Generalizing Image Captions for Image-Text Parallel Corpus

Polina Kuznetsova, Vicente Ordonez, Alexander Berg,
Tamara Berg and Yejin Choi
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
{pkuznetsova,vordonezroma,aberg,tlberg,ychoi}@cs.stonybrook.edu

Abstract

The ever growing amount of web images and their associated texts offers new opportunities for integrative models bridging natural language processing and computer vision. However, the potential benefits of such data are yet to be fully realized due to the complexity and noise in the alignment between image content and text. We address this challenge with contributions in two folds: first, we introduce the new task of *image caption generalization*, formulated as visually-guided sentence compression, and present an efficient algorithm based on dynamic beam search with dependency-based constraints. Second, we release a new large-scale corpus with 1 million image-caption pairs achieving tighter content alignment between images and text. Evaluation results show the intrinsic quality of the generalized captions and the extrinsic utility of the new image-text parallel corpus with respect to a concrete application of image caption transfer.

1 Introduction

The vast number of online images with accompanying text raises hope for drawing synergistic connections between human language technologies and computer vision. However, subtleties and complexity in the relationship between image content and text make exploiting paired visual-textual data an open and interesting problem.

Some recent work has approached the problem of composing natural language descriptions for images by using computer vision to retrieve images with similar content and then transferring text from the retrieved samples to the query image (e.g. Farhadi et al. (2010), Ordonez et al. (2011), Kuznetsova et al. (2012)). Other work (e.g. Feng and Lapata (2010a), Feng and Lapata (2010b)) uses computer vision to bias summarization of text associated with images to produce descriptions. All of these approaches rely on existing text that describes visual content, but many times existing image descriptions contain significant amounts of extraneous, non-visual, or otherwise non-desirable content. The goal of this paper is to develop techniques to automatically clean up visually descriptive text to make it more directly usable for applications exploiting the connection between images and language.

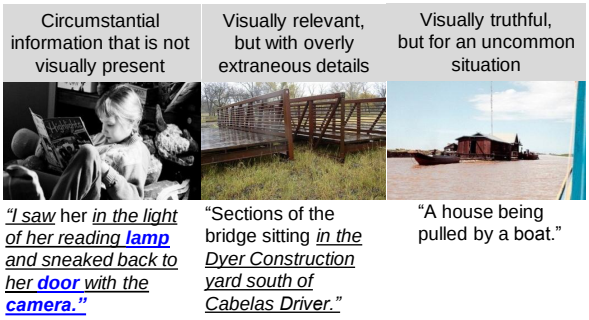

Figure 1: Examples of captions that are not readily applicable to other visually similar images.

As a concrete example, consider the first image in Figure 1. This caption was written by the photo owner and therefore contains information related to the context of when and where the photo was taken. Objects such as "lamp", "door", "camera" are not visually present in the photo. The second image shows a similar but somewhat different issue. Its caption describes visible objects such as *"bridge"* and *"yard"*, but "Cabelas Driver" are overly specific and not visually detectable. The

Dependency Constraints with Examples			Additional Dependency Constraints
Constraints	Sentence	Dependency	
advcl*(←)	Taken when it was running...	taken←running	acomp*(↔), advmod(←), agent*(←), attr(↔)
amod(←)	A wooden chair in the living room	chair← wooden	auxpass(↔), cc*(↔),complm(←), cop*(↔)
aux(↔)	This crazy dog was jumping...	jumping↔was	csubj*/csubjpass*(↔),expl(↔), mark*(↔)
ccomp*(→)	I believe a bear was in the box...	believe→was	infmod*(↔), mwe(↔), nsubj*/nsubjpass*(↔)
prep(←)	A view from the balcony	view←from	npadvmod(←), nn(←), conj*(↔), num*(←)
det(↔)	A cozy street cafe...	cafe↔A	number(↔), parataxis(←), ↔
dobj*(↔)	A curious cow surveys the road...	surveys↔road	partmod*(←), pcomp*(↔), purpcl*(←)
iobj*(↔)	...rock gives the water the color	gives↔water	possessive(↔), preconj*(←), predet*(←)
neg(↔)	Not a cloud in the sky...	cloud↔Not	prt(↔), quantmod(←), rcmod(←), ref(←)
pobj*(↔)	This branch was on the ground...	on↔ground	rel*(↔), tmod*(←), xcomp*(→), xsubj(→)

Table 1: Dependency-based Constraints

text of the third image, *"A house being pulled by a boat"*, pertains directly to the visual content of the image, but is unlikely to be useful for tasks such as caption transfer because the depiction is unusual.[1] This phenomenon of information gap between the visual content of the images and their corresponding narratives has been studied closely by Dodge et al. (2012).

The content misalignment between images and text limits the extent to which visual detectors can learn meaningful mappings between images and text. To tackle this challenge, we introduce the new task of *image caption generalization* that rewrites captions to be more visually relevant and more readily applicable to other visually similar images. Our end goal is to convert noisy image-text pairs in the wild (Ordonez et al., 2011) into pairs with tighter content alignment, resulting in new simplified captions over 1 million images. Evaluation results show both the intrinsic quality of the generalized captions and the extrinsic utility of the new image-text parallel corpus. The new parallel corpus will be made publicly available.[2]

2 Sentence Generalization as Constraint Optimization

Casting the generalization task as *visually-guided* sentence compression with lightweight revisions, we formulate a constraint optimization problem that aims to maximize content selection and local linguistic fluency while satisfying 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 remaining semantically compatible with respect to a given image-text pair (e.g., preserving predicate-argument relations). More formally, we maximize the following objective function:

$$F(y;x) = \Phi(y;x,v) + \Psi(y;x)$$

$$\text{subject to } \mathcal{C}(y;x,v)$$

where $x = \{x_i\}$ is the input caption (a sentence), v is the accompanying image, $y = \{y_i\}$ is the output sentence, $\Phi(y;x,v)$ is the content selection score, $\Psi(y;x)$ is the linguistic fluency score, and $\mathcal{C}(y;x,v)$ is the set of hard constraints. Let $l(y_i)$ be the index of the word in x that is selected as the i'th word in the output y so that $x_{l(y_i)} = y_i$. Then, we factorize $\Phi(\cdot)$ and $\Psi(\cdot)$ as:

$$\begin{aligned}\Phi(y;x,v) &= \sum_i \phi(y_i,x,v) = \sum_i \phi(x_{l(y_i)},v)\\ \Psi(y;x) &= \sum_i \psi(y_i,...,y_{i-K})\\ &= \sum_i \psi(x_{l(y_i)},...,x_{l(y_{i-K})})\end{aligned}$$

where K is the size of local context.

Content Selection – Visual Estimates:

The computer vision system used consists of 7404 visual classifiers for recognizing leaf level WordNet synsets (Fellbaum, 1998). Each classifier is trained using labeled images from the ImageNet dataset (Deng et al., 2009) – an image database of over 14 million hand labeled images organized according to the WordNet hierarchy. Image similarity is represented using a Spatial Pyramid Match Kernel (SPM) (Lazebnik et al., 2006) with Locality-constrained Linear Coding (Wang et al., 2010) on shape based SIFT features (Lowe, 2004).

[1] Open domain computer vision remains to be an open problem, and it would be difficult to reliably distinguish pictures of subtle visual differences, e.g., pictures of *"a water front house with a docked boat"* from those of *"a floating house pulled by a boat"*.

[2] Available at `http://www.cs.stonybrook.edu/~ychoi/imgcaption/`

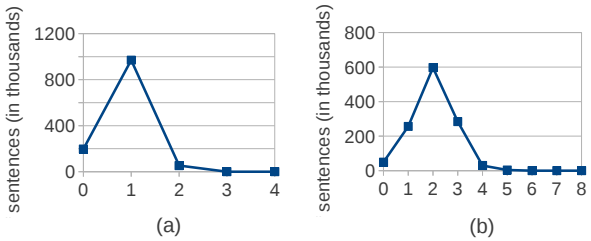

Figure 2: Number of sentences (y-axis) for each *average* (x-axis in (a)) and *maximum* (x-axis in (b)) number of words with future dependencies

Models are linear SVMs followed by a sigmoid to produce probability for each node.[3]

Content Selection – Salient Topics:

We consider Tf.Idf driven scores to favor salient topics, as those are more likely to generalize across many different images. Additionally, we assign a very low content selection score ($-\infty$) for proper nouns and numbers and a very high score (larger then maximum idf or visual score) for the 2k most frequent words in our corpus.

Local Linguistic Fluency:

We model linguistic fluency with 3-gram conditional probabilities:

$$\psi(x_{l(y_i)}, x_{l(y_{i-1})}, x_{l(y_{i-2})}) = p(x_{l(y_i)}|x_{l(y_{i-2})}, x_{l(y_{i-1})}) \quad (1)$$

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).

Dependency-driven Constraints:

Table 1 defines the list of dependencies used as constraints driven from the typed dependencies (de Marneffe and Manning, 2009; de Marneffe et al., 2006). The direction of arrows indicate the direction of inclusion requirements. For example, $dep(X \longleftarrow Y)$, denotes that "X" must be included whenever "Y" is included. Similarly, $dep(X \longleftrightarrow Y)$ denotes that "X" and "Y" must either be included together or eliminated together. We determine the uni- or bi-directionality of these constraints by manually examining a few example sentences corresponding to each of these typed dependencies. Note that some dependencies such as $det(\longleftrightarrow)$ would hold regardless of the particular lexical items, while others, e.g., $dobj(\longleftrightarrow)$ may or may not be necessary depending on the context. Those dependencies that we determine as largely context dependent are marked with * in Table 1.

[3]Code was provided by Deng et al. (2012).

Method-1 (M1)	v.s.	Method-2 (M2)	M1 wins over M2
SALIENCY		ORIG	76.34%
VISUAL		ORIG	81.75%
VISUAL		SALIENCY	72.48%
VISUAL		VISUAL W/O CONSTR	83.76%
VISUAL		NGRAM-ONLY	90.20%
VISUAL		**HUMAN**	19.00%

Table 2: Forced Choice Evaluation (LM Corpus = Google)

One could consider enforcing all dependency constraints in Table 1 as hard constraints so that the compressed sentence must not violate any of those directed dependency constraints. Doing so would lead to overly conservative compression with least compression ratio however. Therefore, we relax those that are largely context dependent as soft constraints (marked in Table 1 with *) by introducing a constant penalty term in the objective function. Alternatively, the dependency based constraints can be learned statistically from the training corpus of paired original and compressed sentences. Since we do not have such in-domain training data at this time, we leave this exploration as future research.

Dynamic Programming with Dynamic Beam:

The constraint optimization we formulated corresponds to an NP-hard problem. In our work, hard constraints are based only on typed dependencies, and we find that long range dependencies occur infrequently in actual image descriptions, as plotted in Figure 2. With this insight, we opt for decoding based on dynamic programming with dynamically adjusted beam.[4] Alternatively, one can find an approximate solution using Integer Linear Programming (e.g., Clarke and Lapata (2006), Clarke and Lapata (2007), Martins and Smith (2009)).

3 Evaluation

Since there is no existing benchmark data for image caption generalization, we crowdsource evaluation using Amazon Mechanical Turk (AMT). We empirically compare the following options:

[4]The required beam size at each step depends on how many words have dependency constraints involving any word following the current one – beam size is at most 2^p, where p is the max number of words dependent on any future words.

Figure 3: Example Image Caption Transfer

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

Table 3: Image Description Transfer: performance in BLEU and F1 with *strict* & *semantic* matching.

- ORIG: original uncompressed captions
- HUMAN: compressed by humans (See § 3.2)
- SALIENCY: linguistic fluency + saliency-based content selection + dependency constraints
- VISUAL: linguistic fluency + visually-guided content selection + dependency constraints
- x W/O CONSTR: method x without dependency constraints
- NGRAM-ONLY: linguistic fluency only

3.1 Intrinsic Evaluation: Forced Choice

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 2. We observe that VISUAL (full model with visually guided content selection) performs the best, being selected over 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 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 VISUAL, with and without dependency-based constraints. As expected, the algorithm with constraints is preferred in the majority of cases.

3.2 Extrinsic Evaluation: Image-based Caption Retrieval

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 description, i.e., a visually descriptive caption. Following Ordonez et al. (2011), we produce a caption for a query image by finding top k most similar images within the 1M image-text corpus (Ordonez et al., 2011) and then transferring their captions to the query image. To compute evaluation measures, we take the average scores of BLEU(1) and F-score (unigram-based with respect to content-words) over $k = 5$ candidate captions.

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 im-

Figure 4: Good (left three, in blue) and bad examples (right three, in red) of generalized captions

ages in the whole dataset using an unweighted sum of gist similarity and tiny image similarity.

Gold standard (human compressed) captions are obtained using AMT for 1K images. The results are shown in Table 3. *Strict matching* gives credit only to identical words between the gold-standard caption and the automatically produced caption. However, words in the original caption of the query image (and its compressed caption) do not overlap exactly with words in the retrieved captions, even when they are semantically very close, which makes it hard to see improvements even when the captions of the new corpus are more general and transferable over other images. Therefore, we also report scores based on *semantic matching*, which gives partial credits to word pairs based on their lexical similarity.[5] The best performing approach with semantic matching is VISUAL (with LM = Image corpus), improving BLEU, Precision, F-score substantially over those of ORIG, demonstrating the extrinsic utility of our newly generated image-text parallel corpus in comparison to the original database. Figure 3 shows an example of caption transfer.

4 Related Work

Several recent studies presented approaches to automatic caption generation for images (e.g., Farhadi et al. (2010), Feng and Lapata (2010a), Feng and Lapata (2010b), Yang et al. (2011), Kulkarni et al. (2011), Li et al. (2011), Kuznetsova et al. (2012)). The end goal of our work differs in that we aim to revise original image captions into descriptions that are more general and align more closely to the visual image content.

In comparison to prior work on sentence compression, our approach falls somewhere between unsupervised to distant-supervised approach (e.g., Turner and Charniak (2005), Filippova and Strube (2008)) in that there is not an in-domain training corpus to learn generalization patterns directly. Future work includes exploring more direct supervision from human edited sample generalization (e.g., Knight and Marcu (2000), McDonald (2006)) Galley and McKeown (2007), Zhu et al. (2010)), and the inclusion of edits beyond deletion, e.g., substitutions, as has been explored by e.g., Cohn and Lapata (2008), Cordeiro et al. (2009), Napoles et al. (2011).

5 Conclusion

We have introduced the task of image caption generalization as a means to reduce noise in the parallel corpus of images and text. Intrinsic and extrinsic evaluations confirm that the captions in the resulting corpus align better with the image contents (are often preferred over the original captions by people), and can be practically more useful with respect to a concrete application.

Acknowledgments

This research was supported in part by the Stony Brook University Office of the Vice President for Research. Additionally, Tamara Berg is supported by NSF #1054133 and NSF #1161876. We thank reviewers for many insightful comments and suggestions.

[5]We take Wu-Palmer Similarity as similarity measure (Wu and Palmer, 1994). When computing BLEU with semantic matching, we look for the match with the highest similarity score among words that have not been matched before. Any word matched once (even with a partial credit) will be removed from consideration when matching next words.

References

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. In *Linguistic Data Consortium*.

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.

James Clarke and Mirella Lapata. 2007. Modelling compression with discourse constraints. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1–11, Prague, Czech Republic, June. Association for Computational Linguistics.

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.

Joao Cordeiro, Gael Dias, and Pavel Brazdil. 2009. Unsupervised induction of sentence compression rules. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pages 15–22, Suntec, Singapore, August. Association for Computational Linguistics.

Marie-Catherine de Marneffe and Christopher D. Manning. 2009. Stanford typed dependencies manual.

Marie-Catherine de Marneffe, Bill Maccartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Language Resources and Evaluation Conference 2006*.

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*.

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*.

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.

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*.

Christiane D. Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Yansong Feng and Mirella Lapata. 2010a. How many words is a picture worth? automatic caption generation for news images. In *Association for Computational Linguistics*.

Yansong Feng and Mirella Lapata. 2010b. Topic models for image annotation and text illustration. In *Human Language Technologies*.

Katja Filippova and Michael Strube. 2008. Dependency tree based sentence compression. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *AAAI/IAAI*, pages 703–710.

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*.

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.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching. In *Conference on Computer Vision and Pattern Recognition*, June.

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.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November.

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.

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.

Courtney Napoles, Chris Callison-Burch, Juri Ganitkevitch, and Benjamin Van Durme. 2011. Paraphrastic sentence compression with a character-based metric: Tightening without deletion. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 84–90, Portland, Oregon, June. Association for Computational Linguistics.

Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

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.

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.

Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, T. Huang, and Yihong Gong. 2010. Locality-constrained linear coding for image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China, August. Coling 2010 Organizing Committee.