quad \text{iff} \quad & \beta_{ijh}^o = \beta_{ikp}^o \\ & = \beta_{(k+1)jq}^o = 1, \end{aligned} \tag{6.4}$$

Where  $r = h \rightarrow pq \in R$  and  $k \in [i, j)$ . Index  $k$  points to the boundary of split between two children as shown in Figure 6.2 for the sequence  $G_{02}$ .

**Auxiliary Variables:** For notational convenience, we also include:

$$\begin{aligned} \gamma_{ijk}^o = 1 \quad \text{iff} \quad & \sum_{s \in S} \beta_{ijs}^o \\ & = \sum_{s \in S} \beta_{iks}^o \\ & = \sum_{s \in S} \beta_{(k+1)js}^o = 1 \end{aligned} \tag{6.5}$$

We model tree composition as maximization of the following objective function<sup>2</sup>:

<sup>2</sup>Note that we indicate object index  $o$  as a superscript in both, scores and variables

$$\begin{aligned}
F = & \sum_o \left( \sum_i F_i^o \times \sum_{k=0}^{N-1} \alpha_{ik}^o \right. \\
& + \sum_{ij} F_{ij}^o \times \sum_{k=0}^{N-2} \alpha_{ijk}^o \\
& \left. + \sum_{ij} \sum_{k=i}^{j-1} \sum_{r \in R} F_r \times \beta_{ijk_r}^o \right)
\end{aligned} \tag{6.6}$$

This objective is comprised of three types of weights (confidence scores):  $F_i^o, F_{ij}^o, F_r$ .<sup>3</sup>  $F_i^o$  represents the phrase selection score based on visual similarity, described in Section 3.3 of Chapter 3.  $F_{ij}^o$  quantifies the sequence cohesion across phrase boundaries and was described in Section 3.6.4 of Chapter 3. For this, we use  $n$ -gram scores ( $n \in [2, 5]$ ) between adjacent phrases computed using the Google Web 1-T corpus (Brants and Franz., 2006). Finally,  $F_r$  quantifies PCFG rule scores (log probabilities) estimated from the 1M image caption corpus (Ordonez et al., 2011) parsed using Stanford parser (Klein and Manning, 2003).

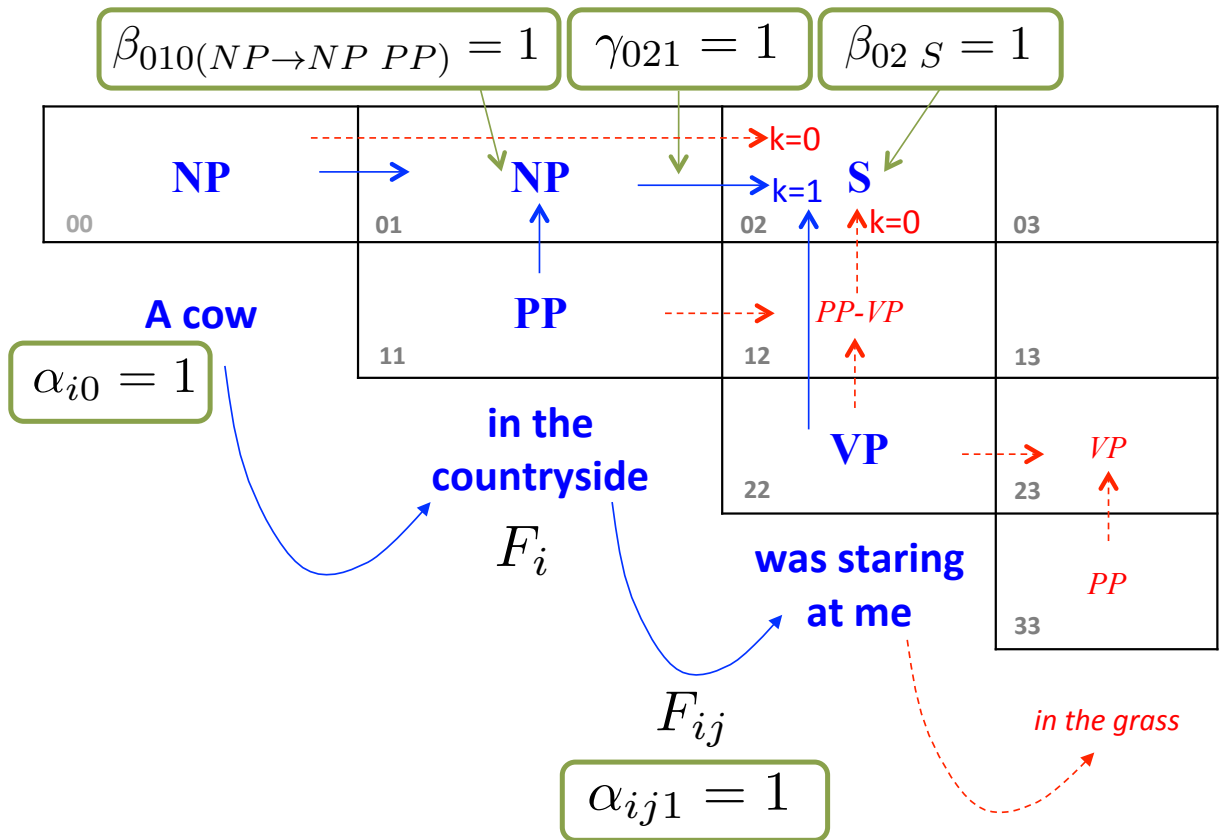
One can view  $F_i^o$  as a *content selection* score, while  $F_{ij}^o$  and  $F_r$  correspond to *linguistic fluency* scores capturing sequence and tree structure respectively. If we set positive values for all of these weights, the optimization function would be biased toward verbose production, since selecting an additional phrase will increase the objective function. To control for verbosity, we set scores corresponding to linguistic fluency, i.e.,  $F_{ij}^o$  and  $F_r$  using negative values (smaller absolute values for higher fluency), to balance dynamics between content selection and linguistic fluency.

Negative scores encourage ILP to variate number of variable assigned to 1. This allows us to avoid overloaded descriptions akin to [A cow ] [ in the countryside] [was staring at me] [in the grass] and generate simpler description, such as [A cow] [ in the countryside] [was staring at me], where not all four types of phrases are selected. Additionally, for this “trick” to work, we have allow empty cells at the top row of tree matrix (For instance cell 03 in Figure 6.3 does not have any tags selected for the final output).

For tree-driven composition we do not prepend the first sentence in a generated description

---

<sup>3</sup>All weights are normalized using z-score.



**Figure 6.3:** CKY-style representation of decision variables as defined in Section ?? for the tree example in Fig 6.2. Non-terminal symbols in boldface (in blue) and solid arrows (also in blue) represent the chosen PCFG rules to combine the selected set of phrases. Non-terminal symbols in smaller font (in red) and dotted arrows (also in red) represent possible other choices that are not selected.

with a *cognitive phrase*, described in Section 3.6.6.

### 6.3.2 Constraints

**Soundness Constraints:** We need constraints to enforce consistency between different types of variables (Equations 6.2, 6.4, 6.5):

$\forall i, j \in P, k \in [0, N),$

$$\forall_{ijk}, \alpha_{ijk} \leq \alpha_{ik} \quad (6.7)$$

$$\alpha_{ijk} \leq \alpha_{j(k+1)}$$

$$\alpha_{ijk} + (1 - \alpha_{ik}) + (1 - \alpha_{j(k+1)}) \geq 1$$

$\forall i \in [0, N), j \in [i + 1, N), k \in [i, j), r = h \rightarrow pq, r \in R,$

$$\beta_{ijk}^o \leq \beta_{ijh}^o \quad (6.8)$$

$$\beta_{ijk}^o \leq \beta_{ikp}^o$$

$$\beta_{ijk}^o \leq \beta_{(k+1)jq}^o$$

$$\beta_{ijk}^o + \quad (6.9)$$

$$(1 - \beta_{ijh}^o) +$$

$$(1 - \beta_{ikp}^o) +$$

$$(1 - \beta_{(k+1)jq}^o) \geq 1$$

Similarly we ensure that  $\gamma_{ijk}^o = \sum_{s \in S} \beta_{ijs}^o \cdot \sum_{s \in S} \beta_{iks} \cdot \sum_{s \in S} \beta_{(k+1)js}$

$\forall i \in [0, N), j \in [i + 1, N), k \in [i, j),$

$$\gamma_{ijk}^o \leq \sum_{s \in S} \beta_{ijs}^o \quad (6.10)$$

$$\gamma_{ijk}^o \leq \sum_{s \in S} \beta_{iks}$$

$$\gamma_{ijk}^o \leq \sum_{s \in S} \beta_{(k+1)js}$$

$$\gamma_{ijk}^o + \quad (6.11)$$

$$(1 - \sum_{s \in S} \beta_{ijs}^o) +$$

$$(1 - \sum_{s \in S} \sum_{s \in S} \beta_{iks}) +$$

$$(1 - \sum_{s \in S} \sum_{s \in S} \beta_{(k+1)js}) \geq 1$$

Note that  $\beta_{ijk}^o$  and  $\gamma_{ijk}^o$  are defined only for non-diagonal cells of CKY matrix.

**Consistency between Tree Leafs and Sequences:** In Figure 6.3 we can see both, the tree matrix, constructed by the algorithm ( $\beta$  variables), and chosen phrases themselves, i.e. “a cow”, “in the countryside”, etc. ( $\alpha$  variables). If we just optimize objective given in equation 6.6, without any constraints enforcing connection between  $\alpha$  and  $\beta$  variables, we can face a situation when the leafs of the tree are  $NP$ ,  $PP$ ,  $VP$ , but the phrase choice and ordering, independent from the tree, is “in the countryside”, “a cow”. Latter doe not correspond to sequence  $NP$ ,  $PP$ ,  $VP$ , which means the constructed tree does not represent the generated description. We need to connect phrases with their non-terminal symbol and the leafs of the resulting parse tree. This, the ordering of phrases implied by  $\alpha^o$  variables must be consistent with the ordering of phrases implied by the  $\beta^o$  variables. This can be achieved by aligning the leaf cells (i.e.,  $\beta_{kks}$ ) in the CKY-style matrix with  $\alpha_{ik}^o$  variables as follows:

$$\forall_{ik}, \alpha_{ik}^o \leq \sum_{s \in S^i} \beta_{kks}^o \quad (6.12)$$

$$\forall_k, \sum_i \alpha_{ik}^o = \sum_{s \in S} \beta_{kks}^o \quad (6.13)$$

Where  $S^i$  refers to the set of PCFG non-terminals that are compatible with the phrase type of  $p_i$ . For example,  $S^i = \{NN, NP, \dots\}$  if  $p_i$  corresponds to an “object” (noun-phrase). Thus, Equation 6.12 enforces the correspondence between phrase types and non-terminal symbols at the tree leafs. Equation 6.13 enforces the constraint that the number of selected phrases and instantiated tree leafs must be the same.

**Sequence Congruence Constraints:** To generate informative descriptions for sequence driven ILP, we choose to include at least two phrases for each sentence:

$$\forall s, \sum_{ij} \alpha_{i1}^o = 1 \quad (6.14)$$

$$\forall s, \sum_{ij} \alpha_{i2}^o = 1 \quad (6.15)$$



We require only contiguous slots to be filled:

$$\forall k = 3, \dots, N - 1, \quad \sum_i \alpha_{i(k+1)}^o \leq \sum_i \alpha_{ik}^o \quad (6.16)$$

**Tree Congruence Constraints:** To ensure that each CKY cell has at most one symbol we require

$$\forall ij, \quad \sum_{s \in S} \beta_{ijs}^o \leq 1 \quad (6.17)$$

We also require that

$$\forall i, j > i, h, \quad \beta_{ijh}^o = \sum_{k=i}^{j-1} \sum_{r \in R_h} \beta_{ijk_r}^o \quad (6.18)$$

Where  $R_h = \{r \in R : r = h \rightarrow pq\}$ . We enforce these constraints only for non-leaves. This constraint forbids instantiations where a non-terminal symbol  $h$  is selected for cell  $ij$  without selecting a corresponding PCFG rule.

We also ensure that we produce a valid tree structure. For instance, if we select 3 phrases as shown in Figure 6.3, we must have the root of the tree at the corresponding cell 02.

$$\forall k \in [1, N), \quad \sum_{s \in S} \beta_{kks}^o \leq \sum_{t=k}^{N-1} \sum_{s \in S} \beta_{0ts}^o \quad (6.19)$$

We also require cells that are not selected for the resulting parse structure to be empty:

$$\forall ij \quad \sum_k \gamma_{ijk}^o \leq 1 \quad (6.20)$$

Additionally, we penalize solutions without the  $S$  tag at the parse root as a soft-constraint<sup>4</sup>

**Linguistic and Discourse constraints:** Similarly to the system, described in Chapter 3 we include linguistically motivated constraints. I.e. we enforce a noun-phrase to be selected to ensure semantic relevance to the image. We also allow at most phrase of each type (NP,VP,PPstuff,PPscene). Additionally, we disallow plural (singular) form of a noun be chosen together with singular (plural) form of a verb. We also allow at

<sup>4</sup>We encode soft-constraints as negative terms in the objective function.

most one prepositional scene phrase for the whole description. Finally, we add constraints that prevent the inclusion of more than one phrase with identical head words. All constraints are similar to those, defined in Section 3.6.2. We do not use the constraint, which forbids a VP at the beginning of a description (equation 3.25), as the parse-tree structure should take care of that.

### 6.3.3 Remarks

We find that handling of sentence boundaries is important if the ILP formulation is based only on sequence structure, but with the integration of tree-based structure, we need not handle sentence boundaries (see Table A.1). Although, we experimented with begin-end statistics, it works only in a very small local context and short descriptions, while ILP-CKY formulation is able to address grammar issue on a more global level.

An interesting aspect of description generation explored in this Chapter is that building blocks of composition are tree fragments, rather than individual words. There are three practical benefits: (1) *syntactic and semantic expressiveness*, (2) *correctness*, and (3) *computational efficiency*. Because we extract nice segments from human-written captions, we are able to use expressive language, and less likely to make syntactic or semantic errors. Our phrase extraction process can be viewed at a high level as visually-grounded or visually-situated paraphrasing.

Also, because the unit of operation is tree fragments, the ILP formulation encoded in this work is computationally lightweight. If the unit of composition was words, the ILP instances would be significantly more computationally intensive, and more likely to suffer from grammatical and semantic errors.

## 6.4 Evaluation of Tree-driven Composition Approach

Our goal is to assess usefulness of tree-driven approach. Additionally we also evaluate tree-driven approach, which composes pruned tree fragments against the one that operated over not pruned fragments. Latter is performed in order to make sure that caption generalization helps to improve descriptions generated by a tree-driven approach as well.

As usual, recall, that we use the 1M captioned image corpus of Ordonez et al. (2011). Out of 1M captions we select 1K test images and generate image description for them using the rest of the images for phrase extraction.

We experiment with the following approaches:

- **SEQ.V.2+LINGRULE+COGN** (Kuznetsova et al., 2012): This method performs structural inference for the objective function over the sequence structure and generation is driven by linguistically motivated constraints rather than a parse tree structure (Chapter 3). For a stronger baseline, we include captions with cognitive phrases into evaluation as the phrases help with grammatical aspect of the description <sup>5</sup> and make captions look more human-like. For all other descriptions, we do not use cognitive phrases.
- **SEQ**: System, described in this Chapter, with the tree structure part suppressed (comparable to **SEQ.V.2+LINGRULE+COGN**, but does not use some of the linguistic rules – see Table A.1).
- **SEQ+PRUNING**: SEQ applied to the compressed captions using **TREEPRUNING**.
- **SEQ+TREE**: Tree-driven system, described in this Chapter.
- **SEQ+TREE+PRUNING** SEQ+TREE applied to the compressed captions using **TREEPRUNING**.

---

<sup>5</sup>they encourage a better phrase ordering as we always put such a phrase at the beginning. This forces descriptions to start with more probable for a beginning of a sentence words.

Method	Bleu	Bleu w/o penalty	Meteor		
			P	R	M
SEQ.V.2+LINGRULE+COGN	0.1518	0.1518	0.130	0.170	0.095
SEQ	0.1375	0.1375	0.117	<b>0.184</b>	0.094
SEQ+TREE	0.1492	0.1492	0.126	0.136	0.082
SEQ+PRUNING	<b>0.1772</b>	0.1772	0.153	0.156	<b>0.101</b>
SEQ+TREE+PRUNING	0.1404	<b>0.1892</b>	<b>0.163</b>	0.119	0.088

**Table 6.1:** Automatic Evaluation of Generated Descriptions

## 6.4.1 Automatic Evaluation

We perform automatic evaluation using BLEU (Papineni et al., 2002)<sup>6</sup> and METEOR (Denkowski and Lavie, 2011).<sup>7</sup> As in Section 5.6.3, we remove all punctuation and convert captions to lower case. We use 1K test images from the captioned image corpus,<sup>8</sup> and assume the original captions as the gold standard captions to compare against. The results in Table 6.1 show that using a tree structure improves BLEU without brevity penalty only<sup>9</sup>, perhaps due to shorter generated descriptions, as longer generated sentences are less likely to maintain a proper tree structure.

## 6.4.2 Human Evaluation: Forced Choice

Similarly as in previous Chapters, we supplement automatic evaluation with human evaluation. We present two options generated from two competing systems, and ask turkers to choose the one that is better with respect to: *relevance*, *grammar*, and *overall*. Results are shown in Table 6.2 with 3 turker ratings per image. We filter out turkers based on a control question. We then compute the selection rate (%) of preferring method-1 over method-2. Similarly, as in Chapter 5 Section 5.6.2, we vary the set of dependable users based on their Cohen’s kappa score ( $\kappa$ ) against other users. Again, it turns out, filtering users based on  $\kappa$  does not make a big difference in determining the winning method.

<sup>6</sup>We use the NIST implementation: <ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a-20091001.tar.gz>







<sup>7</sup>With equal weight between precision and recall in Table 6.1.

<sup>8</sup>Except for those for which image URLs are broken, or CPLEX did not return a solution.

<sup>9</sup>penalizes shorter descriptions

Method-1	Method-2	w/Images	Criteria	Method-1 preferred over Method-2 (%)		
				all turkers	turkers w/ $\kappa > 0.55$	turkers w/ $\kappa > 0.6$
SEQ+TREE	SEQ	+	Rel	72	72	72
SEQ+TREE	SEQ	-	Gmar	83	83	83
SEQ+TREE	SEQ	+	All	68	69	66
SEQ+TREE	SEQ.V.2+LINGRULE+COGN	+	All	62	63	62
SEQ+TREE	HUMAN	+	All	17	16	16
SEQ+TREE+PRUNING	SEQ+PRUNING	-	Gmar	76	77	78
SEQ+TREE+PRUNING	SEQ+PRUNING	+	All	73	75	75
SEQ+TREE+PRUNING	SEQ+TREE	+	Rel	68	72	72
SEQ+TREE+PRUNING	SEQ+TREE	-	Gmar	41	38	41
SEQ+TREE+PRUNING	SEQ+TREE	+	All	63	64	66
SEQ+TREE+PRUNING	SEQ.V.2+LINGRULE+COGN	+	All	67	75	77
SEQ+TREE+PRUNING	HUMAN	+	All	24	19	19

**Table 6.2:** Human Evaluation of Generated Descriptions: posed as a binary question “*which of the two options is better?*” with respect to *Relevance* (Rel), *Grammar* (Gmar), and *Overall* (All). We show images for each question. According to Pearson’s  $\chi^2$  test, all results are statistically significant.

Relevance problems	 6	<p><u>Seq.v.1+LingRule+Cogn</u>: This is what happens when a train lost against the mountain background at Horseshoe Curve in April 1986 under a rock.</p> <p><u>Seq.v.2+LingRule+Cogn</u>: This is a train lost against the mountain background at Horseshoe Curve in April 1986 under a rock.</p> <p><u>Seq+Tree</u>: A train is almost lost against the mountain background at Horseshoe Curve in April 1986.</p> <p><u>Seq+Tree+Pruning</u>: The railroad is the mountain background</p> <p>Human: Found this guy right outside my sleeping bag waking up in OKC.</p>
	 7	<p><u>Seq.v.1+LingRule+Cogn</u>: This is a shoulder bag with a blended rainbow effect.</p> <p><u>Seq.v.2+LingRule+Cogn</u>: This is a shoulder bag featuring many vintage buttons in different materials with a blended rainbow effect.</p> <p><u>Seq+Tree</u>: 60 bags featuring many vintage buttons in different materials.</p> <p><u>Seq+Tree+Pruning</u>: 60 bags inspired by game arcade.</p> <p>Human: A softshell turtle spotted during a walk in Ft. Worth, TX USA.</p>
Grammar problems	 8	<p><u>Seq.v.1+LingRule+Cogn</u>: This is a shot of the road fixed to the table the cup.</p> <p><u>Seq.v.2+LingRule+Cogn</u>: One of the road fixed to the table the cup.</p> <p><u>Seq+Tree</u>: The cup is just around the corner.</p> <p><u>Seq+Tree+Pruning</u>: The cup in the middle of the road.</p> <p>Human: Cup by Corning, plate marked Sterling vitrified china, East Liverpool, OH, G-3.</p>
	 9	<p><u>Seq.v.1+LingRule+Cogn</u>: Of apples the butterfly feeding in Judy flower garden by a tree.</p> <p><u>Seq.v.2+LingRule+Cogn</u>: Of apples the butterfly feeding in Judy flower garden by a tree.</p> <p><u>Seq+Tree</u>: This butterfly was on the sidewalk in the middle of a busy downtown street.</p> <p><u>Seq+Tree+Pruning</u>: This butterfly in the grass.</p> <p>Human: At a butterfly house somewhere in North Wales.</p>
Cognitive problems	 10	<p><u>Seq.v.1+LingRule+Cogn</u>: I like the way the clouds walking in the poppy field under cloudy sky.</p> <p><u>Seq.v.2+LingRule+Cogn</u>: One of the clouds walking in the poppy field under a cloudy sky.</p> <p><u>Seq+Tree</u>: The clouds were walking in the poppy field.</p> <p><u>Seq+Tree+Pruning</u>: The clouds were in the sky.</p> <p>Human: A bike in a field dreams of unconventional places.</p>
		<p><u>Seq.v.1+LingRule+Cogn</u>: Here you can see a cross by the frog in the sky.</p> <p><u>Seq.v.2+LingRule+Cogn</u>: This is a cross standing by the frog.</p> <p><u>Seq+Tree</u>: A cross stands guard over the cemetery that sank beneath the ocean during the eruption of Vulcan Daan in 1871.</p> <p><u>Seq+Tree+Pruning</u>: A cross in the sky.</p> <p>Human: Kites were flying all around the Washington monument during the cherry blossom festival. It was beautiful.</p>

**Figure 6.4:** Examples where SEQ+TREE+PRUNING (blue font underlined with a dashed line) improved captions over SEQ+LINGRULE (red font underlined).

## 6.5 Discussion

As expected, tree-based systems significantly outperform sequence-based counterparts. For example, SEQ+TREE is strongly preferred over SEQ, with a selection rate of 83%. Somewhat surprisingly, improved grammaticality seems to also improve relevance scores (72%), possibly because it is harder to appreciate the semantic relevance of automatic captions when they are less comprehensible. Also as expected, compositions based on pruned tree fragments significantly improve relevance (68–72%), while slightly deteriorating grammar (38–41%).



**Seq+Pruning:** Have a unique look in this area across the street from the current building of the cow hand side of the photo a streetlight.

**Seq+Tree+Pruning:** A streetlight is in a government building at the corner.

**Human:** Greenpeace guy in the green lamp.



**Seq+Pruning:** Flying above our boat a hawk in the sky.

**Seq+Tree+Pruning:** The bird flying above our boat.

**Human:** A black headed gull flies low over a lake in Cannon Hill Park, Birmingham, England.



**Seq+Pruning:** Blooming blue flowers in the grass.

**Seq+Tree+Pruning:** Little flowers blooming in the grass.

**Human:** Blue flower in the shadow.



**Seq+Pruning:** Swimming pool in the summer the duck.

**Seq+Tree+Pruning:** The duck sitting in the water.

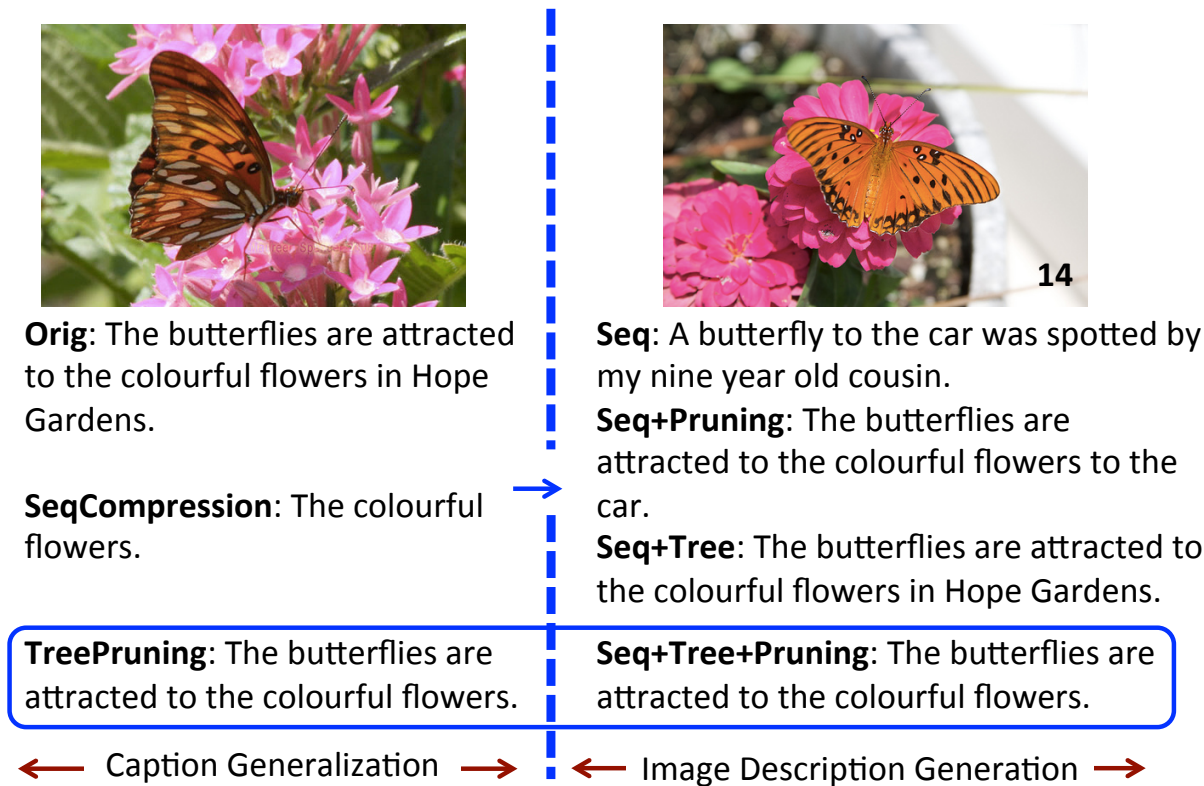
**Human:** A male mallard enjoying his reflection in the water.

**Figure 6.5:** Examples where SEQ+TREE+PRUNING (blue font underlined with a dashed line) improved captions over SEQ+PRUNING (red font underlined).

Notably, the captions generated by SEQ+TREE+PRUNING are preferred over the original (owner generated) captions 19–24% of the time. One such example is included in Figure 6.6: “The butterflies are attracted to the colorful flowers.”.

SEQ+TREE+PRUNING was able to resolve some bad descriptions generated by SEQ+LINGRULE (Figures 3.11 and 6.4). We can see that object detection errors should be resolved by Computer Vision techniques rather than on NLP side.

In Figure 6.5 we can see that SEQ+TREE+PRUNING solved grammar problems for some



**Figure 6.6:** An example of a description preferred over human gold standard. Image description is improved due to caption generalization.

of the problematic descriptions, generated by SEQ+PRUNING (Figure 5.8).

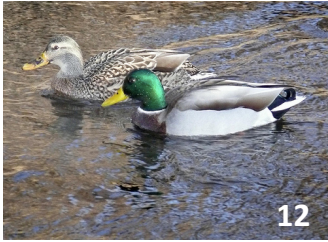
Additional examples (good and bad) are provided in Figures 6.8 and 6.9. Many of these captions are highly expressive while remaining semantically plausible, thanks to the expressive, but somewhat predictable descriptions online users write about their photos. Even among the bad examples (Figure 6.9) one can find highly creative captions with not literal but metaphorical relevance: *“Monarch in her bedroom before the wedding ceremony.”*<sup>10</sup>. More such creative examples are given in Appendix C.

More good examples are provided in Appendix B.

We also provide random 70 images with descriptions generated by various methods in Appendix D.

<sup>10</sup>“Monarch” can be a type of butterfly.





12

**Seq.v.1+LingRule+Cogn:** Photo of a mandarin duck having a fight over a chip a boy had thrown into the water of the pond.

**Seq.v.2+LingRule+Cogn:** At De Wolfe Point state park in water arrived weeks ago and were sitting around on the snow until the pond thawed these ducks.

**Seq+Tree+Pruning:** The duck was having a feast.

**Human:** Maybe the most common bird in the neighborhood, not just the most common water fowl in the neighborhood! Ralston Creek Trail, 12-16-09.



13

**Seq.v.1+LingRule+Cogn:** This is the view from the clock tower converted into an office along Pacific Coast Highway in Seal Beach CA Approximately 16200 PCH from Mauthausen.

**Seq.v.2+LingRule+Cogn:** View from the top of the clock tower converted into an office along Pacific Coast Highway in Seal Beach CA Approximately 16200 PCH from Mauthausen.

**Seq+Tree+Pruning:** Tower in the town.

**Human:** Clock tower in downtown



14

**Seq.v.1+LingRule+Cogn:** Here you can see the butterflies attracted to the colorful flowers in Hope Gardens.

**Seq.v.2+LingRule+Cogn:** I liked the way into their life attracted to the colorful flowers in Hope Gardens the butterflies.

**Seq+Tree+Pruning:** The butterflies are attracted to the colorful flowers.

**Human:** A butterfly on a flower near the Hammocks Clubhouse in Bald Head Island, North Carolina.



15



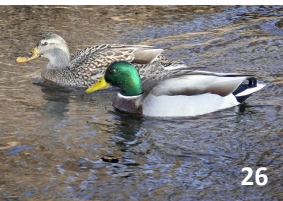



**Seq.v.1+LingRule+Cogn:** The flower in a field near Flagstaff dancing with the wind by the road side. The flowers in a field buds under the microscope.

**Seq.v.2+LingRule+Cogn:** Found this flower taken in Madrid March 2006 near Flagstaff. A native flower found in Venezuela.

**Seq+Tree+Pruning:** Beautiful flower is in the sun. Yellow flower lost in grass.

**Human:** Yellow flower near Morava river.

**Figure 6.7:** SEQ+TREE+PRUNING Descriptions (blue font underlined with a dashed line): Examples, where Generalization Helped to Remove Extraneous Information (red font underlined).

Highly expressive Interesting choice of verb phrases		<p><b>Human:</b> Some flower on a bar in a hotel in Grapevine, TX.</p> <p><b>Seq+Tree+Pruning:</b> The flower was so vivid and attractive.</p>
		<p><b>Human:</b> Sailboat in Waikiki ocean.</p> <p><b>Seq+Tree+Pruning:</b> A small sailboat rests in the water.</p>
		<p><b>Human:</b> Maybe the most common bird in the neighborhood, not just the most common water fowl in the neighborhood!</p> <p><b>Seq+Tree+Pruning:</b> The duck was having a feast.</p>
Poetic		<p><b>Human:</b> Spring in a white dress.</p> <p><b>Seq+Tree+Pruning:</b> Blue flowers have no scent. Small white flowers have no idea what they are.</p>
Informative		<p><b>Human:</b> Tower bridge London in black and white.</p> <p><b>Seq+Tree+Pruning:</b> The tower built on each side.</p>
		<p><b>Human:</b> Shot in Blackpool on Tesco's car park with use of a 5M candle power light.</p> <p><b>Seq+Tree+Pruning:</b> Red car in the middle of the road.</p>

**Figure 6.8:** SEQ+TREE+PRUNING Descriptions: Good Examples. Descriptions preferred over human gold standard are contoured.

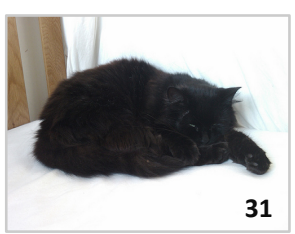

<p><b>Semantic dissonance due to generalization error</b></p>	 <p>30</p>	<p><b>Human:</b> A delightful clock in the town centre of St Helier with the iconic Jersey cow at the base.</p> <p><b>Seq+Tree+Pruning:</b> Not the clock face in the world.</p>
	 <p>31</p>	<p><b>Human:</b> Our cat sleeping in the cot.</p> <p><b>Seq+Tree+Pruning:</b> Our cat is sitting in <del>the</del> bird feeder and <del>actually</del> eats the sun flower seed.</p>
<p><b>Completely wrong</b></p>	 <p>32</p>	<p><b>Human:</b> The floor of the market area in Tirumala was decorated with these rangolis.</p> <p><b>Seq+Tree+Pruning:</b> In a tree ball from the ground train station.</p>
<p><b>Extraneous information</b></p>	 <p>33</p>	<p><b>Human:</b> In the flower bed by the large gate, and various other places in the garden.</p> <p><b>Seq+Tree+Pruning:</b> Random flowers offered to me by two little girls.</p>
<p><b>Vision detection error</b></p>	 <p>34</p>	<p><b>Human:</b> My orange is in a very blue state.</p> <p><b>Seq+Tree+Pruning:</b> Just an apple in the sky.</p>
<p><b>Literally not relevant, but metaphorically creative!</b></p>	 <p>35</p>	<p><b>Human:</b> A butterfly in a field in the Santa Monica mountains.</p> <p><b>Seq+Tree+Pruning:</b> Monarch in her bedroom before the wedding ceremony.</p>

Figure 6.9: Description Generation: Bad Examples.

## CHAPTER 7

### CONCLUSION

#### 7.1 Summary of Contributions

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

Our approach is inspired by previous work in NLP and Computer Vision, in particular, image-matching caption-retrieval approaches. These approaches obtain ready captions from images visually similar to the target image based on global similarity measures. It is unrealistic to expect that the caption retrieved based on global image similarity would fully and precisely describe the target image content. Thus, given an image with no relevant text available, we retrieve matching images based on various aspects, such as objects, actions, stuff and scene. We then extract parts of human-written captions available for the retrieved images and selectively glue those part together.

We have designed a sequence-driven and a novel tree-driven phrase-level description composition methods. Both approaches require complex operations of phrase selection and reordering, thus, we cast the generation task as a constrained optimization problem. The tree-based method takes into account much more complex operations, including parse tree construction, than the sequence-based method does. Our sequence-driven approach outperformed various strong baselines, whilst tree-driven approach performed the best. By integrating both the tree structure and the sequence structure, we have significantly improved the quality of the composed image captions over several competitive baselines.

Additionally, we introduced a new task of *image caption generalization*, aiming to improve existing human-written captions, parts of which serve as building blocks for the new image description. We have presented two approaches to this task, modeling generalization as a

sentence compression problem. Evaluation results showed empirical benefits of automatically compressed captions for the *image description generation* task.

## 7.2 Future Research Directions

We envision several ways to expand our work, which are described below.

### 7.2.1 Expanding Generation Techniques to Creative Language Generation

We described a generation task which concerns image captions. However, we can apply the same techniques to other tasks. If we are for instance to generate a plausible text which satisfies some properties, we can use the same formulation of an optimization problem as in Chapter 6. We only will need to slightly modify the system to fit into a new problem. Mainly we will need to change scoring functions.

For image captions we used vision scores and ngram scores to generate natural language text. For another task, for instance generation of creative text, we still would want to use ngram scores, however vision score we would replace with creativity measures.

Computer Vision researchers are already interested in predicting interestingness and aesthetics of images (e.g., Dhar et al. (2011), Datta et al. (2006)). We would like to concentrate on the interestingness of the language. Our initial study on creativity measures (Kuznetsova et al., 2013a) revealed a few potential measures for text creativity, which we will describe in the remainder of this section. However, this field of research is still in its initial stage.

We found that compositional distributional semantic provides helpful techniques for quantifying creativity. In recent years there has been a swell of work on compositional distributional semantics that captures the compositional aspects of language understanding, such as sentiment analysis (e.g., Yessenalina and Cardie (2011), Socher et al. (2011)) and language

modeling (e.g., Mitchell and Lapata (2009), Baroni and Zamparelli (2010), Guevara (2011), Clarke (2012), Rudolph and Giesbrecht (2010)). Inspired by the thought that the key novelty then lies in the *compositional* operation itself, we explore influence of this operation on creativity. Here we think of a composition as an act of putting together a set of words in an unexpected way, rather than the rareness of individual words being used.

We collected a dataset of creative and not creative word pairs<sup>1</sup> by asking users of Amazon Mechanical Turk to label those word pairs on the scale of creativity from 1 to 5. Score 5 corresponds to creative and score 1 to not creative.

We then used the work of Huang et al. (2012), which provided vector-space models for the words. We explored compositional aspect of creativity by using vector operations between vectors of words in a word pair.

We consider the following compositional vector operations inspired by recent studies for compositional distributional semantics (e.g., Guevara (2011), Clarke (2012), Mitchell and Lapata (2008), Widdows (2008)).

- **add:**  $\vec{w}_1 + \vec{w}_2$
- **diff:**  $abs(\vec{w}_1 - \vec{w}_2)$
- **mult:**  $\vec{w}_1 .* \vec{w}_2$
- **min:**  $\min\{\vec{w}_1, \vec{w}_2\}$
- **max:**  $\max\{\vec{w}_1, \vec{w}_2\}$

All operations take two input vectors  $\in R^n$ , and output a vector  $\in R^n$ . Each operation is applied element-wise. We then perform binary classification over the composed vectors using linear SVM.

As an alternative to explicit vector compositions, we also probe implicit operations based on non-linear combinations of semantic dimensions using kernels (e.g., Schölkopf and Smola (2002), Shawe-Taylor and Cristianini (2004)), in particular:

- **Polynomial:**  $K(x, y) = (\gamma x^T y + r)^d, \gamma > 0$

---

<sup>1</sup>more details can be found in Kuznetsova et al. (2013a)



**Figure 7.1:** Creative (blue bold) and not creative (red italic) word pairs graph.

- RBF:  $K(x, y) = \exp(-\gamma \|x - y\|^2)$ ,  $\gamma > 0$
- Laplacian:  $K(x, y) = \exp(-\gamma \|x - y\|)$ ,  $\gamma > 0$

We were able to achieve accuracy as high as **69.54%**.

**Visualization** To gain additional insight, we project word pairs represented in their vector concatenations onto 2-dimensional space using t-Distributed Stochastic Neighbour Embedding (van der Maaten and Hinton (2008)). Figure 7.1 shows some of the interesting regions of the projection: some regions are relatively futile in having creative phrases (e.g., regions involving simple adjectives such as “good”, “bad”, regions corresponding to legal terms), while some regions are relatively more fruitful (e.g., regions involving abstract adjectives such as “infinite”, “universal”, “fundamental”). There are also many other regions (e.g., in the vicinity of “true”, “perfect” or “intelligent” in Figure 7.1) where the separation between creative and not creative phrases are not as prominent. In those regions, compositional aspects would play a bigger role in determining creativity than memorizing fruitful semantic subspaces.

We also explored various information measures, such as entropy and KL-divergence. More

details on these measures can be found in Kuznetsova et al. (2013a).

## 7.2.2 Improving Image Descriptions

**Aggregation of Multiple Objects in the Descriptions** We could improve multiple objects handling. Recall, that we generate a description for each object. In order to avoid overloaded descriptions, we apply content planning. Latter, however, only selects a subset of objects to describe and orders them. We do not aggregate multiple objects into a single sentence. Thus, some of the descriptions, might sounds robotic, for example, “A flower in the field. A butterfly on the flower.” as opposed to a more naturally sounding “A butterfly on a flower in the field.”. Thus, we can improve our captions by aggregating multiple objects into a single sentence, which can require complex grouping rules and identification of object relations.

**Named Entities** The *image caption generalization task* removes information less likely to be transferable to another image. In many cases this information includes proper names, for instance, “in Florida”. However, named entities, such as “Eiffel Tower” are more likely to be named by people when describing an exact image content. Thus, we need to keep those entities in the description. In the scope of this work we did not differentiate between various proper names. In future, it would be useful to classify proper names into categories and selectively keep some of the proper names corresponding to named entities well known by people.

**Action Detectors** Another item in future work which can improve the generated descriptions is incorporating action detectors. In this work we simply extract verb phrases from images with similar object detections. It could be beneficial to have separate action detectors for determining specific actions depicted in the images. Some of the existing research work on images and videos already explored action detection problem (e.g. Guadarrama et al. (2013), Delaitre et al. (2010))



**Entry-Level Categories** The work of Ordonez et al. (2013) introduced a problem of entry-level categories into an image annotation task. That work aims to find the representative word for a noun category detected by computer vision techniques. This representative word people would use to describe an object depicted on the image. For example, if for an image of a “dolphin” we have an object detection of a “grampus griseus”, we can use WordNet(Fellbaum, 1998) hierarchy and language statistics to infer “dolphin”. This information is useful for image description generation. For example, we can compress or rewrite human captions or retrieve captions which contain more representative words for an object category. If we have a detection of a “bird”, it is safer to use phrases, describing a general “bird”, rather than a specific type of a bird, “heron” for example. However, using only general terms should not be a rule. For example, it would be odd to use “animal” to describe a “cat” or use a “bird” to describe “penguin”.

### 7.2.3 Expanding Sentence Compression Algorithm

In Chapter 5 we proposed a novel sentence compression algorithm, which is driven by a parse tree structure. We can apply the proposed algorithm to the general task of sentence compression and compare to existing compression approaches (e.g. Knight and Marcu (2000), Turner and Charniak (2005), Cohn and Lapata (2007), Filippova and Altun (2013), Cohn and Lapata (2008)). There are also several ways to extend an algorithm, shown in Figures 5.3 and 5.4 to allow more complex operations beyond branch deletion, such as *branch reordering* and *word substitutions*. Additionally to handle word relations, overlooked by PCFG trees, such as subject/direct object, we can enhance the algorithm with *typed dependency* based rules and/or constraints.

APPENDIX A

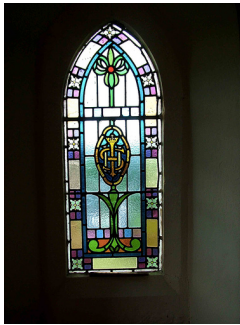
ILP SYSTEM VARIATIONS

OPTION OR CONSTRAINT	DESCRIPTION	SEQ.v.1/2+LINGRULE	SEQ	SEQ+TREE
Variables				
$\alpha$	Variables for Sequence Structure	+	+	+
$\beta$	Variables for Tree Structure	-	-	+
$\gamma$	Auxiliary Tree Structure Variables	-	-	+
Soundness Constraints				
C6.7	Constraints for $\alpha$ variables	+	+	+
C6.8	Constraints for $\beta$ variables	-	-	+
C6.10	Constraints for $\gamma$ variables	-	-	+
Consistency between Tree Leafs and Sequences				
C6.12-C6.13	-	-	-	+
Sequence Congruence Constraints				
C6.17-C6.19	-	+	+	+
Tree Congruence Constraints				
C6.17-C6.20	-	-	-	+
Linguistic and Discourse Constraints				
C3.25	Disallow VP at the beginning	+	-	-
C3.18	enforce a noun-phrase to be selected	+	+	+
C3.19	allow at most one phrase of each type	+	+	+
C3.20-C3.21	enforce plural/singular agreements	+	+	+
C3.22	enforce consistency between VP and cogn phrase	+	-	-
C3.23,C3.24	scene phrase and head words constraints	+	+	+
Various Statistics and Options				
Sentence boundaries	Begin/end of a sentence statistics	+	-	-
Cognitive phrases	-	+/-	-	-

**Table A.1:** ILP system variations. We use “+” to denote usage of an option/constraint, “-” to denote its omission and “+/-” to denote possibility of a usage depending on a system version. Various system versions are explored in evaluation Sections 3.7, 5.6 and 6.4. All constraints indices are consistent with equation numbering.

## APPENDIX B

### ADDITIONAL GOOD EXAMPLES OF GENERATED DESCRIPTIONS



**HMM.v.2+Cogn:** View of a stained glass window colored in the same building in the western infirmary. Glass window depicts Faith Hope and Charity in the church of near the Sydney Town Hall and the Queen Victoria building in church.

**Seq.v.2+LingRule+Cogn:** View of a stained glass window sold us on the house alone in church. This stained glass window in burnt down building depicting St Swithun and St Ethelwood.

**Seq:** This window in the tower of the castle gardens at Dunraven Bay to get a view of the beach near the Sydney Town Hall and the Queen Victoria building.

**Seq+Tree:** This window depicts Faith , Hope , and charity , in the church ofst John the Baptist in Boldre , Hampshire , England.

**Seq+Pruning:** This window depicts the church of the buildings in the western infirmary.

**Seq+Tree+Pruning:** [This window depicts the church.](#)

**Human:** This stained glass window is in the porch of Kilcash Church, Slieve na Mon.



**HMM.v.2+Cogn:** This is a picture of a cow moved off to the side after we stopped in the grass in pasture.

**Seq.v.2+LingRule+Cogn:** These is the picture of the cows moved off to the side after we stopped in pasture.

**Seq:** This cow taking over the bike path in the grass across Tabor Lake , near Prince George , BC , October 21 , 2008.

**Seq+Tree:** This cow was in the middle of the road and actually moved off to the side after we stopped.

**Seq+Pruning:** Taking over the bike path in the grass this cow.

**Seq+Tree+Pruning:** [This cow standing in the middle of a dirt road.](#)

**Human:** Top of the moo to you! A genuine Irish cow in a genuine Irish field just outside of Galway.



**HMM.v.2+Cogn:** I think this is a classic wooden cockpit boat moored in the river in the sky in the ocean.

**Seq.v.2+LingRule+Cogn:** This is one of the little boat moored in the river under the bridge.

**Seq:** Under the bridge in the sky the little boat.

**Seq+Tree:** The little boat is in the sun.

**Seq+Pruning:** Rests in the water under the bridge a small sailboat.

**Seq+Tree+Pruning:** [A small sailboat rests in the water.](#)

**Human:** Sailboat in Waikiki ocean.



**HMM.v.2+Cogn:** One of her puppies being led by a homeless guy on 16th street in Denver against building at watch hill.

**Seq.v.2+LingRule+Cogn:** This is a dog being led by a homeless guy on 16th street in Denver at watch hill.

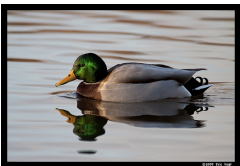
**Seq:** Her puppies are also white from our hotel , which had a walkway to the building over the road on one of the buildings in the apartment complex.

**Seq+Tree:** Her puppies are ready for a nap after 2 hours in the snow and mountains.

**Seq+Pruning:** A dog to the building over the road being a homeless guy on the buildings in the apartment complex.

**Seq+Tree+Pruning:** [A dog being a homeless guy.](#)

**Human:** A street person and his faithful canine companion in Asheville, NC.



**HMM.v.2+Cogn:** This is a picture of a duck swimming in the rain in the summer.

**Seq.v.2+LingRule+Cogn:** One of the duck sitting in the water.

**Seq:** The duck sitting in the water under the coffee table.

**Seq+Tree:** The duck sitting in the water.

**Seq+Pruning:** Swimming pool in the summer the duck.

**Seq+Tree+Pruning:** [The duck sitting in the water.](#)

**Human:** A male mallard enjoying his reflection in the water



**HMM.v.2+Cogn:** These is the picture of the flowers growing in the mountains in the grass in my back yard.

Flowers growing in a glade between rock outcrops in the grass in my back yard. The flower arranged in a hemisphere over green grass from the bus stop.

**Seq.v.2+LingRule+Cogn:** I think this is such a gorgeous flower seen in a mountain valley at Philmont Scout Ranch New Mexico in the Sangre de Cristo Mountains from the bus stop. These spectacular flowers growing in our tomato plant pot over green grass.

**Seq:** Reminds me of a spitting lizard a blue flower in the grass. This flower in the lawn to us saying `` i remember you from the bus stop ''.

**Seq+Tree:** A blue flower was standing all by itself this morning and seemed a perfect shot in the sunshine. This flower is growing in our tomato plant pot.

**Seq+Pruning:** Spectacular flowers in the field. Flowers in the field.

**Seq+Tree+Pruning:** [Spectacular flowers in the field.](#)

**Human:** This sweet little flower was growing in an overgrown field next to a pile of rubble.

## APPENDIX C

### LITERALLY NOT RELEVANT, BUT METAPHORICALLY CREATIVE

#### EXAMPLES OF GENERATED DESCRIPTIONS



**HMM+Cogn.v.2:** These is the picture of the flowers found in a field near a gas station in the water. Fiery flowers found in a field near a gas station in the water.  
**Seq+LingRule+Cogn.v.2:** These is found in the field near the gas station fiery flowers. The most ridiculous looking flower in the water lures me to it.  
**Seq:** Showing against the green background with the silhouette of flower the bright orange flowers in the water. Bathed in red light the bright orange flowers.  
**Seq+Tree:** Fiery flowers showing against a green background with a silhouette of flower. The bright orange flowers spotted in St Nic 's Churchyard , Nottingham.  
**Seq+Pruning:** Fiery flowers in the water.  
**Seq+Tree+Pruning:** [Fiery flowers bathed in red light.](#)  
**Human:** Red flames in the sky.



**HMM+Cogn.v.2:** View of a tabby cat taken over the girls blankies in blue.  
**Seq+LingRule+Cogn.v.2:** View of a tabby cat taken over the girls blankies.  
**Seq:** The cat in the sink at the beach house wonder where the chihuahuas wentthe cats have taken over the girls blankies.  
**Seq+Tree:** The cat sitting in one of his patented `` edge of the couch " positions.  
**Seq+Pruning:** A kitten in the sink at the beach house was a butterfly.  
**Seq+Tree+Pruning:** [A kitten was a butterfly.](#)  
**Human:** The cat of my hostfamily in Michigan in a plastic bag



**HMM.v.2+Cogn:** One of the butterfly clinging to a blue Mist flower at Grapevine Botanical Garden in October before the wedding ceremony.  
**Seq.v.2+LingRule+Cogn:** This is a butterfly clinging to a blue Mist flower at Grapevine Botanical Garden in October.  
**Seq:** Monarch in her bedroom before the wedding ceremony.  
**Seq+Tree:** Winged fairy with swarovski crystals and carved Monarch in her bedroom before the wedding ceremony.  
**Seq+Pruning:** Monarch in her bedroom before the wedding ceremony.  
**Seq+Tree+Pruning:** [Monarch in her bedroom before the wedding ceremony.](#)  
**Human:** A butterfly in a field in the Santa Monica mountains.

## APPENDIX D

### EXAMPLES OF GENERATED DESCRIPTIONS<sup>1</sup>

Image 1



**Human:** New dining room table with a built in stained glass window.1/4"plate glass covers the stained glass and makes a continuous surface  
**HMM.v.1+Cogn:** Here you can see the chair provided in the same building in the back. Leather chairs are in my building in the back. The red chair missing her brother in my building in the back. The table space in the house in the same building in the back.  
**Seq.v.1+LingRule+Cogn:** Love the table frightened vance in my building.  
**HMM.v.2+Cogn:** One of the chair provided in the same building in the back. Leather chairs are in the same building in the back. The red chair missing her brother in my building in the back. The table space in the house in the same building in the back.  
**Seq.v.2+LingRule:** The table frightened vance in the back. Leather chairs surrounded by cookbooks in my building.  
**Seq.v.2+LingRule+Cogn:** One of the table frightened vance in the back. Leather chairs surrounded by cookbooks in my building.  
**Seq:** The table on one of the doors in the same building as the sixth-story restaurant. Made for my dad a high chair in the little basket on the back of the golf.  
**Seq+Tree:** The table is carved from a single tree that fell in the creek. A high chair provided and , in the " stalls " , small tables for refreshments.  
**Seq+Pruning:** Came with house glass in the building the table on the back of the golf. A high chair in the building.  
**Seq+Tree+Pruning:** The table is carved from a single tree that fell in the creek. A high chair in the building.

Image 2



**Human:** Blue and yellow flowers in a very green garden.  
**HMM.v.1+Cogn:** These was taken on the flowers growing together in the center of a big green bush in Tucson Arizona in the grass at the Elkton Community Education Center. Many flowers growing in our fish pond in a field at the Elkton Community Education Center. Small blue flowers found in a conservation zone in a local cemetery in the grass at Gurre Hegn. These purple flowers found in a conservation zone in a local cemetery in the grass at Gurre Hegn.  
**Seq.v.1+LingRule+Cogn:** This is a shot of a blue flower found in a conservation zone in a local cemetery over green grass. A pink flower growing together in the center of a big green bush in Tucson Arizona at Gurre Hegn.  
**HMM.v.2+Cogn:** This is a picture of the flower growing through the slats of a wooden bridge in the Smoky Mountains in the grass by the river. Many flowers were in the grass by the river. Blue flowers appear in the spring in a field by the river. These purple flowers found in a conservation zone in a local cemetery in the grass at Gurre Hegn.  
**Seq.v.2+LingRule:** A blue flower found in a conservation zone in a local cemetery at Gurre Hegn. A pink flower growing together in the center of a big green bush in Tucson Arizona over green grass.  
**Seq.v.2+LingRule+Cogn:** Taken in front of a blue flower found in a conservation zone in a local cemetery at Gurre Hegn. A pink flower growing together in the center of a big green bush in Tucson Arizona over green grass.  
**Seq:** Flower in the grass. Many flowers in the grass with cactus joint bugs and some with soft yellow blooms.  
**Seq+Tree:** A pink flower found in a conservation zone in a local cemetery.  
**Seq+Pruning:** Blue flowers in the grass.  
**Seq+Tree+Pruning:** Blue flowers creeping plant out in front of my house. Many flowers in the grass.

Image 3



**Human:** In the flower market in Bangkok  
**HMM.v.1+Cogn:** These was taken on the flowers looked so pure and fresh bathing in the rain in the library flower beds. Pink flower growing in Susan by Arlenk. Water lily growing in Susan by Arlenk. A flower growing in the back garden not sure sadly what it exact name in the library flower beds.  
**Seq.v.1+LingRule+Cogn:** This is a picture of this flower growing in Susan. The white water lily looked so pure and fresh bathing in the rain by Arlenk.  
**HMM.v.2+Cogn:** One of the flower covered with water plants in the library flower beds. Pink flower growing in Susan of red and pink. All the flowers growing on the street sign at the corner near our house in the library flower beds. A flower is in the library flower beds.  
**Seq.v.2+LingRule:** Many blue flowers growing in Susan. The white water lily looked so pure and fresh bathing in the rain.  
**Seq.v.2+LingRule+Cogn:** Found this flower growing in Susan. The white water lily looked so pure and fresh bathing in the rain.  
**Seq:** These flowers found it today underneath other plants in the plant pot. Purple flowers growing in Susan 's flower gardens.  
**Seq+Tree:** These flowers covered the entire side of a building in downtown half Moon Bay , CA. Purple flowers growing in Susan 's flower gardens.  
**Seq+Pruning:** This flower was the road side on the climb up the Canyon in the France. These flowers covered the side of a building in the plant pot.  
**Seq+Tree+Pruning:** This flower was the road side on the climb up the Canyon in the France. These flowers covered the side of a building.

Image 4



**Human:** Flower in sand  
**HMM.v.1+Cogn:** I liked the way flower growing by the road side on the climb up the Canyon Des Ecouges in the Vercors in France in a tree in the tree. Adjoining flowers found growing in the sand on Straddy in the trees in their new summer dress and straw hat.  
**Seq.v.1+LingRule+Cogn:** I like the way the flower found growing in the sand on Straddy under her Christmas tree. The rose pots my mom has out around our house in the tree.  
**HMM.v.2+Cogn:** This is a flower growing by the road side on the climb up the Canyon Des Ecouges in the Vercors in France in a tree in the tree. Adjoining flowers found growing in the sand on Straddy in the trees in their new summer dress and straw hat.  
**Seq.v.2+LingRule:** The rose pots my mom has out around our house in the tree. The flower was in the trees.  
**Seq.v.2+LingRule+Cogn:** One of the rose growing by the road side on the climb up the Canyon Des Ecouges in the Vercors in France of gray. The flower was in the trees.  
**Seq:** A flower was in the trees to the ground by a tree trunk. The flower was in the trees.  
**Seq+Tree:** A big-ass pink flower pots my mom has out around our house. The flower is on a plant in front of our home.  
**Seq+Pruning:** Pots my mom house a flower in the trees to the ground by a tree trunk. The flower in the trees.  
**Seq+Tree+Pruning:** A flower was the road side on the climb up the Canyon in the France. The flower is on a plant in front of our home.

Image 5



**Human:** Photos from an afternoon in Chestertown, MD, including a public sail on the 1767 replica ship Sultana  
**HMM.v.1+Cogn:** There are the birds seen near the port in Aruba in water in the water. Yellow bird munching on a fish in Sua in water in the water.  
**Seq.v.1+LingRule+Cogn:** Here you can see a yellow bird in water. An interesting looking bird munching on a fish in Sua by the water.  
**HMM.v.2+Cogn:** One of the bird flying in water in the water. Yellow bird munching on a fish in Sua in water in the water.  
**Seq.v.2+LingRule:** Yellow bird seen near the port in Aruba at the beach by the water. An interesting looking bird munching on a fish in Sua by the water.  
**Seq.v.2+LingRule+Cogn:** This is a picture of the yellow bird seen near the port in Aruba at the beach by the water. An interesting looking bird munching on a fish in Sua by the water.  
**Seq:** Looked a bit sick by the water in the pools a green heron. Birds in the water.  
**Seq+Tree:** A green heron seen near the port in Aruba standing on the back of a chair. Birds in the water.  
**Seq+Pruning:** An interesting bird looked a bit sick by the water in the pools. Asked for a dumbo octopus in the sky with a few birds flying by the water an interesting bird.  
**Seq+Tree+Pruning:** An interesting bird looked a bit sick. An interesting bird asked for a dumbo octopus in the sky with a few birds flying.

Image 6

**Human:** Found this guy right outside my sleeping bag waking up in OKC.  
**HMM.v.1+Cogn:** This was taken on the MBTA Blue Line St Louis train lost against the mountain background at Horseshoe Curve in April 1986 in sky under a rock.  
**Seq.v.1+LingRule+Cogn:** This is what happens when a train lost against the mountain background at Horseshoe Curve in April 1986 under a rock.  
**HMM.v.2+Cogn:** This was taken on the MBTA Blue Line St Louis train lost against the mountain background at Horseshoe Curve in April 1986 in sky

<sup>1</sup>IF IMAGES IN THE FIGURES OF THE DISSERTATION ARE MARKED WITH NUMBERS, THOSE CORRESPOND TO INDICES OF IMAGES IN THIS APENDIX.



**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

under a rock.  
 The railroad lost against the mountain background at Horseshoe Curve in April 1986 under a rock.  
 This is a train lost against the mountain background at Horseshoe Curve in April 1986 under a rock.  
 A train under a rock in the sky.  
 A train is almost lost against the mountain background at Horseshoe Curve in April 1986.  
 The railroad on a chilly autumn morning in the sky is the mountain background.  
 The railroad is the mountain background.

Image 7



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

A softshell turtle spotted during a walk in Ft. Worth, TX USA.  
 Love the handmade bags affixed to mail box in tree with a blended rainbow effect.  
 This is a shoulder bag with a blended rainbow effect.  
 Love the handmade bags affixed to mail box in tree with a blended rainbow effect.  
 Of street water in the tree affixed to mail box trash bags.  
 This is a shoulder bag featuring many vintage buttons in different materials with a blended rainbow effect.  
 60 bags in the tree affixed to mail box around the roof of the bridge of the Henry Leith , Wongat Island , Madang.  
 60 bags featuring many vintage buttons in different materials.  
 60 bags inspired by game arcade in the water photo with a blended rainbow effect.  
 60 bags inspired by game arcade.

Image 8



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

Cup by Corning, plate marked Sterling vitrified china, East Liverpool, OH, G-3  
 Taken out of the coffee cup fixed to the table on the street in the butterfly room.  
 This is a shot of the road fixed to the table cup.  
 This is a picture of a coffee cup fixed to the table on the street in the butterfly room.  
 On the street the cup is just around the corner.  
 One of the road fixed to the table cup.  
 Fixed to the table and the person on the street in a big girl bed the cup.  
 The cup is just around the corner.  
 Tea sign that the cup is the corner on the street near my home in the big girl in the butterfly room.  
 The cup in the middle of the road.

Image 9



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

At a butterfly house somewhere in North Wales  
 Found this yellow butterfly feeding in Judy flower garden in an orange tree santando nell albero.  
 Of apples the butterfly feeding in Judy flower garden by a tree.  
 Love the yellow butterfly feeding in Judy flower garden in an orange tree santando nell albero.  
 Unknown butterfly was on the sidewalk in the middle of a busy downtown street santando nell albero.  
 Of apples the butterfly feeding in Judy flower garden by a tree.  
 A butterfly under the tree at the big girl sitting by the river with her daddy was on the sidewalk in the middle of a busy downtown street.  
 This butterfly was on the sidewalk in the middle of a busy downtown street.  
 This butterfly in the grass under the tree.  
 This butterfly in the grass.

Image 10



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

A bike in a field dreams of unconventional places  
 This is a picture of the 1 ball stuck in the wall at Fort Sumter in the sky under a cloudy sky. A bike used by the Portland Zoo Bombers in the sky under a cloudy sky. A sad bike parked against a wall in Damascus Syria in the sky under a cloudy sky. The clouds were in the sky under a cloudy sky.  
 I like the way the clouds walking in the poppy field under cloudy sky.  
 One of the ball stuck in the wall at Fort Sumter in the sky in the back ground. A bike used by the Portland Zoo Bombers in the sky under a cloudy sky. The bike ride for charity from Blackpool Tower to Eiffel Tower in Paris via many long roads and high hills in the sky in the back ground. The clouds were in the sky in the back ground.  
 By river embracing dog in the sky were the clouds.  
 One of the clouds walking in the poppy field under a cloudy sky.  
 Under cloudy sky the clouds were reflected in the tide pools on the beach of the tower de.  
 The clouds were walking in the poppy field.  
 The clouds were in the sky of the tower de.  
 The clouds were in the sky.

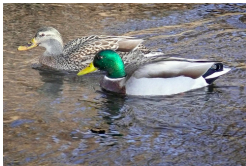
Image 11



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

Kites were flying all around the Washington monument during the cherry blossom festival. It was beautiful.  
 I like the way the big cross contrasting against sky in the sky by the frog.  
 Here you can see a cross by the frog in the sky.  
 This is one of the big cross contrasting against sky in the sky by the frog.  
 By the frog in the sky amazing a cross.  
 This is a cross standing by the frog.  
 A cross in the sky as the storm was simply amazing.  
 A cross stands guard over the cemetery that sank beneath the ocean during the eruption of Vulcan Daan in 1871.  
 A cross in the sky as the storm.  
 A cross in the sky.

Image 12



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**

Maybe the most common bird in the neighborhood, not just the most common water fowl in the neighborhood!Ralston Creek Trail, 12-16-09.  
 I liked the way duck swimming in a pond near the small town of Tnkai in Lithuania under water in the metroparks. Duck has awesome blue patches in its feathers under water for Cambridge.  
 Photo of a mandarin duck having a fight over a chip a boy had thrown into the water of the pond.  
 This is a picture of a duck having a fight over a chip a boy had thrown into the water in the water in the lake. Duck has awesome blue patches in its feathers under water for Cambridge.  
 At De Wolfe Point state park in water arrived weeks ago and were sitting around on the snow until the pond thawed these ducks.  
 One of a mandarin duck having a fight over a chip a boy had thrown into the water for Cambridge.  
 Swimming by in a pool at `` De Efteling '' , a recreational park in the Netherlands under the rocks of the pond a duck.  
 A duck swimming by in a pool at `` De Efteling '' , a recreational park in the Netherlands.

**Seq+Pruning:** The duck was having a feast of the pond in golden water.  
**Seq+Tree+Pruning:** The duck was having a feast.

Image 13



**Human:** Clock tower in downtown  
**HMM.v.1+Cogn:** Found tower shown through a small window near the top of this 12th century church tower in the southern sky in the town. Some tower dwarfing the Boston skyline in the distance mostly with blue sky from Mauthausen.  
**Seq.v.1+LingRule+Cogn:** This is the view from the clock tower converted into an office along Pacific Coast Highway in Seal Beach CA Approximately 16200 PCH from Mauthausen.  
**HMM.v.2+Cogn:** This is a picture of the tower setting sun in the sky in front. Some tower dwarfing the Boston skyline in the distance mostly with blue sky from Mauthausen.  
**Seq.v.2+LingRule:** Clock tower stands in stone from Mauthausen.  
**Seq.v.2+LingRule+Cogn:** View from the top of the clock tower converted into an office along Pacific Coast Highway in Seal Beach CA Approximately 16200 PCH from Mauthausen.  
**Seq:** Converted into an office along Pacific Coast Highway in Seal Beach , CA Approximately 16200 PCH on a lazy Sunday morning under a pale pastel sky some tower.  
**Seq+Tree:** Some tower converted into an office along Pacific Coast Highway in Seal Beach , CA Approximately 16200 PCH.  
**Seq+Pruning:** Tower in the town.  
**Seq+Tree+Pruning:** Tower in the town.

Image 14



**Human:** A butterfly on a flower near the Hammocks Clubhouse in Bald Head Island, North Carolina.  
**HMM.v.1+Cogn:** I like the way the orange butterfly made with a mirror framed effect in Photoshop with Canon G10. The butterflies attracted to the colorful flowers in Hope Gardens with Canon G10.  
**Seq.v.1+LingRule+Cogn:** Here you can see the butterflies attracted to the colorful flowers in Hope Gardens.  
**HMM.v.2+Cogn:** View from the top of the orange butterfly made with a mirror framed effect in Photoshop with Canon G10. The butterflies made with a mirror framed effect in Photoshop to the car.  
**Seq.v.2+LingRule:** The butterflies into their life.  
**Seq.v.2+LingRule+Cogn:** I liked the way into their life attracted to the colorful flowers in Hope Gardens the butterflies.  
**Seq:** A butterfly to the car was spotted by my nine year old cousin.  
**Seq+Tree:** The butterflies are attracted to the colorful flowers in Hope Gardens.  
**Seq+Pruning:** The butterflies are attracted to the colorful flowers to the car.  
**Seq+Tree+Pruning:** The butterflies are attracted to the colorful flowers.

Image 15



**Human:** Yellow flower near Morava river  
**HMM.v.1+Cogn:** I love the way the flowers taken in Madrid March 2006 in a potato field near Flagstaff. Nice yellow flowers found in Venezuela in the field near Flagstaff.  
**Seq.v.1+LingRule+Cogn:** The flower in a field near Flagstaff dancing with the wind by the road side. The flowers in a field buds under the microscope.  
**HMM.v.2+Cogn:** One of the flowers floating in it in the field in California. Nice yellow flowers found in Venezuela in the field near Flagstaff.  
**Seq.v.2+LingRule:** The flowers in the field buds under the microscope. The flower in the field dancing with the wind by the road side to the river access.  
**Seq.v.2+LingRule+Cogn:** Found this flower taken in Madrid March 2006 near Flagstaff. A native flower found in Venezuela.  
**Seq:** This flower in the grass is along the path to the river access. Yellow flower in the field.  
**Seq+Tree:** The flowers close up picture of an orange flower taken in Madrid , March 2006. Just an average yellow flower found in Venezuela.  
**Seq+Pruning:** Beautiful flower in the field.  
**Seq+Tree+Pruning:** Beautiful flower is in the sun. Yellow flower lost in grass.

Image 16



**Human:** Yellow flower in my field  
**HMM.v.1+Cogn:** These was taken on the flowers growing in a rock garden in the field in two sorts. This little flower sprouted up in defiance in the field in two sorts. A full open flower sprouted up in defiance in the field in gardens. Bright yellow flowers growing in a rock garden in the field in gardens.  
**Seq.v.1+LingRule+Cogn:** This is a photo of this little flower sprouted up in defiance against grass. Bright yellow flowers growing in a rock garden at Volcan Mombacho.  
**HMM.v.2+Cogn:** These is the picture of the flowers growing in a rock garden in the field in the grass. The flowers are in the gargantuan bed between our house and the neighbors in the field in the grass. A full open flower sprouted up in defiance in the field in gardens. The flowers are in the gargantuan bed between our house and the neighbors in the field in the grass.  
**Seq.v.2+LingRule:** This little flower sprouted up in defiance at Volcan Mombacho. Bright yellow flowers growing in a rock garden against grass.  
**Seq.v.2+LingRule+Cogn:** This is a picture of this little flower sprouted up in defiance at Volcan Mombacho. Bright yellow flowers growing in a rock garden against grass.  
**Seq:** Against grass the flowers growing in a rock garden.  
**Seq+Tree:** The flowers growing in a rock garden.  
**Seq+Pruning:** The flowers in the field near my house. Bright yellow flowers on grass in the cloud forest at Volcan Mombacho.  
**Seq+Tree+Pruning:** The flowers in the field near my house.

Image 17



**Human:** We had to cross this bridge in order to see the Leaning Tower of Pisa. This area was really peaceful.  
**HMM.v.1+Cogn:** Love the castle known for being the home of Hamlet in the Shakespeare play among the green and the sky across the water.  
**Seq.v.1+LingRule+Cogn:** Love the castle across the water among the green and the sky.  
**HMM.v.2+Cogn:** One of the castle known for being the home of Hamlet in the Shakespeare play among the green and the sky across the water.  
**Seq.v.2+LingRule:** Near the river in the sky is well worth a visit the castle.  
**Seq.v.2+LingRule+Cogn:** One of the castle known for being the home of Hamlet in the Shakespeare play across the water.  
**Seq:** A castle in the sky across the water is on the hill.  
**Seq+Tree:** The castle is on the hill.  
**Seq+Pruning:** A castle in the sky across the water is well worth a visit.  
**Seq+Tree+Pruning:** The castle is well worth a visit.

Image 18



**Human:** Cat in the cat tree - Black and White  
**HMM.v.1+Cogn:** Found kitty resting in her bed and looking out the window from the baptismal pool.  
**Seq.v.1+LingRule+Cogn:** This is what happens when a cat resting in her bed and looking out the window.  
**HMM.v.2+Cogn:** This is a picture of a black cat resting in her bed and looking out the window in the dark.  
**Seq.v.2+LingRule:** The cat is.  
**Seq.v.2+LingRule+Cogn:** This is a cat resting in her bed and looking out the window.  
**Seq:** Bailey resting in her bed and looking out the window my cat. Takes over my bed during the day a cat in the dark.  
**Seq+Tree:** Black cat walk by the window in the kitchen. A cat takes over my bed during the day.  
**Seq+Pruning:** NO SOLUTION

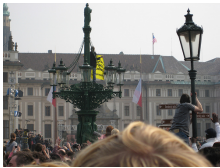
Seq+Tree+Pruning: NO SOLUTION

Image 19



**Human:** Flat bed Chisholms truck on display at the vintage vehicle rally at Astley Green Colliery near Leigh Lanes  
**HMM.v.1+Cogn:** This is the first cellar door left back bedroom in center and clothes dryer to the right to the building in the house. This HUGE screen hanging on the wall outside a burned down building in the house. My truck parked on first avenue in the east village by the glass buildings in the house.  
**Seq.v.1+LingRule+Cogn:** Found trucks parked on first avenue in the east village.  
**HMM.v.2+Cogn:** This was taken at the door closed to the building at the top. LCD screen hanging on the wall to the building at the top. My truck parked on first avenue in the east village by the glass buildings in the house.  
**Seq.v.2+LingRule:** Trucks parked on first avenue in the east village at the East Chase Mall.  
**Seq.v.2+LingRule+Cogn:** These is the picture of the trucks parked on first avenue in the east village at the East Chase Mall.  
**Seq:** Truck from an old building while the interior is a new structure painted in a patriotic scheme of red white and blue by the pink fountain water at the East Chase mall.  
**Seq+Tree:** In the Ditta ranch truck by the glass buildings.  
**Seq+Pruning:** Double door to the building at the East stop in the bathroom.  
**Seq+Tree+Pruning:** Double door stop in the bathroom.

Image 20



**Human:** Greenpeace guy in the green lamp  
**HMM.v.1+Cogn:** These is the view from the signs found in a Clearwater Florida park from the current building of the photo. A streetlight have a unique look in this area of NYC onto the nearby brick office building in picturesque Rockport MA.  
**Seq.v.1+LingRule+Cogn:** These is the view from the signs onto the nearby brick office building.  
**HMM.v.2+Cogn:** These is the picture of the signs found in a Clearwater Florida park to the building in the background. A streetlight have a unique look in this area of NYC onto the nearby brick office building in picturesque Rockport MA.  
**Seq.v.2+LingRule:** Light have a unique look in this area of NYC off the Barnes town green.  
**Seq.v.2+LingRule+Cogn:** These is the picture of the signs in picturesque Rockport MA onto the nearby brick office building.  
**Seq:** Light of the house in the background.  
**Seq+Tree:** A streetlight is in a garden outside a government building at the corner of Bay and Wellesley Streets in Toronto.  
**Seq+Pruning:** Have a unique look in this area across the street from the current building of the cow hand side of the photo a streetlight.  
**Seq+Tree+Pruning:** A streetlight is in a government building at the corner.

Image 21



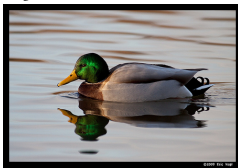
**Human:** A black headed gull flies low over a lake in Cannon Hill Park, Birmingham, England.  
**HMM.v.1+Cogn:** This is the first such a majestic bird flying over one of the Santee Lakes in the sky in this shot.  
**Seq.v.1+LingRule+Cogn:** Found this hawk flying over one of the Santee Lakes in this shot.  
**HMM.v.2+Cogn:** I think this is such a majestic bird flying over one of the Santee Lakes in the sky in this shot.  
**Seq.v.2+LingRule:** At Sion Hill Hall in the sky flying over one of the Santee Lakes a hawk.  
**Seq.v.2+LingRule+Cogn:** This is a hawk flying over one of the Santee Lakes in this shot.  
**Seq:** Frigate birds flying above our boat a hawk on the road to roxbysuch a majestic bird in flight over the sky.  
**Seq+Tree:** A hawk flying over Crusoe No 7 Reservoir , Kangaroo Flat , Victoria.  
**Seq+Pruning:** Flying above our boat a hawk in the sky.  
**Seq+Tree+Pruning:** The bird flying above our boat.

Image 22



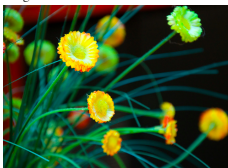
**Human:** Blue flower in the shadow  
**HMM.v.1+Cogn:** I like the way the flower blooming in my driveway in the grass in the wildflower bed. Beautiful pink flower painted in Photoshop elements in the grass in sussex nj.  
**Seq.v.1+LingRule+Cogn:** I liked the way flower painted in Photoshop elements by a farmers field. The little flowers painted in Photoshop elements over the sink.  
**HMM.v.2+Cogn:** One of the flower hanging around in the grass in the garden gray veined white butterfly. Beautiful pink flower painted in Photoshop elements in the grass in sussex nj.  
**Seq.v.2+LingRule:** The little flowers painted in Photoshop elements over the sink. Flower painted in Photoshop elements by a farmers field.  
**Seq.v.2+LingRule+Cogn:** One of the flower painted in Photoshop elements over the sink. Flower painted in Photoshop elements by a farmers field.  
**Seq:** Blooming in the middle of the woods in sussex the flower by a farmers field.  
**Seq+Tree:** The little flowers growing in Susan 's flower gardens.  
**Seq+Pruning:** Blooming blue flowers in the grass.  
**Seq+Tree+Pruning:** Little flowers blooming in the grass.

Image 23



**Human:** A male mallard enjoying his reflection in the water  
**HMM.v.1+Cogn:** I liked the way duck swimming in the rain in the summer.  
**Seq.v.1+LingRule+Cogn:** This is a picture of the duck sitting in the water.  
**HMM.v.2+Cogn:** This is a picture of a duck swimming in the rain in the summer.  
**Seq.v.2+LingRule:** The duck in London.  
**Seq.v.2+LingRule+Cogn:** One of the duck sitting in the water.  
**Seq:** The duck sitting in the water under the coffee table.  
**Seq+Tree:** The duck sitting in the water.  
**Seq+Pruning:** Swimming pool in the summer the duck.  
**Seq+Tree+Pruning:** The duck sitting in the water.

Image 24



**Human:** Some flower on a bar in a hotel in Grapevine, TX  
**HMM.v.1+Cogn:** I liked the way flower surrounded by a little bit in the grass in the ground. Flowers growing in my grandparents back garden in some grass to the bridge.  
**Seq.v.1+LingRule+Cogn:** I like the way the flower surrounded by a little bit in some grass. A tiny pink flower stands out against a bed of bright green foliage in some grass.  
**HMM.v.2+Cogn:** This is a picture of the flower surrounded by a little bit in the grass in the field. Flowers growing in my grandparents back garden in some grass to the bridge.  
**Seq.v.2+LingRule:** The flower to the bridge surrounded by a little bit in the grass. A tiny pink flower in some grass stands out against a bed of bright green foliage.  
**Seq.v.2+LingRule+Cogn:** One of the flower surrounded by a little bit to the bridge in the field. A tiny pink flower stands out against a bed of bright green foliage in some grass.  
**Seq:** Through out the walk to the bridge the flower in the grass was so vivid and attractive. 've no idea what they are in the field behind my house cute pink flowers.  
**Seq+Tree:** The flower is opening in the garage border. Cute pink flowers growing in my grandparents.  
**Seq+Pruning:** Through out the walk to the bridge the flower in the grass was so vivid and attractive.  
**Seq+Tree+Pruning:** The flower was so vivid and attractive.



Image 25



**Human:** Sailboat in Waikiki ocean  
**HMM.v.1+Cogn:** This is what happens when a classic wooden cockpit boat moored in the river in the sky in the ocean.  
**Seq.v.1+LingRule+Cogn:** This is the view from the little boat moored in the river under the bridge.  
**HMM.v.2+Cogn:** I think this is a classic wooden cockpit boat moored in the river in the sky in the ocean.  
**Seq.v.2+LingRule:** The little boat rests in the water in Cape Porpoise Maine under the bridge.  
**Seq.v.2+LingRule+Cogn:** This is one of the little boat moored in the river under the bridge.  
**Seq:** Under the bridge in the sky the little boat.  
**Seq+Tree:** The little boat is in the sun.  
**Seq+Pruning:** Rests in the water under the bridge a small sailboat.  
**Seq+Tree+Pruning:** A small sailboat rests in the water.

Image 26



**Human:** Spring in a white dress  
**HMM.v.1+Cogn:** I liked the way flower found in a conservation zone in a local cemetery in the grass in my heart. A flower found in shady places in the woods in the grass in my heart. These small white flowers be seen in my photostream in the grass in my heart. Flowers are all over the forest floor in the grass in my heart. Flowers taken at a garden in Manchester in the grass in my heart. Two blue passion flowers die in the grass in my heart. A flower found in shady places in the woods in the grass in my heart.  
**Seq.v.1+LingRule+Cogn:** These was taken on the flowers against grass found in shady places in the woods.  
**HMM.v.2+Cogn:** This is a picture of the flower seen in this hole in the grass by the river. A flower found in shady places in the woods in the grass in my heart. These small white flowers be seen in my photostream in the grass in my heart. Flowers die in the grass by the river. Flowers growing along the trail in the grass by the river. Two blue passion flowers die in the grass by the river. Flowers are all over the forest floor in the grass by the river.  
**Seq.v.2+LingRule:** A flower found in shady places in the woods outside our house. Flower found in a conservation zone in a local cemetery against grass.  
**Seq.v.2+LingRule+Cogn:** This is a flower found in shady places in the woods outside our house. Flower found in a conservation zone in a local cemetery against grass.  
**Seq:** The flowers in the grass will never die. 've no idea what they are in the tree these small white flowers against grass.  
**Seq+Tree:** The flowers will never die. Other flowers are found in shady places in the woods.  
**Seq+Pruning:** Blue flowers in the grass have no scent. 've no idea what they are small white flowers in the grass.  
**Seq+Tree+Pruning:** Blue flowers have no scent. Small white flowers 've no idea what they are.

Image 27



**Human:** Tower bridge london in black and white  
**HMM.v.1+Cogn:** These is the view from the signs found in a Clearwater Florida park of a building for the blue sky hehe. The lighthouse placed symmetrical against the street of a building for the blue sky hehe.  
**Seq.v.1+LingRule+Cogn:** Love the tower placed symmetrical against the street of a building.  
**HMM.v.2+Cogn:** These is the picture of the signs found in a Clearwater Florida park of the building from the City Hall. The lighthouse placed symmetrical against the street of a building for the blue sky hehe.  
**Seq.v.2+LingRule:** In a house the tower placed symmetrical against the street.  
**Seq.v.2+LingRule+Cogn:** One of the tower placed symmetrical against the street for the blue sky hehe.  
**Seq:** A tower in the middle of a building is amazing : many towers built with the same face on each side for the blue sky hehe.  
**Seq+Tree:** A tower is watching over the parade in 2009.  
**Seq+Pruning:** The tower in the middle of a building built on each side by grant.  
**Seq+Tree+Pruning:** The tower built on each side.

Image 28



**Human:** Shot in Blackpool on Tesco's car park with use of a 5,000,000 candle power light.  
**HMM.v.1+Cogn:** Saw this orange car covered in orange velvet in the street by Varioseif.  
**Seq.v.1+LingRule+Cogn:** I love the way this car covered in orange velvet.  
**HMM.v.2+Cogn:** Love this orange car covered in orange velvet in the street by Varioseif.  
**Seq.v.2+LingRule:** On Kauai of the main road covered in orange velvet this car.  
**Seq.v.2+LingRule+Cogn:** Found this car covered in orange velvet of the road.  
**Seq:** Car in the middle of the road - Pyongyang - North Korea over the 18th green at Bandon Dunes gc from our second floor room at the Inn.  
**Seq+Tree:** Over the 18th green at Bandon Dunes gc from our second floor room at the Inn car in the middle of the road - Pyongyang - North Korea.  
**Seq+Pruning:** Red car in the middle of the road at the Inn.  
**Seq+Tree+Pruning:** Red car in the middle of the road.

Image 29



**Human:** A delightful clock in the town centre of St Helier with the iconic Jersey cow at the base.  
**HMM.v.1+Cogn:** Found clock running in clockwise direction of Hoskins building in art museum. Wonderful pink flowers found in Stkholm of Hoskins building in art museum. The main shop window is near my workplace of Hoskins building in art museum.  
**Seq.v.1+LingRule+Cogn:** I liked the way flower on a building. The main shop window is near my workplace.  
**HMM.v.2+Cogn:** This was taken on the clock running in clockwise direction in the house in a new shopping area. Wonderful pink flowers found in Stkholm of Hoskins building in art museum. Window overlooking a canal in the house in a new shopping area.  
**Seq.v.2+LingRule:** The main shop window is near my workplace of Hoskins building.  
**Seq.v.2+LingRule+Cogn:** This is one of the main shop window located across the street from the historic Redford Theatre in Detroit of Hoskins building.  
**Seq:** The flowers in our room for a boat on the Chicago River near the Sun-Times building are over the roof of my shed which was an old wash house. Window on a building in Riga , Latvia is near my workplace.  
**Seq+Tree:** The flowers are over the roof of my shed which was an old wash house. By a somewhat lower building window in our room.  
**Seq+Pruning:** Not the clock running on a building in a new shopping area.  
**Seq+Tree+Pruning:** Not the clock face in the world.

Image 30



**Human:** Our cat sleeping in the cot  
**HMM.v.1+Cogn:** I liked the way cat sitting in the bird feeder and actually eats the sun flower seed in the air.  
**Seq.v.1+LingRule+Cogn:** Saw this cat sitting in the bird feeder and actually eats the sun flower seed.  
**HMM.v.2+Cogn:** This is a picture of the cat sitting in the bird feeder and actually eats the sun flower seed in the air.  
**Seq.v.2+LingRule:** This cat sitting in the bird feeder and actually eats the sun flower seed.  
**Seq.v.2+LingRule+Cogn:** Photo of this cat sitting in the bird feeder and actually eats the sun flower seed.  
**Seq:** Fighting over the basket on the table in a bar this cat.  
**Seq+Tree:** This cat is sitting in the bird feeder and actually eats the sun flower seed.  
**Seq+Pruning:** Our cat is sitting in bird feeder and the sun flower seed of that spot , herself.  
**Seq+Tree+Pruning:** Our cat is sitting in bird feeder and the sun flower seed.

Image 31

**Human:** The floor of the market area in Tirumala was decorated with these rangolis.  
**HMM.v.1+Cogn:** One of the few ball made by my boy in a tree from the under ground train station. A banana stand in a tree from the under ground train



**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

station. The tiny dish lit from behind by a candle in tree from the under ground train station.  
The glazing ball from the under ground train station made by my boy in tree.  
This is a picture of the ball bounced off the Dr Pepper sign in left field in a tree from the under ground train station. A banana stand in a tree from the under ground train station. The tiny dish lit from behind by a candle in tree from the under ground train station.  
In downtown Portland gazing ball made by my boy.  
This is a picture of the gazing ball made by my boy from the under ground train station.  
Wooden plate from the under ground train station in a tree.  
Wooden plate made in Transilvania , hanged on natural wooden wall.  
Ball from the ground train station in a tree made by my boy.  
In a tree ball from the ground train station.

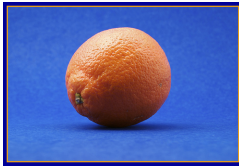
Image 32



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

In the flower bed by the large gate, and various other places in the garden  
On the side of a flower growing in the lava fields in the grass in a pink flower. Flowers offered to me by two little girls in the grass in a pink flower.  
This is a picture of this flower offered to me by two little girls in the grass. Random flowers growing in the lava fields of river and sea stones.  
This is a flower growing in the lava fields in the field in the middle. Flowers offered to me by two little girls in the grass in a pink flower. Random flowers growing in the lava fields of river and sea stones. This flower offered to me by two little girls in the grass.  
Love the random flowers growing in the lava fields of river and sea stones. This flower offered to me by two little girls in the grass.  
Yellow flowers on the plant stand offered to me by two little girls in the field.  
A flower growing in the lava fields.  
Random flowers offered to me by two little girls in the field.  
Random flowers offered to me by two little girls.

Image 33



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

My orange is in a very blue state.  
On the side of a silver ball suspended over the ocean under blue sky over glass bud vase probably Hazel Atlas depression era.  
I think this is a just an apple under blue sky.  
This is a silver ball suspended over the ocean under blue sky over glass bud vase probably Hazel Atlas depression era.  
Just an apple holding on a tree in the winter over glass bud vase probably Hazel Atlas depression era.  
I think this is just an apple stuck in the grooves above the house.  
Holding on a tree in the winter of the sky above the house just an apple.  
Just an apple shot in a light box with a reverse lens.  
Just an apple of the sky of the sunset in the heaven.  
Just an apple in the sky.

Image 34



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

A butterfly in a field in the Santa Monica mountains.  
Found the butterfly clinging to a blue Mist flower at Grapevine Botanical Garden in October before the wedding ceremony.  
The butterfly before the wedding ceremony clinging to a blue Mist flower at Grapevine Botanical Garden in October.  
One of the butterfly clinging to a blue Mist flower at Grapevine Botanical Garden in October before the wedding ceremony.  
Monarch butterfly clinging to a blue Mist flower at Grapevine Botanical Garden in October.  
This is a butterfly clinging to a blue Mist flower at Grapevine Botanical Garden in October.  
Monarch in her bedroom before the wedding ceremony.  
Winged fairy with swarovski crystals and carved Monarch in her bedroom before the wedding ceremony.  
Monarch in her bedroom before the wedding ceremony.  
Monarch in her bedroom before the wedding ceremony.

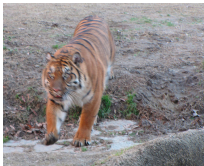
Image 35



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

Butterfly bracelet Turquoise and Fire agate set in sterling silver  
Love the green apple displayed in a bowl in the sky in the window. My window over the entry to the restaurant against a blue sky in rose gold plated sterling silver.  
The most realistic blue apple flies in medium against the the blue sky.  
Love the green apple floating in water in the sky in the window. My window window over the entry to the restaurant against a blue sky in rose gold plated sterling silver.  
Green apple flies in medium against the the blue sky.  
Love the green apple floating in water against the the blue sky.  
Green apple in the window under the glorious blue sky.  
Green apple displayed in a bowl.  
Green apple in the sky.  
Green apple in the sky.

Image 36



**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**  
**Seq+Pruning:**  
**Seq+Tree+Pruning:**

Here is the tiger again but in motion he/she was running towards the fence.  
Here you can see the female tiger playing in the water by a missions cmpound.  
Here you can see the female tiger playing in the water near the Mendenhall Glacier.  
This is a tiger playing in the water in the Rockies Canada.  
In the Rockies Canada the female tiger.  
This is one of the female tiger playing in the water.  
A tiger looks down from a high wall in his enclosure at Disney 's Animal Kingdom near the Mendenhall Glacier.  
A tiger looks down from a high wall in his enclosure at Disney 's Animal Kingdom.  
A tiger in the Canada.  
A tiger looks down from a high wall in his enclosure at Disney 's Animal Kingdom.

Image 37

**Human:**  
**HMM.v.1+Cogn:**  
**Seq.v.1+LingRule+Cogn:**  
**HMM.v.2+Cogn:**  
**Seq.v.2+LingRule:**  
**Seq.v.2+LingRule+Cogn:**  
**Seq:**  
**Seq+Tree:**

Kitchen table and built in kitchen desk as seen from family room angle.  
View from the rocking chair came in by Seneca Lalonde.  
Taken in front of her high chair came in by Seneca Lalonde.  
Taken in front of the rocking chair came in by Seneca Lalonde.  
Her high chair under the stairs and wood laminate floor.  
This was in her high chair came in on the wall.  
Came in on the wall her high chair.  
Her high chair came in.



**Seq+Pruning:** High chair on the wall.  
**Seq+Tree+Pruning:** High chair on the wall.

Image 38



**Human:** Black and white for my christmas tree! Bought this wonderful decoration in Metz  
**HMM.v.1+Cogn:** One of the few ball embedded in the floor of the women for a little four year old. The orange ball suspended about the civic square in Wellington for a little four year old. Leftover banana sitting on a large cutting board in my kitchen pretty in my livingroom.  
**Seq.v.1+LingRule+Cogn:** I think this is just normal fruit.  
**HMM.v.2+Cogn:** This is a picture of the ball embedded in the floor of the women at the beach. Ball laying in the grass for a little four year old. Leftover banana sitting on a large cutting board in my kitchen pretty in my livingroom.  
**Seq.v.2+LingRule:** Just normal fruit at the beach.  
**Seq.v.2+LingRule+Cogn:** I think this is just normal fruit sitting on a large cutting board in my kitchen for a little four year old.  
**Seq:** Hang out all year round the orange ball in my pink library.  
**Seq+Tree:** The orange ball hang out all year round.  
**Seq+Pruning:** Laying in the grass the ball for a little four year.  
**Seq+Tree+Pruning:** The orange ball laying in the grass.

Image 39



**Human:** Small unidentified orange flower in Cusco, Peru.  
**HMM.v.1+Cogn:** Found these pretty yellow flowers taken in Tunisia by darker yellow leaves in flower. A bright orange died due to the summer drought by darker yellow leaves in flower.  
**Seq.v.1+LingRule+Cogn:** Found a bright orange died due to the summer drought in flower. Flower taken in Tunisia.  
**HMM.v.2+Cogn:** One of these pretty yellow flowers taken in Tunisia by darker yellow leaves in flower. Pineapples are all in a tree in flower.  
**Seq.v.2+LingRule:** A bright orange died due to the summer drought in flower. Flower taken in Tunisia.  
**Seq.v.2+LingRule+Cogn:** Found a bright orange died due to the summer drought in flower. Flower taken in Tunisia.  
**Seq:** Died due to the summer drought in a tree a good sweet apple. Flower in the garden over the roof and onto a tree glowing against the green leaf.  
**Seq+Tree:** A good sweet apple died due to the summer drought. A yellow flower bathed in beautiful bokeh.  
**Seq+Pruning:** A bright orange stage over the roof and a tree in flower. Flower over the roof and a tree.  
**Seq+Tree+Pruning:** A bright orange stage in flower. Yellow flower over the roof and a tree.

Image 40



**Human:** Passenger train from Wroclaw Główny to Poznan passes bridge over Odra river  
**HMM.v.1+Cogn:** This was a 4 car electric train driving through small Basque Country roads over the building in downtown Ocala Florida.  
**Seq.v.1+LingRule+Cogn:** I liked the way the little train driving through small Basque Country roads in downtown Ocala Florida.  
**HMM.v.2+Cogn:** I think this is a 4 car electric train driving through small Basque Country roads over the building in downtown Ocala Florida.  
**Seq.v.2+LingRule:** In downtown Ocala Florida of the nice deco buildings driving through small Basque Country roads the little train.  
**Seq.v.2+LingRule+Cogn:** This is one of the little train driving through small Basque Country roads in downtown Ocala Florida.  
**Seq:** Tracks while driving through small basque country roads the train by the market place of the house on the island of Kimmen.  
**Seq+Tree:** The train tracks while driving through small basque country roads.  
**Seq+Pruning:** Tracks while driving through small basque country roads the train by the market place of the house on the island of Kimmen.  
**Seq+Tree+Pruning:** The train tracks while driving through small basque country roads.

Image 41



**Human:** The pink and white rock against the blue sky was pretty amazing.  
**HMM.v.1+Cogn:** Found kitty sitting by the garden against a blue sky in the water. Clock tower set against a Tar Heel Blue sky against a blue sky in the water.  
**Seq.v.1+LingRule+Cogn:** The steeple building in armed concrete in 1901 in Sceaux France.  
**HMM.v.2+Cogn:** This is a picture of the cat sitting by the garden in the sky in the water. Clock tower set against a Tar Heel Blue sky against a blue sky in the water.  
**Seq.v.2+LingRule:** Their water tower set against a Tar Heel Blue sky in the field.  
**Seq.v.2+LingRule+Cogn:** One of their water tower set against a Tar Heel Blue sky in the field.  
**Seq:** A beautiful cat is adored by one of her kittens on the left in the water under the blue sky.  
**Seq+Tree:** A beautiful cat is adored by one of her kittens on the left.  
**Seq+Pruning:** Beautiful cat is of her kittens on the left in the water under the blue sky.  
**Seq+Tree+Pruning:** Beautiful cat is of her kittens on the left.

Image 42



**Human:** A girl who fell in love with a monkey  
**HMM.v.1+Cogn:** This is a shot of a wall clock manufactured in England in the background.  
**Seq.v.1+LingRule+Cogn:** Taken in front of a shiny dish.  
**HMM.v.2+Cogn:** Taken in front of the church clock manufactured in England in the background.  
**Seq.v.2+LingRule:** A shiny dish in bulilit ang lilitit advertisement.  
**Seq.v.2+LingRule+Cogn:** Taken in front of a shiny dish in the background.  
**Seq:** The plate on his thumb and a white sprinkle between his finger, doughnut in the background trying to hard.  
**Seq+Tree:** The plate is a swiss communal dish shared at the table in an earthenware pot over a small burner.  
**Seq+Pruning:** This clock on his thumb between his finger in the background.  
**Seq+Tree+Pruning:** This clock on his thumb between his finger in the background.

Image 43



**Human:** On the way up the mountain in the cable car  
**HMM.v.1+Cogn:** Here you can see the hills covered with green tea plants in Sri Lanka in the sky in the shot. Hills covered in quartz in the sky above fish lake.  
**Seq.v.1+LingRule+Cogn:** Here you can see the hills in the sky.  
**HMM.v.2+Cogn:** One of the hills overlooking the sea in the sky in the shot. Hills covered in quartz in the sky above fish lake.  
**Seq.v.2+LingRule:** The hills covered in quartz at the lovely clear blue sky.  
**Seq.v.2+LingRule+Cogn:** One of the hills covered in quartz at the lovely clear blue sky.  
**Seq:** The hill on the banks of lake lucerne in the sky is called haystack.  
**Seq+Tree:** The hill is called haystack.

**Seq+Pruning:** The hill in the shot at the lovely clear blue sky is.  
**Seq+Tree+Pruning:** The hill is a favorite.

Image 44



**Human:** I like the big hot dog right above the sign advertising Thai Food.  
**HMM.v.1+Cogn:** I like the way a car parked in front of that door constantly in this building in Weymouth. Police car parked in the TCC bus stop in this building in Weymouth.  
**Seq.v.1+LingRule+Cogn:** Saw this silver old classic car parked in the TCC bus stop. A car parked in front of that door constantly under the sign.  
**HMM.v.2+Cogn:** This is a car parked in front of that door constantly to the house in the background. Police car parked in the TCC bus stop in this building in Weymouth.  
**Seq.v.2+LingRule:** A car parked in front of that door constantly under the sign. Silver old classic car parked in the TCC bus stop of every building.  
**Seq.v.2+LingRule+Cogn:** This is a car parked in front of that door constantly under the sign. Silver old classic car parked in the TCC bus stop of every building.  
**Seq:** Down under the sign to the house with the scaffolding outside a car was parked in front of that door constantly. Cool silver is everywhere in the highway against building.  
**Seq+Tree:** A car was parked in front of that door constantly. Cool silver is everywhere in the highway.  
**Seq+Pruning:** Down under the sign to the house with a shop a car was parked in front of that door constantly. On every floor of every building a car is in the highway.  
**Seq+Tree+Pruning:** A car was parked in front of that door constantly. A car is in the highway.

Image 45



**Human:** Window in kitchen  
**HMM.v.1+Cogn:** Here you can see door leading into the building of the building in the center window. Bia window let in so much light it is just beautiful inside of the building in the center window.  
**Seq.v.1+LingRule+Cogn:** I think this is an window let in so much light it is just beautiful inside. The entrance door leading into the building.  
**HMM.v.2+Cogn:** This was taken at the door leading into the building of the building in the center window. Bia window let in so much light it is just beautiful inside of the building in the center window.  
**Seq.v.2+LingRule:** New windows above sink designed by Sr. The entrance door leading into the building.  
**Seq.v.2+LingRule+Cogn:** This is a picture of the window let in so much light it is just beautiful inside of the building. The entrance door leading into the building in the center window.  
**Seq:** Block windows by best block glass Block Service of St Louis the window in the building for our new RallITeK Shop customer area. Door of the building in the center window.  
**Seq+Tree:** The window block windows by best block glass Block Service of St Louis. In the center window door of the building.  
**Seq+Pruning:** Front door of the building in our new house.  
**Seq+Tree+Pruning:** Front door leading into the building.

Image 46



**Human:** The bottom floor the lease to a shop. And yes, they have homes at the beach and in the mountain too.  
**HMM.v.1+Cogn:** These is the view from these balconies hanging over the patio area in the sky with a message. Balcony hanging over the patio area in the sky with a message. A balcony hanging over the patio area in the sky with a message.  
**Seq.v.1+LingRule+Cogn:** The balconies were neat in the sky with a message.  
**HMM.v.2+Cogn:** These is the view from the balconies overlooking the swimming pool in the sky with a message. Balcony hanging over the patio area in the sky with a message. Balconies hanging over the patio area in the sky with a message.  
**Seq.v.2+LingRule:** Coveted balcony in Cruz living room against the blue sky.  
**Seq.v.2+LingRule+Cogn:** In the background and wonderful blue cloudy sky coveted balcony with a message.  
**Seq:** Coveted balcony like ufos against the blue sky of a building overlooking the crazy clock in Old Town Square in Prague.  
**Seq+Tree:** Coveted balcony overlooking the crazy clock in Old Town Square in Prague.  
**Seq+Pruning:** The balconies of a building in the sky.  
**Seq+Tree+Pruning:** The balconies overlooking the clock in Old Town.

Image 47



**Human:** Walking in defiantly, ready to order some chicken strips  
**HMM.v.1+Cogn:** This is a picture of the Health promotion bus passing our virtual office in London of the road by red train.  
**Seq.v.1+LingRule+Cogn:** Found this tourist bus passing our virtual office in London by red train.  
**HMM.v.2+Cogn:** This is a picture of the Health promotion bus passing our virtual office in London of the road by red train.  
**Seq.v.2+LingRule:** To the park entrance in the street is for tours to the houses of Hollywood stars a tourist bus.  
**Seq.v.2+LingRule+Cogn:** This is a tourist bus passing our virtual office in London by red train.  
**Seq:** An old double double-decker bus passing by our London virtual office building in the middle of the road to the park entrance.  
**Seq+Tree:** An old double double-decker bus passing by our London virtual office building.  
**Seq+Pruning:** The old buses go to the end in the middle of the road to the park entrance.  
**Seq+Tree+Pruning:** The old buses go to the end.

Image 48



**Human:** Nadia was projected onto a television screen while in Studio 1.  
**HMM.v.1+Cogn:** Love the best seat set in the balcony of our cottage in the sky by the sink. Broken brown couch set in the balcony of our cottage in the sky by the sink.  
**Seq.v.1+LingRule+Cogn:** Love the car seat set in the balcony of our cottage across the darkening sky.  
**HMM.v.2+Cogn:** I think this is the best seat set in the balcony of our cottage in the sky by the sink. Broken brown couch set in the balcony of our cottage in the sky by the sink.  
**Seq.v.2+LingRule:** By the sink not yet in the sky is a sofa bed the car seat.  
**Seq.v.2+LingRule+Cogn:** On the side of the car seat set in the balcony of our cottage by the sink.  
**Seq:** Couch is a sofa bed in the sky through the bedroom window.  
**Seq+Tree:** The sofa set in the balcony of our cottage.  
**Seq+Pruning:** The sofa in the sky is faded.  
**Seq+Tree+Pruning:** The sofa set in the balcony of our cottage.

Image 49



**Human:** Two wood duck drakes on a fresh water pond in central South Carolina.  
**HMM.v.1+Cogn:** Love the some ducks swimming along in the lake near goring.  
**Seq.v.1+LingRule+Cogn:** This is a shot of the white duck.  
**HMM.v.2+Cogn:** Found in some ducks swimming along in the lake near goring.  
**Seq.v.2+LingRule:** The white duck in black and white 1.  
**Seq.v.2+LingRule+Cogn:** This is one of the white duck swimming along in the lake.  
**Seq:** These two ducks are in the white duck and the seagull's shadow over gatehampton bridge near goring.  
**Seq+Tree:** These two ducks are in the white duck and the seagull's shadow.  
**Seq+Pruning:** These two ducks were in a small river off the Ala Wai Canal in Honolulu, Hawaii.  
**Seq+Tree+Pruning:** These two ducks were in a small river off the Ala Wai Canal in Honolulu, Hawaii.

Image 50



**Human:** Sam in his new favorite place-the bathroom shelf 12-5-2007  
**HMM.v.1+Cogn:** Here you can see the neighbors cat sitting by the bakery in 7th District Victory City in the sky under kitchen sink.  
**Seq.v.1+LingRule+Cogn:** Here you can see the neighbors cat sitting by the bakery in 7th District Victory City under kitchen sink.  
**HMM.v.2+Cogn:** This was one of the neighbors cat sitting by the bakery in 7th District Victory City in the sky under kitchen sink.  
**Seq.v.2+LingRule:** The neighbors cat likes to come in and visit under kitchen sink.  
**Seq.v.2+LingRule+Cogn:** This was one of the neighbors cat sitting by the bakery in 7th District Victory City under kitchen sink.  
**Seq:** For Peace Corps volunteers to use the cat sitting in the chair upstairs by the sky.  
**Seq+Tree:** The cat sitting in the chair upstairs.  
**Seq+Pruning:** For Peace Corps volunteers to use the cat sitting in the chair by the sky.  
**Seq+Tree+Pruning:** The cat sitting in the chair.

Image 51



**Human:** Dutch,16th century stained glass window in east wall of south transept showing eight scenes from the life of St Nicholas  
**HMM.v.1+Cogn:** Taken in front of stained glass moved here from elsewhere in church in tree in the east window. Glass window is the oldest episcopal church in N in tree in the east window. Glass window set into the south facing wall of St Nicholas Church in Fyfield Essex England in tree in the east window. Stained glass moved here from elsewhere in church in tree in the east window. The c15 east window set into the south facing wall of St Nicholas Church in Fyfield Essex England in the trees in the east window. This beautiful window moved here from elsewhere in church in tree in the east window. A stained glass window above the West entrance in tree in the east window.  
**Seq.v.1+LingRule+Cogn:** In the east window near Hoop Pine trees stained glass depicting St Swithun and St Ethelwood.  
**HMM.v.2+Cogn:** This photo was taken in a glass moved here from elsewhere in church in the trees in the east window. Stained glass windows in the Notre Dame in the trees in the east window. Glass window set into the south facing wall of St Nicholas Church in Fyfield Essex England in tree in the east window. Glass window is in the north wall of St Helens in the trees in the east window. Stained glass window above the West entrance in the trees in the east window. Stained glass is in the north wall of St Helens in the trees in the east window. A stained glass window window above the West entrance in tree in the east window.  
**Seq.v.2+LingRule:** Glass window viewing level of the Scott Monument in St Mary Sprothorough. A stained glass window moved here from elsewhere in church above the trees.  
**Seq.v.2+LingRule+Cogn:** Taken out of the glass window viewing level of the Scott Monument in St Mary Sprothorough. A stained glass window moved here from elsewhere in church above the trees.  
**Seq:** This beautiful window above the trees in the North wall of St Laurence church in Blackmore , Essex. Above the trees this beautiful window is set into the south facing wall of St Nicholas Church in Fyfield , Essex , England.  
**Seq+Tree:** This beautiful window is set into the south facing wall of St Nicholas Church in Fyfield , Essex , England.  
**Seq+Pruning:** This window in the building is the south facing wall.  
**Seq+Tree+Pruning:** This window is the south facing wall.

Image 52



**Human:** A rare glimpse inside lauras tower in shrewsbury castle on open heritage sunday  
**HMM.v.1+Cogn:** This is one of the glass windows looking onto the river by the building in the hallways. Stained glass window in the church in Cropredy in the building in the hallways. Stained glass sitting room by the building in the hallways.  
**Seq.v.1+LingRule+Cogn:** The window seen in the preceding photoand by the building for one glorious night.  
**HMM.v.2+Cogn:** One of the window falling on the building through the doorway. Stained glass window in the church in Cropredy in the building through the doorway. Stained glass sitting room by the building in the hallways.  
**Seq.v.2+LingRule:** Glass windows seen in the preceding photoand for one glorious night.  
**Seq.v.2+LingRule+Cogn:** This is one of the glass windows seen in the preceding photoand for one glorious night.  
**Seq:** Sitting room in the building through the doorway the window.  
**Seq+Tree:** The window was the kitchen , then in front theres windows looking onto the river.  
**Seq+Pruning:** Sitting room the windows in the building through the doorway.  
**Seq+Tree+Pruning:** The window was the kitchen , then in front theres windows looking onto the river.

Image 53



**Human:** Osprey fishing over Packer Lake. The fish were sick with &quot;ick&quot; and the osprey were having a field day!  
**HMM.v.1+Cogn:** These is the photo of the largest birds hovering over our boat as we cruised along in the sky on the pond surface. Frigate bird hovering over our boat as we cruised along in the sky on the pond surface.  
**Seq.v.1+LingRule+Cogn:** These is the photo of the largest birds hovering over our boat as we cruised along in the sky. The largest birds hovering over our boat as we cruised along on the pond surface.  
**HMM.v.2+Cogn:** This is a bird sitting in the treetops above the Minnesota River in the sky in the background. Frigate bird hovering over our boat as we cruised along in the sky on the pond surface.  
**Seq.v.2+LingRule:** The largest birds hovering over our boat as we cruised along on the pond surface. The largest birds hovering over our boat as we cruised along in the sky.  
**Seq.v.2+LingRule+Cogn:** One of the largest birds hovering over our boat as we cruised along on the pond surface. The largest birds hovering over our boat as we cruised along in the sky.  
**Seq:** Flying over California Poppy Fields this bald eagle in the sky above the female while she deposited eggs on the pond surface. Frigate birds flying above our boat big bird in the sky.  
**Seq+Tree:** This bald eagle flying over California Poppy Fields. Big bird flying over California Poppy Fields.  
**Seq+Pruning:** Flying above our boat a small birds in the sky for the hummingbirds. Little bird in the sky flying over California Poppy Fields.  
**Seq+Tree+Pruning:** A small birds flying over California Poppy Fields. Little bird flying over California Poppy Fields.

Image 54



**Human:** With experiments in orange on the wall  
**HMM.v.1+Cogn:** I like the way the woman biting down on my medal with Seattle tallest building by the ocean. 3 mayan women walking to the beach with their daughters in hand with Seattle tallest building by the ocean.  
**Seq.v.1+LingRule+Cogn:** I like the way the woman biting down on my medal by the ocean. The woman walking to the beach with his daughters in hand.  
**HMM.v.2+Cogn:** One of the woman walking to the beach with his daughters in hand of the house in the background. 3 mayan women walking to the beach with their daughters in hand with Seattle tallest building by the ocean.  
**Seq.v.2+LingRule:** One less strong person looks really like a ball in winter wear by the ocean. The woman walking to the beach with his daughters in hand of the Star House building.  
**Seq.v.2+LingRule+Cogn:** One of the woman biting down on my medal by the ocean. The woman walking to the beach with his daughters in hand of the Star House building.  
**Seq:** The woman in front of the Star House building , while tourists await to cross the street to the garage is me biting down on my ` medal ' - . Walking to the beach with his daughters in hand the woman of the house and a soffit.  
**Seq+Tree:** The woman is me biting down on my ` medal ' - . A man walking to the beach with his daughters in hand.  
**Seq+Pruning:** One person pulls the floor of the house and the screen in the rooms wooden bungalow. A man of the house and the screen became comfy in the cute granny chair.  
**Seq+Tree+Pruning:** One person is me biting down on my ` medal ' - . A man became comfy in the cute granny chair.

Image 55

**Human:** A flower along the road near Kailua-Kona s business district.



**HMM.v.1+Cogn:** I liked the way cat wrapped up in a towel on my lap over grass in her keyboard box.  
**Seq.v.1+LingRule+Cogn:** Saw this yellow flower blooming beautifully in my office garden over grass.  
**HMM.v.2+Cogn:** This is a picture of the cat wrapped up in a towel on my lap over grass in her keyboard box.  
**Seq.v.2+LingRule:** In a rock this yellow flower shown all season.  
**Seq.v.2+LingRule+Cogn:** Found this yellow flower blooming beautifully in my office garden of my camera.  
**Seq:** Over grass has shown all season a flower in the background , confident in her keyboard box.  
**Seq+Tree:** This flower was blooming beautifully in my office garden.  
**Seq+Pruning:** Yellow flower in the field.  
**Seq+Tree+Pruning:** Yellow flower found in our friends.

Image 56



**Human:** Young man sitting in the park under the tree near his bicycle  
**HMM.v.1+Cogn:** I liked the way bikes made in Wright Brothers bike shop in Dayton before in tree in New Babbage. The bike trail across the street from my sister house in Madison in the apple tree in a tree. Green vinyl record wall clock is green against her green walls in a plum tree in New Babbage.  
**Seq.v.1+LingRule+Cogn:** The bike of the pine trees in New Babbage rests against a colorful wall Hue Vietnam.  
**HMM.v.2+Cogn:** One of the bike sitting in barn in the grass for a little bit. The bike sitting in barn in the grass for a little bit. Green vinyl record wall clock is green against her green walls in a plum tree in New Babbage.  
**Seq.v.2+LingRule:** For a little bit in the grass rests against a colorful wall Hue Vietnam the bike.  
**Seq.v.2+LingRule+Cogn:** One of the bike ordered for the rapper Juvenile wife who live in Slidell on the rock.  
**Seq:** Black bike rests against a colorful wall , Hue , Vietnam of pilot Mountain from the Jomeokee trail showing the mountain rising above the tops of the pine trees in a tree , Shenandoah National Park.  
**Seq+Tree:** Black bike trail across the street from my sister 's house in Madison.  
**Seq+Pruning:** My bike in a tree.  
**Seq+Tree+Pruning:** My bike ride in many roads and high hills.

Image 57



**Human:** H happily rests his amput on a warm Gatorade bottle of water (a small bottle wrapped in a rag).  
**HMM.v.1+Cogn:** I liked the way cat basking in the sun on crepe paper under the Christmas tree in my recliner. Cats tv 2 years later by isewcute in my recliner.  
**Seq.v.1+LingRule+Cogn:** Taken in front of my cat sitting in a shoe box. Cat likes hanging around in my recliner.  
**HMM.v.2+Cogn:** One of the cat sitting in a shoe box in the market. Cats tv 2 years later by isewcute in my recliner.  
**Seq.v.2+LingRule:** Cat likes hanging around. My cat is in the bag.  
**Seq.v.2+LingRule+Cogn:** This is a picture of the cat sleeping locations. My cat is in the bag.  
**Seq:** The cat curled up in an old wooden bowl. The cat is in the bag on top of the main light source.  
**Seq+Tree:** The cat curled up in an old wooden bowl. The cat is in the bag.  
**Seq+Pruning:** The cat has the locations in the butterfly exhibit.  
**Seq+Tree+Pruning:** The cat has the locations.

Image 58



**Human:** The water was like glass and crystal clear! The Garden Wall in pictured in the distance.  
**HMM.v.1+Cogn:** Love this Little boat docked off of Cape Porpoise Pier by the lighthouse in the open ocean in the lake.  
**Seq.v.1+LingRule+Cogn:** Found this that boat docked off of Cape Porpoise Pier by the lighthouse in the open ocean.  
**HMM.v.2+Cogn:** Love the Little boat docked off of Cape Porpoise Pier by the lighthouse in the open ocean in the lake.  
**Seq.v.2+LingRule:** Near chong rock in the water docked off of Cape Porpoise Pier by the lighthouse that boat.  
**Seq.v.2+LingRule+Cogn:** Found that boat docked off of Cape Porpoise Pier by the lighthouse in the lake on the water.  
**Seq:** Coming home that boat in the water by the road.  
**Seq+Tree:** That boat was a perpetual tenant at the shores of the beach in front of my cabin.  
**Seq+Pruning:** Boat around the lake by the road was a tenant at the shores of the beach in front of my cabin.  
**Seq+Tree+Pruning:** That boat was a tenant at the shores of the beach in front of my cabin.

Image 59



**Human:** I adjusted all the colors in this photo except the wine bottle and glass  
**HMM.v.1+Cogn:** Here you can see the bottle done in orange and pink in the bathtub.  
**Seq.v.1+LingRule+Cogn:** View from big wet bottle done in orange and pink near DK.  
**HMM.v.2+Cogn:** One of the bottle done in orange and pink in the bathtub.  
**Seq.v.2+LingRule:** Big wet bottle in classroom.  
**Seq.v.2+LingRule+Cogn:** Found a big wet bottle done in orange and pink.  
**Seq:** Lit by incoming kitchen light a bottle in the bathtub.  
**Seq+Tree:** A bottle left on table right near the sign about cleaning up your rubbish - ironic.  
**Seq+Pruning:** Left on table right near the sign about cleaning up your rubbish - ironic a bottle in the bathtub.  
**Seq+Tree+Pruning:** A bottle left on table right near the sign about cleaning up your rubbish - ironic.

Image 60



**Human:** A lonely horse stand in a field next to glendalough church and tower etc.  
**HMM.v.1+Cogn:** I liked the way horse had one blue eye in her field like Jason and myself.  
**Seq.v.1+LingRule+Cogn:** Found winning shire horses had one blue eye like Jason and myself.  
**HMM.v.2+Cogn:** This is a picture of the horse had one blue eye in her field like Jason and myself.  
**Seq.v.2+LingRule:** In the foreground of the field had one blue eye winning shire horses.  
**Seq.v.2+LingRule+Cogn:** Found the winning shire horses had one blue eye like Jason and myself.  
**Seq:** A horse in a field from the house up to the high plateau above the lake.  
**Seq+Tree:** A horse was tied right next to the road in the Rocha region between La Pedrera and Cabo Polonio.  
**Seq+Pruning:** A horse in a field from the house up to the high plateau above the lake had bear bells.  
**Seq+Tree+Pruning:** Some horses had bear bells.

Image 61

**Human:** North window of the aisle has stained glass by Arthur S Walker 1951[Full information on the Stoke Gifford set homepage]  
**HMM.v.1+Cogn:** Taken in front of stained glass building against the old square brick building in Dearborn Michigan. Stained glass free to the building in Dearborn Michigan. The classic style window seen inside the cathedral of Cologne Germany to the building in Dearborn Michigan.



**Seq.v.1+LingRule+Cogn:** Love the four stained seen inside the cathedral of Cologne Germany. The classic style window seen inside the cathedral of Cologne Germany.

**HMM.v.2+Cogn:** These is the picture of the windows building to the building in the sky. Stained glass free to the building in Dearborn Michigan. Glass windows broken to the building in the sky.

**Seq.v.2+LingRule:** The window free in Dearborn Michigan. The classic style window seen inside the cathedral of Cologne Germany to the building.

**Seq.v.2+LingRule+Cogn:** One of four stained seen inside the cathedral of Cologne Germany in Dearborn Michigan. The classic style window seen inside the cathedral of Cologne Germany to the building.

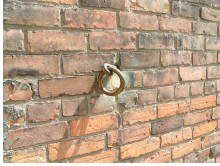
**Seq:** A window of the sharp ,sleek , modern glass tower against the old , square , brick building known as the *Dom <* , seen in April 2000.

**Seq+Tree:** The random window known as the *Dom <* , seen in April 2000.

**Seq+Pruning:** By Jim Monroe the window to the building known as the *Dom <* , seen in April 2000.

**Seq+Tree+Pruning:** The window known as the *Dom <* , seen in April 2000.

Image 62



**Human:** This ring is on a downtown building in Sudbury, Ontario. It is a horse hitch.

**HMM.v.1+Cogn:** Love the table had springtime napkins in yellow pink and turquoise of dylan old house.

**Seq.v.1+LingRule+Cogn:** Love the table of dylan old house.

**HMM.v.2+Cogn:** One of the table had springtime napkins in yellow pink and turquoise of dylan old house.

**Seq.v.2+LingRule:** Water table had springtime napkins in yellow pink and turquoise.

**Seq.v.2+LingRule+Cogn:** One of the table had springtime napkins in yellow pink and turquoise.

**Seq:** Outdoors near Arezzo , Tuscany , Italy the table in my friend Rachel.

**Seq+Tree:** The table had springtime napkins in yellow , pink and turquoise.

**Seq+Pruning:** The table underneath the bridge outside of dylan 's old house becomes its legs.

**Seq+Tree+Pruning:** The table becomes its legs.

Image 63



**Human:** I with my four mother in laws - in sky blue , pink ,green and blue and green

**HMM.v.1+Cogn:** One of the few boy helping her father cathing fish in the lake in a nearby building in the red. A boy running in pants at Global Green house by about 7 pm. A happy girl bought at the game against the Colorado Rockies in a nearby building in the red. A little girl dreaming in a nearby building by about 7 pm. The girl is on the phone in a nearby building in the red.

**Seq.v.1+LingRule+Cogn:** The first flower girl bought at the game against the Colorado Rockies in a nearby building.

**HMM.v.2+Cogn:** This is a picture of the boy helping her father cathing fish in the lake in the house in the red. A boy running in pants at Global Green house by about 7 pm. A happy girl stayed in the house in the red. She stayed in the house in the red. The girl is on the phone in the house in the red.

**Seq.v.2+LingRule:** The little girl by about 7 pm at Global Green house. The white people keep innocent in this colorful world at Global Green house.

**Seq.v.2+LingRule+Cogn:** This is a picture of the old man running in pants at Global Green house. Boy in the door at Global Green house.

**Seq:** The girl in the house is on the phone. Old man in the house by about 7 pm.

**Seq+Tree:** The girl is on the phone. Old man got him some traditional Malay clothes and he looks cute in it.

**Seq+Pruning:** Shows off his rare green Cubs hat that he bought at the game against the Colorado Rockies in the house the first flower. Cross here in the house the first flower.

**Seq+Tree+Pruning:** The first flower shows off his rare green Cubs hat that he bought at the game against the Colorado Rockies. The first flower was walking in the middle of the street.

Image 64



**Human:** Colorful sign in Old Town

**HMM.v.1+Cogn:** One of the few ball sitting in the grass in the front window. The fruit given to me by Sharon in 2008sushi bar. ME plates put pizza on in 2008sushi bar.

**Seq.v.1+LingRule+Cogn:** The white shifter ball.

**HMM.v.2+Cogn:** One of the ball sitting in the grass in the front window. The fruit given to me by Sharon in a glass. ME plates put pizza on in 2008sushi bar.

**Seq.v.2+LingRule:** A toy ball in train car.

**Seq.v.2+LingRule+Cogn:** This is a toy ball sitting in the grass.

**Seq:** Ball in a glass made from logs.

**Seq+Tree:** In a glass ball made from logs.

**Seq+Pruning:** Ball sitting in the grass.

**Seq+Tree+Pruning:** Ball in a glass.

Image 65



**Human:** Our kitten, Nala, siting all proper by the front door of our house.

**HMM.v.1+Cogn:** View of a tabby cat taken over the girls blankies to the building in our hotel room. My cat morning to the building upon completion.

**Seq.v.1+LingRule+Cogn:** The stray cat morning. The feral cats sit down near the window.

**HMM.v.2+Cogn:** View of a tabby cat taken over the girls blankies to the building in our hotel room. Cat morning to the building upon completion.

**Seq.v.2+LingRule:** The stray cat by the glass buildings. The cat in our hotel room taken over the girls blankies.

**Seq.v.2+LingRule+Cogn:** This photo was taken in the morning in our hotel room my cat. The cat to the building taken over the girls blankies.

**Seq:** Sit down near the window of our house the cat. My cat in charge of the many doors to the building , this one siting by the garden.

**Seq+Tree:** The cat sitting in one of his favorite places in the spare bedroom.

**Seq+Pruning:** The cat sitting in his favorite places in spare bedroom to the building by trees and trees. Cat sitting by the garden of our house over the next days in our hotel room.

**Seq+Tree+Pruning:** The cat sitting in his favorite places in spare bedroom. In our hotel room cat to the building by trees and trees.

Image 66



**Human:** Around market house

**HMM.v.1+Cogn:** I like these cars got stuck in the sand in the water by the river. My old blue car waiting for the ferry by the water in anaglyph 3d stereo red blue cyan glasses. This car is really cool it one of the RARE cars here in Syria in the river by the river.

**Seq.v.1+LingRule+Cogn:** One of the few these fun cars by the water waiting for the ferry. Old car shows a nice contrast against the gray church wall by the water.

**HMM.v.2+Cogn:** These is the picture of the cars got stuck in the sand in the water by the river. My old blue car waiting for the ferry by the water in anaglyph 3d stereo red blue cyan glasses. This car parked in front of the Museum in the water by the river.

**Seq.v.2+LingRule:** Many road rally cars take you anywhere you need to go around North Hills bags and all of the pool. Old car shows a nice contrast against the gray church wall by the water.

**Seq.v.2+LingRule+Cogn:** One of these fun cars waiting for the ferry of the pool. Old car shows a nice contrast against the gray church wall by the water.

**Seq:** Old car spotted on the way to the park by the river in the cooling water. The cars will take you anywhere in the river.

**Seq+Tree:** His car parked in ballater. The cars will take you anywhere.

**Seq+Pruning:** Old car in the river by the truck towards the cars waiting for the ferry. The cars by the water will need to go around North Hills bags.

**Seq+Tree+Pruning:** Old car shows a contrast the gray church wall. These cars will need to go around North Hills bags.

Image 67



**Human:** This is a copy of the Capitol building in Washington...I know which one I d rather be standing outside!  
**HMM.v.1+Cogn:** I like these cars got in the way on the street with a mobile phone. Wooden cottage type lighthouse seen above the courtyard at the entrance to the studios on the street in Wintersville Ohio.  
**Seq.v.1+LingRule+Cogn:** I like these cars got in the way of Everton.  
**HMM.v.2+Cogn:** One of our car got in the way on the street with a mobile phone. Wooden cottage type lighthouse seen above the courtyard at the entrance to the studios on the street in Wintersville Ohio.  
**Seq.v.2+LingRule:** Near the Times building cars in the water.  
**Seq.v.2+LingRule+Cogn:** These is under the blue sky got in the way cars.  
**Seq:** Left of the cars involved in the front straight crash at the green flag on the street with a mobile phone that blue car.  
**Seq+Tree:** That blue car is a KFC delivery guy.  
**Seq+Pruning:** Blue car is a guy on the street for a pizza place in Wintersville , Ohio.  
**Seq+Tree+Pruning:** Blue car is a guy.

---

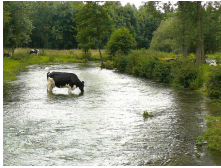
Image 68



**Human:** This cow was walking in the fields behind our home in the eveninglight...Added some drama with Topaz...  
**HMM.v.1+Cogn:** These is the first cows showed up in the sky in the field grass. This cow eating grass in rim in the sky in the field grass.  
**Seq.v.1+LingRule+Cogn:** Taken from beautiful cows showed up in the field grass.  
**HMM.v.2+Cogn:** One of the cow showed up in the sky in the field grass. This cow eating grass in rim in the sky in the field grass.  
**Seq.v.2+LingRule:** Beautiful cows showed up in the field grass.  
**Seq.v.2+LingRule+Cogn:** The beautiful cows showed up in the field grass.  
**Seq:** Beautiful cows in the sky looking down on me from near the low Bradfield bus.  
**Seq+Tree:** Beautiful cows being led down the gorge's river in the Rhodopi mountains , Bulgaria.  
**Seq+Pruning:** Cows in the field grass.  
**Seq+Tree+Pruning:** The cow was the plants in the water.

---

Image 69



**Human:** A lush scene enlivened by the cows that walked across the river from time to time to test the grass on the other side  
**HMM.v.1+Cogn:** Here you can see the cows peeping at me in the trees in the building.  
**Seq.v.1+LingRule+Cogn:** Here you can see the cows in the building under tree.  
**HMM.v.2+Cogn:** One of the cows peeping at me in the trees in the building.  
**Seq.v.2+LingRule:** In tall grass the cows showed up near the rock pools.  
**Seq.v.2+LingRule+Cogn:** One of the cows eating waste in India of the Hoh River.  
**Seq:** Herded by 3 cowboys on horses the cows in the field near the rock pools.  
**Seq+Tree:** A cow is eating waste in India.  
**Seq+Pruning:** The cows in the field by the end of my street.  
**Seq+Tree+Pruning:** Holy cow is eating waste in India.

---

Image 70



**Human:** K-os in her white shirt and Fiddy in his pantaloons  
**HMM.v.1+Cogn:** Taken in front of my cat posed in the window of the Nathaniel of Colorado hat shop in downtown Mancos this morning Nov in the sky in the car.  
**Seq.v.1+LingRule+Cogn:** One of cute young cat posed in the window of the Nathaniel of Colorado hat shop in downtown Mancos this morning Nov over spring break.  
**HMM.v.2+Cogn:** Taken in front of all these extra cat posed in the window of the Nathaniel of Colorado hat shop in downtown Mancos this morning Nov in the sky in the car.  
**Seq.v.2+LingRule:** Cute young cat posed in the window of the Nathaniel of Colorado hat shop in downtown Mancos this morning Nov over spring break.  
**Seq.v.2+LingRule+Cogn:** Found a cute young cat posed in the window of the Nathaniel of Colorado hat shop in downtown Mancos this morning Nov over spring break.  
**Seq:** Beautifully framed by matching gray tree bark this cat in the car under a blue sky.  
**Seq+Tree:** This cat hanging out under the palm tree in southern California.  
**Seq+Pruning:** Hanging under the palm tree in the sky this cat.  
**Seq+Tree+Pruning:** This cat hanging under the palm tree.



## BIBLIOGRAPHY

- [Aker and Gaizauskas2008] Ahmet Aker and Robert Gaizauskas. 2008. Evaluating automatically generated user-focused multi-document summaries for geo-referenced images. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, pages 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Aker and Gaizauskas2010] A. Aker and R. Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In *ACL*.
- [Aletras and Stevenson2013] Nikolaos Aletras and Mark Stevenson. 2013. Representing topics using images. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 158–167, Atlanta, Georgia, June. Association for Computational Linguistics.
- [Arkin1998] R.C. Arkin. 1998. *Behavior-based Robotics*. Bradford book. MIT Press.
- [Barnard et al.2003] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M Blei, and Michael I Jordan. 2003. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.
- [Baroni and Zamparelli2010] Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Stroudsburg, PA, USA.
- [Barzilay and McKeown2005] Regina Barzilay and Kathleen McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.
- [Belz and Reiter2006] Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics.
- [Bigham et al.2006] Jeffrey P Bigham, Ryan S Kaminsky, Richard E Ladner, Oscar M Danielsson, and Gordon L Hempton. 2006. Webinsight:: making web images accessible. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 181–188. ACM.
- [Bigham et al.2008] Jeffrey P Bigham, Craig M Prince, and Richard E Ladner. 2008. We-banywhere: a screen reader on-the-go. In *Proceedings of the 2008 international cross-disciplinary conference on Web accessibility (W4A)*, pages 73–82. ACM.
- [Borodin et al.2010] Yevgen Borodin, Jeffrey P. Bigham, Glenn Dausch, and I. V. Ramakrishnan. 2010. More than meets the eye: a survey of screen-reader browsing strategies. In *W4A*, page 13.

- [Brants and Franz.2006] Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. In *Linguistic Data Consortium*.
- [Brooks1985] Rodney A. Brooks. 1985. A robust layered control system for a mobile robot. Technical report, Cambridge, MA, USA.
- [Brooks1990] Rodney A. Brooks. 1990. Elephants don't play chess. *Robot. Auton. Syst.*, 6(1-2):3–15, June.
- [Chomsky1968] Noam Chomsky. 1968. *Language and mind*. Harcourt Brace Jovanovich New York.
- [Clarke and Lapata2006] James Clarke and Mirella Lapata. 2006. Constraint-based sentence compression: An integer programming approach. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 144–151, Sydney, Australia, July. Association for Computational Linguistics.
- [Clarke and Lapata2008] James Clarke and Mirella Lapata. 2008. Global inference for sentence compression an integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.
- [Clarke2012] Daoud Clarke. 2012. A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1):41–71.
- [Cocke1969] John Cocke. 1969. *Programming Languages and Their Compilers: Preliminary Notes*. Courant Institute of Mathematical Sciences, New York University.
- [Cohn and Lapata2007] Trevor Cohn and Mirella Lapata. 2007. Large margin synchronous generation and its application to sentence compression. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 73–82, Prague, Czech Republic, June. Association for Computational Linguistics.
- [Cohn and Lapata2008] Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 137–144, Manchester, UK, August. Coling 2008 Organizing Committee.
- [Cohn and Lapata2009] Trevor Cohn and Mirella Lapata. 2009. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674.
- [Dalal and Triggs2005] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA. IEEE Computer Society.
- [Datta et al.2006] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *Computer Vision—ECCV 2006*, pages 288–301. Springer Berlin Heidelberg.

- [de Marnee and Manning2008] Marie-Catherine de Marnee and Christopher D. Manning. 2008. Stanford typed dependencies manual.
- [Delaitre et al.2010] Vincent Delaitre, Ivan Laptev, and Josef Sivic. 2010. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *Proceedings of the British Machine Vision Conference*, pages 97.1–97.11. BMVA Press. doi:10.5244/C.24.97.
- [Deng et al.2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition*.
- [Deng et al.2012] Jia Deng, Jonathan Krause, Alexander C. Berg, and L. Fei-Fei. 2012. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *Conference on Computer Vision and Pattern Recognition*.
- [Denkowski and Lavie2011] Michael Denkowski and Alon Lavie. 2011. 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*.
- [Dhar et al.2011] S. Dhar, V. Ordonez, and T. L. Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '11*, pages 1657–1664, Washington, DC, USA. IEEE Computer Society.
- [Dindo and Zambuto2010] Haris Dindo and Daniele Zambuto. 2010. A probabilistic approach to learning a visually grounded language model through human-robot interaction. In *IROS*, pages 790–796. IEEE.
- [Dodge et al.2012] Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daume III, Alex Berg, and Tamara Berg. 2012. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 762–772, Montréal, Canada, June. Association for Computational Linguistics.
- [Elliott and Keller2013] Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *EMNLP*, pages 1292–1302.
- [Elliott and Keller2014] Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association of Computational Linguistics (ACL-2014)*, Baltimore, Maryland, June.
- [Everingham et al.2010] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June.
- [Farhadi et al.2010] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young1, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: generating sentences for images. In *European Conference on Computer Vision*.

- [Fellbaum1998] Christiane D. Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.
- [Felzenszwalb et al.] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. <http://people.cs.uchicago.edu/~pff/latent-release4/>.
- [Feng and Lapata2008] Yansong Feng and Mirella Lapata. 2008. Automatic image annotation using auxiliary text information. In *Proceedings of ACL-08: HLT*, pages 272–280, Columbus, Ohio.
- [Feng and Lapata2010a] Yansong Feng and Mirella Lapata. 2010a. How many words is a picture worth? automatic caption generation for news images. In *Association for Computational Linguistics*.
- [Feng and Lapata2010b] Yansong Feng and Mirella Lapata. 2010b. How many words is a picture worth? automatic caption generation for news images. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, *ACL*, pages 1239–1249. The Association for Computer Linguistics.
- [Feng and Lapata2010c] Yansong Feng and Mirella Lapata. 2010c. Topic models for image annotation and text illustration. In *Human Language Technologies*.
- [Feng and Lapata2010d] Yansong Feng and Mirella Lapata. 2010d. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839, Los Angeles, California, June. Association for Computational Linguistics.
- [Feng and Lapata2010e] Yansong Feng and Mirella Lapata. 2010e. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 91–99, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Feng and Lapata2013] Yansong Feng and Mirella Lapata. 2013. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812.
- [Filippova and Altun2013] Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *EMNLP*, pages 1481–1491.
- [Galley and McKeown2007] Michel Galley and Kathleen McKeown. 2007. Lexicalized Markov grammars for sentence compression. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 180–187, Rochester, New York, April. Association for Computational Linguistics.
- [Guadarrama et al.2013] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the 14th International Conference on Computer Vision (ICCV-2013)*, pages 2712–2719, Sydney, Australia, December.

- [Guevara2011] Emiliano Guevara. 2011. Computing semantic compositionality in distributional semantics. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, pages 135–144. Citeseer.
- [Hafiz and Tudor1989] F. M. Hafiz and Ian Tudor. 1989. Extensive reading and the development of language skills. *ELT Journal*, 43(1):4–13.
- [Hawes et al.2007] Nick Hawes, Aaron Sloman, Jeremy Wyatt, Michael Zillich, Henrik Jacobsson, Geert-Jan M Kruijff, Michael Brenner, Gregor Berginc, and Danijel Skocaj. 2007. Towards an integrated robot with multiple cognitive functions. In *AAAI*, volume 7, pages 1548–1553.
- [Hoiem et al.2005] Derek Hoiem, Alexei A. Efros, and Martial Hebert. 2005. Geometric context from a single image. In *ICCV*, pages 654–661.
- [Huang et al.2012] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
- [ILOG, Inc2006] ILOG, Inc. 2006. ILOG CPLEX: High-performance software for mathematical programming and optimization. See <http://www.ilog.com/products/cplex/>.
- [Jamieson et al.2010] Michael Jamieson, Afsaneh Fazly, Suzanne Stevenson, Sven J. Dickinson, and Sven Wachsmuth. 2010. Using language to learn structured appearance models for image annotation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(1):148–164.
- [Joshi et al.2006] Dhiraj Joshi, James Z Wang, and Jia Li. 2006. The story picturing engine—a system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):68–89.
- [Kant1998] Immanuel Kant. 1998. *Critique of Pure Reason*. The Cambridge Edition of the Works of Immanuel Kant. Cambridge University Press, New York, NY. Translated by Paul Guyer and Allen W. Wood.
- [Kasami1965] Tadao Kasami. 1965. An efficient recognition and syntax-analysis algorithm for context-free languages. Technical report, Air Force Cambridge Research Lab, Bedford, MA.
- [Klein and Manning2003] Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430. Association for Computational Linguistics.
- [Knight and Marcu2000] Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *AAAI/IAAI*, pages 703–710.
- [Kojima et al.2002] Atsuhiko Kojima, Takeshi Tamura, and Kunio Fukunaga. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *IJCV*, 50.

- [Kovashka et al.2012] Adriana Kovashka, Devi Parikh, and Kristen Grauman. 2012. Whittlesearch: Image Search with Relative Attribute Feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.
- [Krishnamoorthy et al.2013] Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond J. Mooney, Kate Saenko, and Sergio Guadarrama. 2013. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*.
- [Kulkarni et al.2011] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Babytalk: Understanding and generating simple image descriptions. In *Conference on Computer Vision and Pattern Recognition*.
- [Kuznetsova et al.2012] Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 359–368, Jeju Island, Korea, July. Association for Computational Linguistics.
- [Kuznetsova et al.2013a] Polina Kuznetsova, Jianfu Chen, and Yejin Choi. 2013a. Understanding and quantifying creativity in lexical composition. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- [Kuznetsova et al.2013b] Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2013b. Generalizing image captions for image-text parallel corpus. In *The 51st Annual Meeting of the Association for Computational Linguistics - Short Papers*, Sofia, Bulgaria, August. Association for Computational Linguistics.
- [Kuznetsova et al.2014] Polina Kuznetsova, Vicente Ordonez, Tamara Berg, and Yejin Choi. 2014. Treetalk: Composition and compression of trees for image descriptions.
- [Lazebnik et al.2006] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching. In *Conference on Computer Vision and Pattern Recognition*, June.
- [Leong et al.2010] Chee Wee Leong, Rada Mihalcea, and Samer Hassan. 2010. Text mining for automatic image tagging. In *COLING*.
- [Leung and J.1999a] T. K. Leung and Malik J. 1999a. Recognizing surfaces using three-dimensional textons. In *ICCV*.
- [Leung and J.1999b] T. K. Leung and Malik J. 1999b. Recognizing surfaces using three-dimensional textons. In *ICCV*.
- [Li et al.2010] Li-Jia Li, Hao Su, Eric P. Xing, and Fei-Fei Li. 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, pages 1378–1386.
- [Li et al.2011] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, Portland, Oregon, USA, June. Association for Computational Linguistics.

- [Lindberg1981] D.C. Lindberg. 1981. *Theories of Vision from Al-kindi to Kepler*. Chicago History of Science and Medicine. University of Chicago Press.
- [Lowe2004a] D. G. Lowe. 2004a. Distinctive image features from scale invariant keypoints. *IJCV*.
- [Lowe2004b] David G. Lowe. 2004b. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November.
- [Martins and Smith2009] Andre Martins and Noah A. Smith. 2009. Summarization with a joint model for sentence extraction and compression. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.
- [Mason and Charniak2013] Rebecca Mason and Eugene Charniak. 2013. Annotation of on-line shopping images without labeled training examples. In *Proceedings of Workshop on Vision and Language*, Atlanta, Georgia, June. Association for Computational Linguistics.
- [Mason2013] Rebecca Mason. 2013. Domain-independent captioning of domain-specific images. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 69–76, Atlanta, Georgia, June. Association for Computational Linguistics.
- [Matuszek et al.2012a] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. 2012a. A Joint Model of Language and Perception for Grounded Attribute Learning. *ArXiv e-prints*, June.
- [Matuszek et al.2012b] Cynthia Matuszek, Nicholas FitzGerald, Luke S. Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012b. A joint model of language and perception for grounded attribute learning. In *ICML*. icml.cc / Omnipress.
- [McDonald2006] Ryan T. McDonald. 2006. Discriminative sentence compression with soft syntactic evidence. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics.
- [Mitchell and Lapata2008] Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *In Proceedings of ACL-08: HLT*, pages 236–244.
- [Mitchell and Lapata2009] Jeff Mitchell and Mirella Lapata. 2009. Language models based on semantic composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 430–439. Association for Computational Linguistics.
- [Mitchell et al.2012] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé, III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 747–756, Stroudsburg, PA, USA. Association for Computational Linguistics.

- [Monner and Reggia2011] Derek D. Monner and James A. Reggia. 2011. Systematically grounding language through vision in a deep, recurrent neural network. In *Proceedings of the 4th international conference on Artificial general intelligence*, AGI'11, pages 112–121, Berlin, Heidelberg. Springer-Verlag.
- [Mooney2008] Raymond J. Mooney. 2008. Learning to connect language and perception. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1598–1601, Chicago, IL, July. Senior Member Paper.
- [Oliva and Torralba2001] Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*.
- [Ordonez et al.2011] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.
- [Ordonez et al.2013] Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. From large scale image categorization to entry-level categories. In *IEEE International Conference on Computer Vision (ICCV)*.
- [Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.
- [Parikh and Grauman2013] Devi Parikh and Kristen Grauman. 2013. Implied feedback: Learning nuances of user behavior in image search. In *ICCV*, pages 745–752.
- [Pastra et al.2003] Katerina Pastra, Horacio Saggion, and Yorick Wilks. 2003. Nlp for indexing and retrieval of captioned photographs. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 143–146. Association for Computational Linguistics.
- [Perzanowski et al.2001] Dennis Perzanowski, Alan C Schultz, William Adams, Elaine Marsh, and Magda Bugajska. 2001. Building a multimodal human-robot interface. *Intelligent Systems, IEEE*, 16(1):16–21.
- [Petrov and Klein2007] Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *HLT-NAACL*.
- [Petrov et al.2006] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *COLING/ACL*.
- [Roth and Yih2004] D. Roth and W. Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*.
- [Roy2002] Deb K. Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, In review.
- [Rudolph and Giesbrecht2010] Sebastian Rudolph and Eugenie Giesbrecht. 2010. Compositional matrix-space models of language. In *Proceedings of the 48th Annual Meeting of the*



*Association for Computational Linguistics*, pages 907–916. Association for Computational Linguistics.

[Russell et al.1996] Stuart J. Russell, Peter Norvig, John F. Candy, Jitendra M. Malik, and Douglas D. Edwards. 1996. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Schölkopf and Smola2002] Bernhard Schölkopf and Alexander J Smola. 2002. *Learning with kernels*. “The” MIT Press.

[Shawe-Taylor and Cristianini2004] John Shawe-Taylor and Nello Cristianini. 2004. *Kernel methods for pattern analysis*. Cambridge university press.

[Silberer and Lapata2012] Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL ’12*, pages 1423–1433, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Silberer et al.2013] Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In *ACL (1)*, pages 572–582.

[Snow et al.2008] Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’08*, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Socher et al.2011] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics.

[Socher et al.2014] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. In *Transactions of the Association for Computational Linguistics*, pages 207 – 218, April.

[Thomason et al.2014] Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond Mooney. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, Dublin, Ireland, August.

[Torralba et al.2008] Antonio Torralba, Rob Fergus, and William T. Freeman. 2008. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *Pattern Analysis and Machine Intelligence*, 30.

[Tsang1996] Wai-King Tsang. 1996. Comparing the effects of reading and writing on writing performance. *Applied Linguistics*, 17(2):210–233.

- [Turing1950] A. M. Turing. 1950. Computing machinery and intelligence. One of the most influential papers in the history of the cognitive sciences: <http://cogsci.umn.edu/millennium/final.html>.
- [Turner and Charniak2005] Jenine Turner and Eugene Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 290–297, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- [van der Maaten and Hinton2008] L.J.P. van der Maaten and G.E. Hinton. 2008. Visualizing high-dimensional data using t-sne.
- [Wang et al.2010] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, T. Huang, and Yihong Gong. 2010. Locality-constrained linear coding for image classification. In *CVPR*.
- [Widdows2008] Dominic Widdows. 2008. Semantic vector products: Some initial investigations. In *Proceedings of the Second AAI Symposium on Quantum Interaction*.
- [Woodsend and Lapata2010] Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 565–574, Uppsala, Sweden, July. Association for Computational Linguistics.
- [Woodsend and Lapata2011] Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.
- [Woodsend et al.2010] Kristian Woodsend, Yansong Feng, and Mirella Lapata. 2010. Title generation with quasi-synchronous grammar. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10*, pages 513–523, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Xiao et al.2010] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*.
- [Yang et al.2011] Yezhou Yang, Ching Teo, Hal Daume III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 444–454, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.
- [Yao et al.2010] B.Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. 2010. I2t: Image parsing to text description. *Proc. IEEE*, 98(8).
- [Yatskar et al.2013] Mark Yatskar, Svitlana Volkova, Asli elikyilmaz, Bill Dolan, and Luke S. Zettlemoyer. 2013. Learning to relate literal and sentimental descriptions of visual properties. In *HLT-NAACL*, pages 416–425. The Association for Computational Linguistics.
- [Yessenalina and Cardie2011] Ainur Yessenalina and Claire Cardie. 2011. Compositional matrix-space models for sentiment analysis. In *Proceedings of the Conference on Empirical*

*Methods in Natural Language Processing*, pages 172–182. Association for Computational Linguistics.

[Younger1967] Daniel H. Younger. 1967. Recognition and parsing of context-free languages in time  $n^3$ . *Information and Control*, 10(2):189 – 208.

[Yu and Siskind2013] Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 53–63, Sofia, Bulgaria, August. Association for Computational Linguistics.