Polish Corpus of Annotated Descriptions of Images

Alina Wróblewska
Institute of Computer Science, Polish Academy of Sciences
Warsaw, Poland
alina@ipipan.waw.pl

Abstract

The paper presents a new dataset of image descriptions in Polish. The descriptions are morphosyntactically analysed and the pairs of these descriptions are annotated in terms of semantic relatedness and entailment. All annotations are provided by human annotators with strong linguistic background. The dataset can be used for evaluation of various systems integrating language and vision. It is applicable for evaluation of systems designed to image generation based on provided descriptions (text-to-image generation) or to caption generation based on images (image-to-text generation). Furthermore, as selected images are split into thematic groups, the dataset is also useful for validating image classification approaches.

Keywords: language and vision, evaluation dataset, annotated image descriptions, Polish

1. Introduction

Language and vision are two essential modalities that enable interpersonal communication. As broadly understood communication is an important area of artificial intelligence, AI researchers attach importance to image processing, natural language processing, and integration of these two fields. Vision-to-language approaches consist mostly in mapping images to sentences, e.g. Hodosh et al. (2013), or generating image descriptions, e.g. Xu et al. (2015), Karpathy and Fei-Fei (2017). For example, Karpathy and Fei-Fei (2017) propose a model that generates natural language descriptions of whole images or their regions. Based on datasets of images and their sentence descriptions, the model learns about the inter-modal correspondences between language and vision. Regarding language-to-vision approaches, research and scientific experiments are conducted in the areas of text-based image retrieval, e.g. Rasiwasia et al. (2010), or text-based image generation, e.g. Denton et al. (2015). For example, the paper by Denton et al. (2015) describes a "generative parametric model capable of producing high quality samples of natural images". In order to evaluate many of these approaches, the high-quality caption-image datasets are necessary.

This paper presents a new Polish dataset of annotated image descriptions – AIDe[1] (**A**nnotated **I**mage **De**scriptions). The dataset consists of 2K natural language descriptions of 1K images. The dataset is probably too small for training a sophisticated language–vision system. For training purposes, the dataset should be expanded to the greatest possible extent. However, in the age of intensive research on multilingual NLP, e.g. Faruqui and Dyer (2014), it seems to make sense to build even small but high-quality evaluation resources.

The presented dataset can be used for evaluation of various systems integrating language and vision. It is applicable for evaluation of systems designed to generation of images based on provided descriptions (text-to-image generation) or to generation of captions based on images (image-to-text generation). Furthermore, as elected images are split into thematic groups, the dataset is also useful for validating image classification approaches.

The procedure of selecting, describing, and splitting images into thematic groups is described in Section 2. The image descriptions are morphosyntactically annotated (see Section 3.) and the pairs of these descriptions are annotated in terms of semantic relatedness and entailment (see Section 4.). All annotations are provided by human annotators with strong linguistic background.

2. Dataset

2.1. Image Classification

The first step of building the dataset consists in selecting 1K images from the Flickr8k dataset (Hodosh et al., 2013). The selected images are arbitrarily split into 46 thematic groups.[2]

As the classes do not link to any of standard ontologies and each image is classified into only one group, even if some of them could be classified into multiple classes, we decided to reclassify the images according to the heuristics based on WordNet hyperonym hierarchy[3] (Fellbaum, 1998).

[1] http://zil.ipipan.waw.pl/Scwad/AIDe

[2] The images are assigned to the following thematic groups: **people** (kids, different_people), **animals** (dogs, birds, different_animals), and sport and leisure activities: **water activities** (fishing, swimming, surfing, kayaking, boating_or_sailing), **winter activities** (skiing, snowboarding, sledding, ice-skating), **driving and riding** (driving, motorbike_riding, quad-bike_riding, biking, non-motor_vehicle_riding, horse_riding, (inline)roller-skating, skateboarding), **playing** (jumping, jumping_to_water, jumping_on_trampoline, swinging, sliding_down, dancing), **team games** (basketball, football, volleyball, baseball_or_rugby, hockey), **individual activities** (individual_sports, martial_arts, climbing, mountain_hiking, running_or_jogging), **unclassified activities** (flying, photographing, telephoning, kissing, musical_instruments, eating, resting, sunbathing).

The numbers of images within individual thematic groups vary from 6 images in the volleyball and telephoning groups to 94 images in the different_people group. The second largest groups are children and dogs with 50 images each.

[3] We use English categories from WordNet in our dataset, however these classes could be straightforwardly map onto plWordNet (Rudnicka et al., 2012).

We distinguish 3 class types: events (**Event**), entities (**Entity**), i.e. participants and artefacts, and general location (**Out-In**), i.e. outside vs. inside (the building). Each image has to be assigned to at least one class within each of three class groups:

- **Event** classes (24):[4]

sing (6)	swing (21)	game (47)
smoke (6)	lie (22)	run (68)
kiss (8)	climb (26)	walk (82)
fish (16)	fly (28)	play (95)
consume (18)	swim (33)	sit (104)
carry (19)	watch (34)	ride (125)
photograph (19)	go (40)	jump (127)
dance (20)	skate (45)	stand (186)

- **Entity** classes (6):[5]

food (23)	vehicle (166)
instrument (49)	animal (172)
artifact (141)	person (876)

- **Out**side-**In**side classes (2):[6]

inside (211) outside (845)

[4]The number in brackets corresponds to the number of images classified into the particular class. The game class is further divided into 8 subclasses (**Event-hypon**ym):

cricket (2)	ice hockey (5)	soccer (9)
rugby (2)	football (7)	basketball (10)
field hockey (4)	baseball (8)	

[5]Some of the classes within **Entity** group are further divided into the following subclasses (**Entity-hypon**ym):

fish (2)	roller skate (18)
insect (2)	sledge (18)
wheelchair (2)	snowboard (18)
aquatic mammal (4)	bicycle (20)
racket (4)	horse (20)
equipment (5)	motorcycle (20)
reptile (5)	skateboard (20)
weapon (5)	ski (20)
scooter (6)	slide (20)
cat (7)	surfboard (20)
percussion instrument (7)	boat (21)
aircraft (8)	hoofed mammal (22)
ball (8)	car (23)
mammal (8)	guitar (25)
bike (9)	bird (41)
string (9)	dog (71)
trampoline (9)	athlete (82)
kayak (10)	woman (283)
skate (10)	child (323)
wheeled vehicle (14)	man (441)
wind (15)	

[6]**Inside-Outside** classes are sometimes specified with the following subclasses:

stairs (9)	court (19)	street (53)
ring (10)	park (26)	track (101)
apartment (12)	stadium (26)	water (104)
ice rink (16)	pool (28)	field (118)
shore (16)	beach (36)	
playground (17)	mountain (49)	

The classes of **Event** type correspond mostly to the sentence predicates. The generalised **Entities** correlate, in turn, with predicate arguments. Finally, the classes of **Out-In** group can be equated with the location adverbials.

Statistics One image can be classified into more than one class within a thematic group (see Table 1). Especially in the group of Entities, the images are assigned to more than one class (see 2-fold classification).

	1-fold	2-fold	3-fold
Event	808	189	3
Entity	582	409	9
Outside_**In**side	944	56	0

Table 1: Classification statistics: 1-fold, 2-fold, and 3-fold classifications correspond to processes of assigning an image to one, two, or three classes (within one class type), respectively.

Figure 1: An example image from https://www.flickr.com/photos/floridatania/1057089366.

Example The example image (see Figure 1) is classified as follows:

- Event: jump
- Entity: person (subclass of the person class: child)
- Location: outside (subclass: pool).

2.2. Image Descriptions

The chosen images are presented to two authors who, independently of each other, formulate their descriptions based on a short instruction. The authors are instructed to write one single sentence (with a sentence predicate) describing the entities, events and scenes depicted in a displayed image. They should not describe an imaginable context or an interpretation of what may lie behind the scene in the picture. If some details in the picture are not obvious, they should not be described either. Finally, the descriptions should contain Polish diacritics and proper punctuation.

The final set of image descriptions consists of 2K sentences, i.e. two sentences for each image. The descriptions of the same image are not doublets.

Statistics The authors write similarly long sentences. The average length of sentences written by the A author is 12.42 tokens. The B author's description length is 12.49 tokens per sentence on average.

In order to estimate textual similarity between sentences in each pair we apply two measures: Monge-Elkan similarity measure (Monge and Elkan, 1996) and tokens sort ratio based on approximate string matching (Wagner and Fisher, 1974). Monge-Elkan distance is a hybrid measure for computing similarity between two strings of multiple tokens using an internal measure (e.g. Jaro-Winkler or Levenshtein) to estimate similarity between individual tokens. Measured with Monge-Elkan distance, textual similarity between description pairs A and B is 0.811 (average)[7] and between pairs B and A is 0.809 (average)[8].

Approximate string matching token sort splits two strings into tokens, sorts the tokens, and estimates the similarity of the sorted strings. Textual similarity between desription pairs is 0.595 (average)[9] measured with approximate string matching token sort ratio.

Example The image in Figure 1 is described as follows in our dataset:

A. *Chłopiec skacze do basenu z wysokiej trampoliny.*
(Eng. 'A boy is jumping into the pool off a high diving board.')

B. *Chłopiec w niebieskim ubraniu skacze do basenu z trampoliny.*
(Eng. 'A boy in blue clothes is jumping into the pool off a diving board.')

3. Morphosyntactic Annotations

The similar datasets, e.g. for English, typically consist of images combined with multiple captions, e.g. Rashtchian et al. (2010),[10] Hodosh et al. (2013),[11] Lin et al. (2014).[12]

The morphosyntactic annotations of English captions are usually not provided, because English is a resource-rich language and there are plenty of high-quality English NLP tools. Polish, in turn, is a resource-poor language which still suffers from the lack of high-quality NLP tools. In order to avoid propagation of tagging or parsing errors[13] to the evaluation of language-vision systems, we provide morphosyntactic annotations of our Polish image descriptions. Furthermore, with annotated data, it is possible to verify the impact of tagging and/or parsing on the overall quality of language-vision systems.

3.1. Annotation Procedure

Each description is tokenised and morphologically analysed with Morfeusz (Woliński, 2014) and tagged and lemmatised with Concraft (Waszczuk, 2012). The sentences are then parsed with MaltParser (Nivre, 2009) and Mate parser (Bohnet, 2010) trained[14] on Polish Dependency Bank (Wróblewska, 2014).

Two dependency trees and accompanying morphosyntactic annotations (lemmas, part-of-speech tags, morphological features) are manually verified and possibly corrected by two linguists. Finally, two verified dependency trees are unified by the third linguist who is the most experienced in Polish linguistics.

3.2. Dependency Tree Formats

The dependency trees and accompanying morphosyntactic annotations are stored in column-based CoNLL format (Nivre et al., 2007). As the format of Universal Dependencies (UD)[15] becomes more and more common, the dependency trees are also automatically converted into corresponding UD trees.

token	lemma	POS	morph
Chłopiec	CHŁOPIEC	subst	sg|nom|m1
skacze	SKAKAĆ	fin	sg|ter|imperf
do	DO	prep	gen
basenu	BASEN	subst	sg|gen|m3
z	Z	prep	gen|nwok
wysokiej	WYSOKI	adj	sg|gen|f|pos
trampoliny	TRAMPOLINA	subst	sg|gen|f

Figure 2: The morphosyntactic analysis of *Chłopiec skacze do basenu z wysokiej trampoliny.* (Eng. 'A boy is jumping into the pool off a high diving board.').

Example Taking the A caption as an example (see Section 2.2.), its morphosyntactic analysis with part-of-speech tags and morphological features, and its dependency tree are in Figure 2[16] and Figure 3, respectively. The UD-formatted example is in Figures 4 and 5.

[7] Standard deviation of the average similarity scores estimated with Monge-Elkan measure: 0.086.

[8] Standard deviation of the average similarity scores estimated with Monge-Elkan measure: 0.085.

[9] Standard deviation of the average similarity scores estimated with approximate string matching: 0.114.

[10] http://vision.cs.uiuc.edu/pascal-sentences

[11] http://nlp.cs.illinois.edu/HockenmaierGroup/8k-pictures.html

[12] http://cocodataset.org

[13] Some statistics about the quality of Polish NLP tools are collected on http://clip.ipipan.waw.pl/benchmarks.

[14] http://zil.ipipan.waw.pl/PDB/PDBparser

[15] http://universaldependencies.org

[16] Explanation of grammatical classes and categories:
subst – substantive

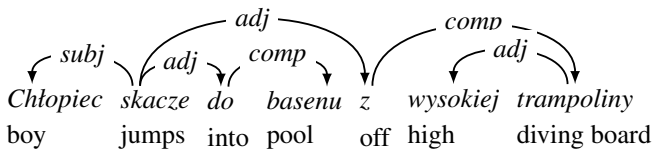

Figure 3: The dependency tree of *Chłopiec skacze do basenu z wysokiej trampoliny.* (Eng. 'A boy is jumping into the pool off a high diving board.').

token	UD-POS	UD-feature
Chłopiec	NOUN	Animacy=Hum,Case=Nom, Gender=Masc,Number=Sing
skacze	VERB	Aspect=Imp,Mood=Ind, Number=Sing,Person=3, Tense=Pres,VerbForm=Fin
do	ADP	AdpType=Prep,Case=Gen
basenu	NOUN	Animacy=Inan,Case=Gen, Gender=Masc,Number=Sing
z	ADP	AdpType=Voc,Case=Gen, Variant=Short
wysokiej	ADJ	Case=Gen,Degree=Pos, Gender=Fem,Number=Sing
trampoliny	NOUN	Case=Gen,Gender=Fem, Number=Sing

Figure 4: The UD-formatted morphosyntactic analysis of *Chłopiec skacze do basenu z wysokiej trampoliny.* (Eng. 'A boy is jumping into the pool off a high diving board.').

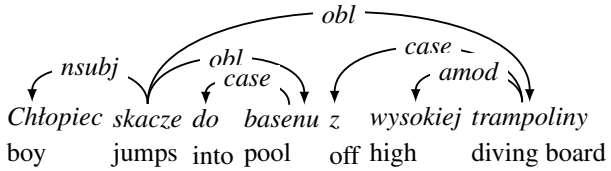

Figure 5: The UD-formatted dependency tree of *Chłopiec skacze do basenu z wysokiej trampoliny.* (Eng. 'A boy is jumping into the pool off a high diving board.').

prep – preposition
fin – finite verb
adj – adjective
sg – singular number
nom – nominative case
gen – genitive case
m1 – human masculine (virile) gender
m3 – inanimate masculine gender
f – feminine gender
ter – third person
imperf – imperfect aspect
nwok – non-vocalic
pos – positive degree (of adjectives)

4. Semantic Annotations

In order to test a system of image generation based on two descriptions, information whether these two descriptions are semantically related, or whether the meaning of one description entails the meaning of the other one, seems to be relevant. Therefore, each sentence pair is human-annotated for relatedness in meaning and entailment. The semantic annotations are derived from Polish CDSCorpus[17] (Wróblewska and Krasnowska-Kieraś, 2017).

4.1. Semantic Relatedness

The relatedness score corresponds to the degree of semantic relatedness between two sentences and is calculated as the average of six human ratings collected for this sentence pair on the 6-point Likert scale (0 to 5). This score indicates the extent to which the meanings of two sentences are related. The score 5 indicates very related descriptions and the score 0, in turn, indicates unrelated descriptions. The scores 1–4 denote that the pair consists of more or less related descriptions. The degree of semantic relatedness is not equivalent to the degree of semantic similarity. Semantic similarity is only a special case of semantic relatedness, semantic relatedness is thus a more general term than the other one.

Statistics Table 2 aggregates the occurrences of 6 possible relatedness scores, calculated as the mean of all 6 individual annotations, rounded to an integer.

relatedness	# of pairs
0	0
1	11
2	99
3	418
4	440
5	32

Table 2: Relatedness scores rounded to integers (total: 1K pairs).

Example The captions A and B (see Example in Section 2.2.) are annotated as quite related (scored 4) in our dataset.

4.2. Entailment Relations

The entailment relation between two descriptions of the same image is labelled with *entailment* or *neutral*.[18] The description pairs are annotated for entailment in both directions (i.e. bidirectional entailment annotations), because an entailment relation between two sentences must not be symmetric. The final entailment label is actually a pair of two labels:

[17] http://zil.ipipan.waw.pl/Scwad/CDSCorpus

[18] There is also the label *contradiction*, but it is not present in the dataset. This is in line with our assumption that two sentences describing one image should not be contradictory.

- *entailment+neutral* points to one-way entailment,[19]
- *entailment+entailment* points to equivalence (two-way entailment),
- *neutral+neutral* points to no entailment.

The label assigned by the majority of 3 human annotators is selected as the valid entailment label.

Statistics Table 3 shows the number of the particular entailment labels in the corpus.

entailment	# of pairs
entailment+entailment	40
entailment+neutral	293
neutral+neutral	667

Table 3: Entailment labels (total: 1K pairs).

Example The descriptions A and B (see Example in Section 2.2.) are labelled *neutral* in both entailment directions.

5. Multilingual Variant of the Dataset

The decision to choose images from Flickr8k was well thought out. Based on Flickr8k dataset and AIDe dataset it is possible to compile a new Polish-English multi-parallel corpus, which could be used for validating e.g. machine translation systems.
In our dataset, we apply original image IDs of Flickr8k. In order to build the multi-parallel corpus, for each image ID, e.g. 1057089366_ca83da0877.jpg,[20] the following sentences should be selected:

- two Polish sentences from AIDe (see A-PL and B-PL in the list in Example),
- five English captions from Flickr8k (see 0-EN to 4-EN in Example)[21].

As Polish is a resource-poor language, each new resource is valuable, even if it is a by-product like in this case.

Example An excerpt of a possible Polish-English multi-parallel corpus:

A-PL Chłopiec skacze do basenu z wysokiej trampoliny.

B-PL Chłopiec w niebieskim ubraniu skacze do basenu z trampoliny.

0-EN A boy descends off the end of a high diving board.

1-EN A child jumps off a high diving board into the pool.

2-EN A kid jumps off the diving board and into the swimming pool below.

3-EN A little kid is jumping off a high dive at the pool.

4-EN The boy is jumping off a high diving board into the pool.

6. Conclusions

The aim of this paper was to present AIDe – Polish dataset of annotated image descriptions. The descriptions were morphosyntactically annotated, i.e. tokens were assigned part-of-speech tags and morphological features, and sentences were represented as dependency trees (also UD-formatted trees). Apart from morphosyntactic annotations, we also provided semantic annotations of description pairs. Image description pairs were annotated with semantic relatedness scores and bidirectional entailment labels. All annotations were provided by human annotators with strong linguistic background.
We decided to augment raw image descriptions with morphosyntactic annotations, in order to provide a dataset which is designed to evaluate the correspondence between language and vision undisturbed by errors at the lower language processing stages. Furthermore, pre-annotated data enable verification of the impact of tagging and/or parsing on the overall quality of language-vision systems.
We are not aware of availability of other image-caption datasets which are annotated with semantic relatedness scores and entailment labels. Hence, our dataset enables to check whether additional semantic information is useful e.g. in image generation based on multiple sentences.
The dataset is small and without increasing its size it is insufficient for training purposes. This high-quality resource is rather intended for evaluation of systems integrating language and vision (text-to-image or image-to-text generation), for image classification, as selected images are split into thematic groups based on WordNet, or even for evaluation of machine translation systems, if corresponding captions are extracted from Flickr8k dataset.

7. Acknowledgements

The research presented in this paper was founded by SONATA 8 grant no 2014/15/D/HS2/03486 from the National Science Centre Poland and by the Polish Ministry of Science and Higher Education as part of the investment in the CLARIN-PL research infrastructure.
The author would like to thank the anonymous reviewers for their careful reading of the paper and their valuable comments and suggestions.

[19] While the actual corpus labels are ordered in the sense that there is a difference between e.g. *entailment+neutral* and *neutral+entailment* (the entailment occurs in different directions), we treat all labels as unordered for the purpose of this summary (e.g. *entailment+neutral* covers *neutral+entailment* as well, representing the same type of relation between two sentences).

[20] This is the ID of the image in Figure 1.

[21] The original IDs of English captions are:
1057089366_ca83da0877.jpg#0
1057089366_ca83da0877.jpg#1
1057089366_ca83da0877.jpg#2
1057089366_ca83da0877.jpg#3
1057089366_ca83da0877.jpg#4

8. Bibliographical References

Bohnet, B. (2010). Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING 2010, pages 89–97.

Denton, E. L., Chintala, S., Szlam, A., and Fergus, R. (2015). Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In C. Cortes, et al., editors, *Advances in Neural Information Processing Systems 28*, pages 1486–1494. Curran Associates, Inc.

Faruqui, M. and Dyer, C. (2014). Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing Image Description As a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*, 47(1):853–899.

Karpathy, A. and Fei-Fei, L. (2017). Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In David Fleet, et al., editors, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer International Publishing.

Monge, A. E. and Elkan, C. P. (1996). The field matching problem: Algorithms and applications. In *Proceedings of The Second International Conference on Knowledge Discovery and Data Mining*, pages 267–270.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Nivre, J. (2009). Non-Projective Dependency Parsing in Expected Linear Time. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 351–359.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting Image Annotations Using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., and Vasconcelos, N. (2010). A New Approach to Cross-modal Multimedia Retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 251–260.

Rudnicka, E., Maziarz, M., Piasecki, M., and Szpakowicz, S. (2012). A strategy of mapping Polish Wordnet onto Princeton Wordnet. In *Proceedings of COLING 2012: Posters*, pages 1039–1048, Mumbai, India, December. The COLING 2012 Organizing Committee.

Wagner, R. A. and Fisher, M. J. (1974). The String-to-String Correction Problem. *Journal of the Association for Computing Machinery*, 21(1):168–173.

Waszczuk, J. (2012). Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of COLING 2012: Technical Papers*, pages 2789–2804.

Woliński, M. (2014). Morfeusz reloaded. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111. ELRA.

Wróblewska, A. and Krasnowska-Kieraś, K. (2017). Polish Evaluation Dataset for Compositional Distributional Semantics Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 784–792, Vancouver, Canada. Association for Computational Linguistics.

Wróblewska, A. (2014). *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In David Blei et al., editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2048–2057.