Pragmatic factors in [automatic] image description

Emiel van Miltenburg
SIKS Dissertation Series No. 2019-25

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Typeset using \LaTeX, using the TeXGyre fonts. The small figure on the previous page comes from the phaistos package, which is based on the Greek Phaistos disk.

Cover photos by Willem van de Poll (1895 - 1970), licensed CC0, by het Nationaal Archief.

Printed by ProefschriftMaken || www.proefschriftmaken.nl

ISBN: 9789463804899

© 2018 Emiel van Miltenburg
Pragmatic factors in (automatic) image description

ACADEMISCH PROEFSCHRIFT

Ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Geesteswetenschappen
op maandag 14 oktober 2019 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door
Cornelis Wilhelmus Johannes van Miltenburg
geboren te Nieuwegein
promotor: prof.dr. P.T.J.M. Vossen
copromotor: dr. D. Elliott
## Contents

1 Introduction

1.1 Describing an image ............................................................. 1
1.2 Automatic image description .................................................. 1
1.3 Defining image descriptions ................................................... 2
1.4 Image description data ......................................................... 2
1.5 A model of the image description process ................................. 3
1.6 Image description systems and the semantic gap ....................... 4
   1.6.1 The semantic gap .......................................................... 6
   1.6.2 The pragmatic gap ......................................................... 7
1.7 Research questions ............................................................ 8
   1.7.1 Characterizing human image descriptions .......................... 9
   1.7.2 Characterizing automatic image descriptions .................... 9
1.8 Methodology ......................................................................... 10
   1.8.1 Corpus analysis ............................................................ 10
   1.8.2 Computational modeling ............................................... 11
1.9 Contributions of this thesis .................................................. 11

2 How people describe images

2.1 Introduction ........................................................................... 17
   2.1.1 Contents of this chapter ................................................... 17
   2.1.2 Publications ................................................................. 18
2.2 Levels of interpretation ........................................................ 19
   2.2.1 The Of/About distinction ................................................. 19
   2.2.2 Barthes’ Denotation and Connotation ............................... 20
   2.2.3 Understanding the semantic gap ..................................... 20
2.3 Pragmatic factors in image description .................................... 21
2.4 Image description datasets .................................................... 24
2.5 Image description as perspective-taking ................................... 25
2.6 Variation ............................................................................... 26
   2.6.1 Clustering entity labels ................................................... 26
## Contents

6.4.3 Different kinds of RNNs ........................................ 102  
6.4.4 Encoding and decoding sentences .............................. 103  
6.4.5 Attention mechanisms ........................................... 104  
6.5 Generative Adversarial Networks .................................. 105  
6.6 Takeaway ................................................................ 106  
6.7 Evaluation .................................................................. 106  
6.7.1 Evaluation of automatic image descriptions ................. 106  
6.8 Error analysis ............................................................. 108  
6.8.1 Coarse-grained analysis .......................................... 108  
6.8.2 Fine-grained analysis .............................................. 108  
6.9 Error categories .......................................................... 109  
6.10 Annotation tasks ........................................................ 110  
6.10.1 Results for the coarse-grained task ............................ 110  
6.10.2 Evaluating the fine-grained annotations ...................... 111  
6.11 Correcting the errors ................................................ 112  
6.12 Takeaway ................................................................ 112  
6.13 Conclusion ................................................................ 114  
6.13.1 Implications for image description research ............... 114  
6.13.2 Next chapter ........................................................... 114

7 Measuring diversity .......................................................... 117  
7.1 Introduction ................................................................. 117  
7.1.1 Contents of this chapter .......................................... 117  
7.1.2 Publications .......................................................... 118  
7.2 Background ................................................................ 118  
7.3 Existing metrics ............................................................ 119  
7.3.1 Systems ................................................................. 120  
7.3.2 Results ................................................................. 120  
7.4 Image description as word recall .................................... 122  
7.4.1 Global recall ......................................................... 123  
7.4.2 Local recall ........................................................... 123  
7.4.3 Global ranking of omitted words ............................... 125  
7.4.4 Local ranking of omitted words ............................... 126  
7.5 Compound nouns and prepositional phrases ..................... 127  
7.6 Discussion and Future Research .................................... 129  
7.6.1 Other metrics ......................................................... 129  
7.6.2 Limitations and human validation ............................. 130  
7.7 Conclusion ................................................................. 130

8 Final conclusion ................................................................ 133  
8.1 What have we learned? ................................................ 133  
8.1.1 Image description from a human perspective .............. 133  
8.1.2 Image description from a machine perspective .......... 136  
8.1.3 How human-like should automatic image descriptions be? 137  
8.2 Application: supporting blind and visually impaired people 138  
8.2.1 Developing sign-language gloves: A cautionary tale ......................... 138  
8.2.2 Existing research on supporting the blind ........................ 139  
8.2.3 Future research supporting blind and visually impaired people ........ 141
Acknowledgments

Although only my name is on the cover of this dissertation, I could not have completed this work without the people around me. First and foremost I would like to thank my supervisors, Piek Vossen and Desmond Elliott, for all their encouragement and support. Working with Piek has taught me the meaning of visionary research, imagining long-term goals and working towards those goals despite the inevitable obstacles. There was certainly no shortage of ideas in our meetings! Having too many ideas tends to put you at the risk of drowning, but Piek always made sure I stayed afloat. Desmond has been a patient guide in the world of Vision and Language. Without him, this thesis would have looked completely different. I could not have wished for better supervisors.

Many thanks also go to the reading committee, for taking the time to read and comment on my thesis. As I am writing this, I am looking forward to the defense! At the event, I am honored to have Hennie van der Vliet and Roxane Segers as my paranymphs. Many thanks in advance.

I am also grateful to all of my co-authors. Next to Desmond and Piek, these are (in alphabetical order): Lora Aroyo, Ákos Kádar, Ruud Koolen, Emiel Krahmer, Alessandro Lopopolo, Roser Morante, Chantal van Son, and Benjamin Timmermans. Collaborating with these people has made me a better writer and researcher. If you spot a particularly good piece of writing in this thesis, there’s a good chance it’s theirs.

I have greatly benefited from working in a very pleasant environment, both in Amsterdam and in Edinburgh. The CLTL (Computational Lexicology and Terminology Lab) has felt like a second home for more than four years. Who says you cannot have two captains on one ship? Even if the waves in education became really big, Captain Hennie was always able to steer us into calmer waters. And Captain Piek made sure there was never any cause for mutiny, with regular events to keep the spirits up. Many thanks to all of the crew for all discussions, banter, and collaboration. My stay at the University of Edinburgh has been much shorter. Ten weeks is really too short a time to spend in such a nice city. Thanks to everyone at EdinburghNLP for making me feel welcome.

After almost five years working on my PhD research in Amsterdam, it was hard to imagine life after the PhD. As it turns out: it exists! There is no black hole, but a wonderful green campus in Tilburg. Thanks to all my new colleagues in Communication and Information Science for welcoming me to the department.

Finally, I would like to thank my friends and family for being there throughout my PhD. And a very big ‘thank you’ to Loes, for the past, the present, and the future.

Utrecht, Summer 2019
Understanding of language by machines

The research for this thesis was carried out within the context of a larger project, called *Understanding of Language by Machines* (ULM). This project is funded through the NWO Spinoza prize, awarded in 2013 to Piek Vossen. The goal of the project is:

“…to develop computer models that can assign deeper meaning to language that approximates human understanding and to use these models to automatically read and understand text. Current approaches to natural language understanding consider language as a closed-world of relations between words. Words and text are however highly ambiguous and vague. People do not notice this ambiguity when using language within their social communicative context. This project tries to get a better understanding of the scope and complexity of this ambiguity and how to model the social communicative contexts to help resolving it.”

(Source: http://www.understandinglanguagebymachines.org/)

The project is led by Piek Vossen, with the help of Selene Kolman (project manager) and Paul Huygen (scientific programmer). Other members are or have been: Tommaso Caselli, Filip Ilievski, Rubén Izquierdo, Minh Lê, Alessandro Lopopolo, Roser Morante, Marten Postma, and Chantal van Son.
Notes

Language in this thesis

Research is almost impossible to carry out alone. Hence, all the content chapters from this thesis are based on collaborative work. Since this thesis is presented as a single-authored monograph, I have made the following choice. The introduction and conclusion are written from a first-person singular perspective (using *I*), but, in acknowledgment of my co-authors, all content chapters are written from a first-person plural perspective (using *we*). I remain solely responsible for any errors in this thesis.

Images and Copyright

Most of the images in this thesis originate from Flickr.com, a social image sharing platform, where amateurs and professional photographers share their work under various licenses. Many of these images are provided under a Creative Commons licence. Where possible, I have tried to use images provided either under such a license, or even images that are part of the Public Domain, with the appropriate attributions. Unfortunately, this was not always possible.

The research presented in this thesis focuses on image descriptions from the Flickr30K and MS COCO datasets, and some of the images from those corpora are fully copyrighted. Furthermore, some images have been deleted from Flickr.com after their publication in either Flickr30K or MS COCO. In those cases, it was not always possible to find and credit the original author (although I did try, using Google’s reverse image search). I have generally tried to avoid using these images, and to look for alternative examples. In some cases, however, I have found that the copyrighted image provided the clearest example.

The use of copyrighted images is somewhat of a legal gray area. Copyright law in the US (where Flickr is based) has a Fair Use exception, that allows for the use of copyrighted images in some cases. Those cases are judged using the following four factors:

1. **The purpose and character of the use.** Here, we could reasonably argue that scholarly work qualifies as ‘transformative use’, where we do not just copy the image, but reflect on the meaning of the image and the associated descriptions from existing image description corpora.
2. **The nature of the copyrighted work.** Here, we could argue that the images were published on Flickr.com already (meant to be seen by others), and used in existing image description datasets.
3. **The Amount and Substantiality of the Portion Taken.** Here, we need to concede that we are not just copying a portion of the image. However, this is unavoidable in discussing image descriptions, which aim to capture the heart of the work.
4. **The effect of the use on the potential market.** This is hard to judge, as the use is not for commercial gain, but purely for research purposes.

The use of copyrighted images is somewhat of a legal gray area. Copyright law in the US (where Flickr is based) has a Fair Use exception, that allows for the use of copyrighted images in some cases. Those cases are judged using the following four factors:

1. **The purpose and character of the use.** Here, we could reasonably argue that scholarly work qualifies as ‘transformative use’, where we do not just copy the image, but reflect on the meaning of the image and the associated descriptions from existing image description corpora.
2. **The nature of the copyrighted work.** Here, we could argue that the images were published on Flickr.com already (meant to be seen by others), and used in existing image description datasets.
3. **The Amount and Substantiality of the Portion Taken.** Here, we need to concede that we are not just copying a portion of the image. However, this is unavoidable in discussing image descriptions, which aim to capture the heart of the work.

---

1. See https://creativecommons.org.
2. See https://fairuse.stanford.edu/overview/public-domain/welcome/
3. See https://fairuse.stanford.edu/overview/fair-use/four-factors/
The Effect of the Use Upon the Potential Market. We do not wish to use the images for any commercial benefit, and do not foresee any effect on the potential market for the images discussed in this thesis.

Dutch law does not have a Fair Use exception. Rather, it provides for a ‘Right to Quote’, which arguably covers our use of the copyrighted images from Flickr. After all: one cannot have a scholarly discussion of the image descriptions from MS COCO or Flickr30K without taking the images into account. Having said this, it seems to me that the current situation is not ideal. I hope that we, as a scientific community, can move toward datasets that are not limited by copyright. If this turns out to be impossible, we should at least require all new datasets to provide a list of authors to be acknowledged when citing relevant parts of that dataset.

For my part, I invite authors of any images that have gone uncredited to contact me, so that I can give credit where credit is due.

---