Alignment and annotation issues with English and German image captions

Oliver Czulo[1], Marcelo Viridiano[2] & Tiago Timponi Torrent[2]
[1]Leipzig University, oliver.czulo@uni-leipzig.de
[2] University of Juiz de Fora, marcelo.viridiano@gmail.com & tiago.torrent@ufjf.br

Keywords: multimodality, frame semantics, image annotation, translation

Current work in the field of Multimodal Machine Translation (Elliott et al. 2016; Lan et al. 2017; Barrault et al. 2018; Nakayama et al. 2020; Torrent et al. 2022) has been focusing on expanding the popular benchmark dataset for sentence-based image description Flickr30k for multiple languages (Young et al. 2014). The dataset comprises image-caption pairs with multiple crowd-sourced descriptions per image. Preliminary experiments (Viridiano et al. 2022) applied the FrameNet annotation methodology to assess frame semantic similarity across languages and across communicative modes to extensions of the Flickr30K dataset – specifically the Multi30k dataset (Elliott et al. 2016) and the Flickr 30K Entities dataset (Plummer et al. 2015). The results make the case for the adoption of annotation practices that recognize and represent the inherently perspectivized nature of multimodal communication.

In this contribution, we will report on issues encountered during a two-step alignment and annotation task of English image captions and their German translations during a project course with students at Leipzig University. These issues concern the alignment of English noun phrases with correspondent phrases in the German translation, and the subsequent assignment of frames and frame elements for image entities correlated via bounding boxes with noun phrases in both English and German captions.

English noun phrases and their corresponding bounding boxes are pre-annotated in Flickr 30k Entities and could not be changed by students, so as to ensure mapability onto the original dataset. The current setup of the data set is guided by the English original captions, with bounding boxes only created if an entity is referred to by a noun phrase. In the alignment, the shortcoming of this shows: For instance, there are multiple cases in which nouns in English are translated by a verb in German, such as in example (a) (see below), which, by the current logic in the dataset, would leave the bounding box linked to *a gathering* devoid of a corresponding linguistic element in German. On the other hand, in some cases, adjectives embedded in noun phrases in English are translated as nouns in German, as in example (b), which by the current logic would require an additional bounding box for the *Schnauzer* 'mustache'.

As outlined in frame semantic translation analysis (Czulo 2017; Torrent et al. 2018) and frame semantic multimodal annotation (Belcavello et al. 2020), formal divergences between originals and translations do not necessarily lead to a difference in semantics, or to one that cannot be explained by means of frame relations, resulting in equally plausible descriptions of images. We will present an initial classification of formal divergences and their impact within the current annotation setting. On top of that, we will make another case for the multi-perspectivity of image captioning and how frame semantic analysis can help us uncover the constraint space for potential interpretations.

Examples

(a) EN: Indians having a gathering with coats and food and drinks.
DE: Indios in Umhängen versammeln sich zum Essen und Trinken.
lit. ‚Indians in coats are gathering for food and drink‘

(b) EN: A mustached man in a white shirt
DE: Ein Mann mit Schnauzer und weißem Hemd
lit. ‚A man with mustache and white shirt‘

References

Barrault, Loïc, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott & Stella Frank. 2018. Findings of the Third Shared Task on Multimodal Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 304–323. Belgium, Brussels: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6402.

Belcavello, Frederico, Marcelo Viridiano, Ely Matos & Tiago Timponi Torrent. 2022. Charon: A FrameNet annotation tool for multimodal corpora. In *Proceedings of the 16th lingusitic annotation workshop (LAW-XVI) within LREC2022*, 91–96. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lawxvi-1.11.

Czulo, Oliver. 2017. Aspects of a primacy of frame model of translation. In S. Hansen-Schirra, Oliver Czulo & Sascha Hofmann (eds.), *Empirical modelling of translation and interpreting* (Translation and Multilingual Natural Language Processing 6), 465–490. Berlin: Language Science Press.

Elliott, Desmond, Stella Frank, Khalil Sima'an & Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. arXiv. http://arxiv.org/abs/1605.00459.

Fillmore, C. J. 2003. Framenet in Action: The Case of Attaching. *International Journal of Lexicography* 16(3). 297–332. https://doi.org/10.1093/ijl/16.3.297.

Lan, Weiyu, Xirong Li & Jianfeng Dong. 2017. Fluency-Guided Cross-Lingual Image Captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, 1549–1557. Mountain View California USA: ACM. https://doi.org/10.1145/3123266.3123366.

Miltenburg, Emiel van. 2016. Stereotyping and Bias in the Flickr30K Dataset. arXiv. http://arxiv.org/abs/1605.06083.

Nakayama, Hideki, Akihiro Tamura & Takashi Ninomiya. 2020. A Visually-Grounded Parallel Corpus with Phrase-to-Region Linking.

Plummer, Bryan A., Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier & Svetlana Lazebnik. 2016. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. arXiv. http://arxiv.org/abs/1505.04870.

Torrent, Tiago Timponi, Michael Ellsworth, Collin F. Baker & Ely Matos. 2018. The Multilingual FrameNet Shared Annotation Task: a Preliminary Report. In Tiago Timponi Torrent, Lars Borin & Collin F. Baker (eds.), *Proceedings of the International FrameNet Workshop 2018: Multilingual Framenets and Constructicons*. Paris, France: European Language Resources Association (ELRA).

Torrent, Tiago Timponi, Ely Edison da Silva Matos, Frederico Belcavello, Marcelo Viridiano, Maucha Andrade Gamonal, Alexandre Diniz da Costa & Mateus Coutinho Marim. 2022. Representing Context in FrameNet: A Multidimensional, Multimodal Approach. *Frontiers in Psychology* 13. 838441. https://doi.org/10.3389/fpsyg.2022.838441.

Viridiano, Marcelo, Tiago Timponi Torrent, Oliver Czulo, Arthur Lorenzi, Ely Matos & Frederico Belcavello. 2022. The case for perspective in multimodal datasets. In *Proceedings of the 1st workshop on perspectivist approaches to NLP @LREC2022*, 108–116. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.nlperspectives-1.14.

Young, Peter, Alice Lai, Micah Hodosh & Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2. 67–78. https://doi.org/10.1162/tacl_a_00166.