Investigating language drift in multi-agent communication about natural images: MSc Thesis Proposal V2

By Polina Tsvilodub, *February 16th 2022*

The following proposal refines the topic for my MSc thesis. More specifically, the thesis will focus on identifying and potentially counteracting language drift which arises in multi-agent communication experiments employing natural language as the communication protocol. The experiments will be based on the MS COCO Captions dataset which two agents will communicate about in a reference game setting [Chen et al., 2015, Lazaridou et al., 2016].

Along with this outline, I put together a thesis schedule, describing the time frame of the project. In this proposal, all work *milestones* are enumerated (preceded by an *M*) in the order I will work on them. You can find respective milestones and corresponding work step numbers on the right of the text. The milestones are referenced in the schedule. This outline is structured according to the suggested flow of the thesis (i.e., from introduction to conclusion). Additionally, a *pseudo algorithm* of the experiment is provided at the end of the write-up. Comments in [Time: green] highlight the aspects that I assume might be time intensive or which would have an effect on project timing, while comments in [Unclear: blue] highlight design decisions which I am unsure about.

Most important dates:

- Register thesis by May 23rd 2022 — M7.1
- Finish thesis draft by August 1st 2022 — M7.2
- Submit thesis by August 31st 2022 — M7.4

Outline:

1. Introduction & Theoretical Background on Multi-Agent Communication

   This section will cover the motivation of current work, as well as review related work in the area of multi-agent communication. — Steps 2, 5; M2.1, M5.1

   - The area of multi-agent communication has gained increased popularity (e.g., [Lazaridou and Baroni, 2020]). This thesis particularly focuses on multi-agent communication experiments employing communication in natural language (i.e., English) (e.g., [Andreas and Klein, 2016, Mao et al., 2016, Lazaridou et al., 2020, Gupta et al., 2021]).
   - The proposed experiment will use a reference game setting, so the relavant background as well as related multi-agent reference game experiments will

be reviewed (e.g., [Lazaridou et al., 2016]). Other tasks used in multi-agent communication experiments will also be reviewed (e.g., [Jaques et al., 2019]).

- The thesis will take multi-agent communication one stop closer towards real-world applications by using real images from the **MS COCO Captions dataset** [Chen et al., 2015]. This dataset features photographic images alongside with five manually annotated captions per image, assigned categories (80 in the entire dataset), object bounding boxes, as well as other features like whether the image is crowded or not.
- Previous work either uses such a dataset (e.g., [Havrylov and Titov, 2017]), or focuses on using natural language. Some experiments combine both by using single natural language labels for the images (e.g., [Lazaridou et al., 2016]), but full captions have not been used yet.
- Therefore, current work will focus in combining natural images and natural language, while particularly zooming in on the quality of the language.

2. Theoretical Background on Image Captioning

   This chapter will cover necessary background in neural image captioning. Step 2 M2.1

   - The experiments conducted in the thesis will use the overall architecture proposed by [Lazaridou et al., 2020]. That is, the captioning module will consist of an LSTM cell, as proposed by [Vinyals et al., 2015], while the image embedding module will be based on a pretrained ResNet-50 [He et al., 2016] (more details below).
   - This chapter will review different approaches to image captioning (e.g., [Karpathy and Fei-Fei, 2015, Vinyals et al., 2015, Vedantam et al., 2017, Kim et al., 2021]). Other related tasks like visual question answering will be mentioned, as well.
   - Both multi-agent communication as well as image captioning might be prone to *language drift*, i.e., the decreasing quality and intelligibility of natural language as task-specific training proceeds. However, this issue hasn't received much attention in the literature (but see e.g., [Lee et al., 2019]), so current work will specifically focus on language drift. The next chapter outlines the experimental setting, followed by an in-depth discussion of language drift in Chapter 4.

3. Own Experiment

   In order to investigate language drift, several multi-agent reference game experiments would be conducted. Their design will closely follow [Lazaridou et al., 2020]. This section will describe the dataset, experimental set up and training procedure. A more formal summary can be found under Algorithm 1. [Train: Steps in orange] indicate parts of the experiment which I assume to require training. Preferably, the experiments Steps 1, 3, 4

would be implemented using TensorFlow [Abadi et al., 2016]. First, "minimal viable product" (MVP) versions of the models would be implemented and trained on a small scale locally, to ensure that the architecture works. Following that the models would be trained at full scale. M1.4 M3.2

- In the reference game, two neural agents will receive pairs of images, and the speaker agent's task will be to produce a message identifying the target image among the two. The listener agent's task will be, given the speaker's message, to pick the correct image among the two. This reference game set up allows to introduce a cognitively plausible task objective—producing a maximally discriminative caption for the target image, highlighting its differences to the distractor image ([Andreas and Klein, 2016, Dai and Lin, 2017, Vedantam et al., 2017, Nie et al., 2020]).
- Each agent will consist of two modules: the visual module which produces a feature vector given raw images, and the language module which produces or decodes messages (i.e., image captions), respectively. Following [Lazaridou et al., 2020], the visual encoding modules for both agents will consist of pre-trained ResNet-50 modules (available both in TensorFlow and Pytorch, [He et al., 2016]). Following [Vinyals et al., 2015], the language module of the speaker will consist of an LSTM which will produce messages conditioned on the image embeddings. KL regularization will be used.
- [Time: The speaker will be trained to produce discriminative captions by using a "multi-task" approach where the agent is optimized to produce both structurally fluent and functionally correct (i.e., discriminative) captions [Lazaridou et al., 2020].] (see Alg. 1) This architecture might be prone to language drift, making it a good avenue for investigating this phenomenon. [Unclear: I am not sure if this is the optimal loss / architecture to use. An alternative would be to use a pretrained captioning module, only fine-tuning it on the task.]
- [Time: Two versions] of this experiment will be conducted. The first version would use **random image pairs** as target and distractor during training. Because the dataset consists of images belonging to 80 different categories, images from different categories might feature completely different scenes, making detailed descriptions superfluous. This first experiment will serve as a baseline, and to investigate whether the captions remain short and reference minimal necessary features, or if the speaker overgenerates. Furthermore, this experiment will be a baseline for checking if the desired set up (i.e., the dataset and natural language) works at all. M1.5, M3.1
- The second version would use **within-category** target and distractor image pairs. To group the images, image pairs would be constructed by using images for which

at least two category labels are the same (each image has several category annotations provided by each of the annotators which might vary). Using more similar images would, ideally, pressure the speaker to produce more detailed, longer captions. This second experiment might give the room necessary for investigating language drift.

- The two experiment versions would be compared with respect to referential success of the agents in order to validate the experiment. Descriptive analyses will also be conducted, like comparing average caption lengths for the two experiments. The caption complexity could also be compared as a function of target/distractor similarity (computed, e.g., [Unclear: as cosine similarity between the image embeddings]).
- [Time: For both experiments, the idea would be to use roughly a quarter of the categories (i.e, 20), yielding around 70.000 images.] One could use about 5-10% of images in each category used for training as validation data (around 3.500 images in total). Two test splits could be created: one split could contain images with categories partially observed during training (easy split), while images from, e.g., five unseen categories can be used as a difficult test split. It is hard to consider the categories as clearcut "novel", though, because many images have different category labels assigned to them. This issue could be tackled by only using images with, e.g., at least three out of five (non)overlapping categories. M1.3
- [Unclear: Maximal message length could be set to the maximal caption length appearing in the dataset. The vocabulary could consist of all the tokens appearing in the entire dataset.]
- [Unclear: Potentially, a different model as a baseline could be useful.]
- [Unclear: For both experiments, the models would be trained until convergence.]

4. Analysing Language Drift

This thesis aims to compare state-of-the-art language drift metrics and to attempt to develop a novel metric. Therefore, this section will review existing language drift metrics (e.g., [Lazaridou et al., 2020, Lu et al., 2020]), provide theoretical background for the proposed metrics and analysis, and present several approaches to developing a new metric. Existing drift metrics which are applicable to the proposed experiment would all be computed for both experiments. More specifically, they could be computed both during the training (e.g., every 50 training steps) to assess the dynamics of language development, as well as for the trained models on the test split. A more formal overview of the required models can be found under Algorithm 2. Existing language drift metrics include: Steps 1, 3, 4

(a) Structural / syntactic language drift: log probability of the generated message m under a pre-trained unconditional language model $P(m)$ [Lazaridou et al., 2020]

(b) Semantic language drift: conditional log probability $P(m|i)$ of the generated message m given the image i. Another measure includes the n-gram overlap of generated messages and the ground-truth captions (ignoring stopwords) [Lazaridou et al., 2020]. Semantic drift is also addressed by [Lee et al., 2019, Lu et al., 2020], but their approaches rather propose specific training methods than measures for identifying language drift, so their proposals wouldn't be considered here. In alternative framings, semantic drift has been measured as the difference between the message semantics and the action taken by the receiver agent [Jacob et al., 2021].

(c) Pragmatic language drift: [Lazaridou et al., 2020] assess this drift as referential failure in absense of structural or semantic drift by comparing human and listener agent referential success given a model which only reranks ground-truth captions. Given that the currently proposed experiment won't have human data, this kind of drift will have to be assessed differently. Although the proposed approach has the advantage of being task-agnostic, in this work, I would propose to focus on a referential task drift which I referred to as *functional* drift.

Similar to the experimental models, first, MVPs of the models for language drift analysis would be implemented and tested, before training them full-scale. M1.7 M3.2

A novel metric would focus on evaluating both structural and **functional drift** of the produced expressions. In this context, funtional drift would refer to the deterioration of language which would make the referential task impossible for humans (e.g., leaving out critical content words). Structural drift, in contrast, might involve mixing up the word order, which nevertheless wouldn't hinder the referential task, if distinctive content words are still present. For instance, the caption "A plate food with" would exemplify functional drift, while the caption "A plate red food with" wouldn't, if Figure 2 was the target image and Figure 1 was the distractor.

Different approaches to developing such a metric could be taken in the thesis.

(a) One idea for identifying functional language drift which would also be stable against compositional alternations within the caption would be to compute the word overlap between the generated captions and the target and distractor ground truth captions, respectively. From a functional perspective, an optimal generated target caption would maximize the overlap with the target ground truth, while minimizing the overlap with the distractor ground truth. This idea is related to the omission score suggested by [Havrylov and Titov, 2017]. [Unclear: This descriptive metric might possibly even be converted into a version of a contrastive loss for training the speaker agent (cf. [Andreas and Klein, 2016,

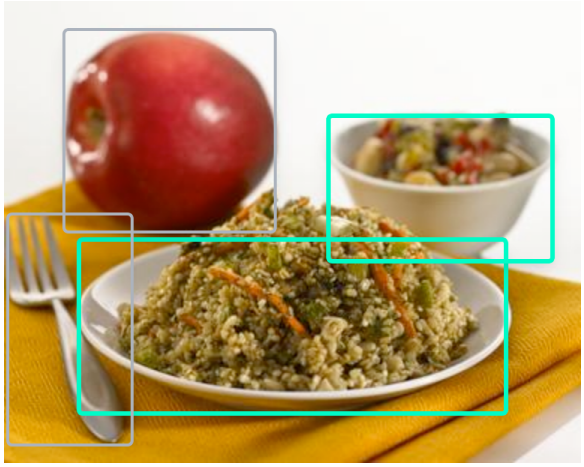

Figure 1: Example image from the MS Coco Captions dataset. Example caption: “The plate is piled with rice next to a whole apple.”. Assigned categories: [fork, bowl, apple, bowl]

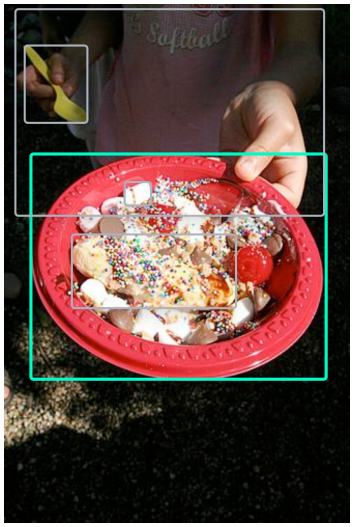

Figure 2: Example image from the MS Coco Captions dataset. Example caption: “A person holding a red bowl filled with cake.”. Assigned categories: [person, spoon, bowl, banana, banana]

Gunel et al., 2020]).]

(b) Alternatively, the idea described above could be formalized by computing the cosine similarity between the caption embeddings instead of word overlap scores.

(c) A more structured approach to identifying whether critical discrimnative components have been captured in the caption might be to consider the labeled objects in the two images, identify the ones that are only present in the target image, and perform image patch to caption alignment, as proposed by [Karpathy and Fei-Fei, 2015]. Higher alignment scores would indicate more discriminative captions. [Unclear: However, the availability of the model and the usefulness of the approach are not completely clear, as the similarity of images even within MS Coco Captions categories might be quite variable.]

(d) Finally, we discussed the idea to measure the drift as the similarity between the original image and an image generated by a pretrained text-to-image model given the generated caption. Some text-to-image architectures include DALL-E, StackGAN++ or other models [Ramesh et al., 2021, Zhang et al., 2018, Zhou et al., 2021]. [Time: However, it seems rather difficult to find models, pretrained on our dataset. Example sources for other datasets might be `https://tinyurl.com/y4pz7ymz` or `https://github.com/ShanHaoYu/Text2Image`. Although comprehensive guides to training these models exist, computational demands for this task might be rather cosmic.] [Unclear: Furthermore, a similarity metric for the original and generated image would be necessary. One could potentially compare embeddings of these images extracted by the ResNet module via cosine similarity, but the interpretability of such a comparison might be unclear]. M1.6, M3.2

(e) Based on this idea, [Time: a different approach to training could be considered,] whereby an image could be generated from the ground truth caption as an intermediate representation (cf. [Lee et al., 2019]). [Unclear: Specifics of the architecture would have to be determined, provided the availability of a text-to-image model.]

If feasible, all these approaches could be tested in similar way as the exisiting metrics (i.e., during training and on the test split), and compared with respect to their accuracy and functional adequacy by manual inspection of caption samples. M3.3, M4.1

5. Conclusion & Discussion

This section will summarize the work, outline its contributions and limitations. More specifically, the generalizability of the considered drift metrics to different tasks and different datasets will be discussed. The susceptibility of the metrics to design decisions regarding the experimental set up, as well as properties of the dataset will be covered. Depending on the results, the reasons for failure or improvements will be Step 6

discussed. Finally, an outlook to directions for future work will be given.

References

[Abadi et al., 2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). {TensorFlow}: A system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283.

[Andreas and Klein, 2016] Andreas, J. and Klein, D. (2016). Reasoning about pragmatics with neural listeners and speakers. *arXiv preprint arXiv:1604.00562*.

[Chen et al., 2015] Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

[Conneau et al., 2017] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

[Dai and Lin, 2017] Dai, B. and Lin, D. (2017). Contrastive learning for image captioning. *arXiv preprint arXiv:1710.02534*.

[Gunel et al., 2020] Gunel, B., Du, J., Conneau, A., and Stoyanov, V. (2020). Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

[Gupta et al., 2021] Gupta, A., Lanctot, M., and Lazaridou, A. (2021). Dynamic population-based meta-learning for multi-agent communication with natural language. *Advances in Neural Information Processing Systems*, 34.

[Havrylov and Titov, 2017] Havrylov, S. and Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Jacob et al., 2021] Jacob, A. P., Lewis, M., and Andreas, J. (2021). Multitasking inhibits semantic drift. *arXiv preprint arXiv:2104.07219*.

[Jaques et al., 2019] Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. (2019). Social influence as intrinsic motivation

for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pages 3040–3049. PMLR.

[Karpathy and Fei-Fei, 2015] Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

[Kim et al., 2021] Kim, W., Son, B., and Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

[Lazaridou and Baroni, 2020] Lazaridou, A. and Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.

[Lazaridou et al., 2016] Lazaridou, A., Peysakhovich, A., and Baroni, M. (2016). Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*.

[Lazaridou et al., 2020] Lazaridou, A., Potapenko, A., and Tieleman, O. (2020). Multi-agent communication meets natural language: Synergies between functional and structural language learning. *arXiv preprint arXiv:2005.07064*.

[Lee et al., 2019] Lee, J., Cho, K., and Kiela, D. (2019). Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*.

[Lu et al., 2020] Lu, Y., Singhal, S., Strub, F., Courville, A., and Pietquin, O. (2020). Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pages 6437–6447. PMLR.

[Mao et al., 2016] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

[Nie et al., 2020] Nie, A., Cohn-Gordon, R., and Potts, C. (2020). Pragmatic issue-sensitive image captioning. *arXiv preprint arXiv:2004.14451*.

[Ramesh et al., 2021] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

[Vedantam et al., 2017] Vedantam, R., Bengio, S., Murphy, K., Parikh, D., and Chechik, G. (2017). Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

[Vinyals et al., 2015] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

[Zhang et al., 2018] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2018). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962.

[Zhou et al., 2021] Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., and Sun, T. (2021). Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*.

Algorithm 1 Basic reference game set up

$|images| = n$
$|caption_i| = t$ for the caption of the $i-$th image
f: LSTM network
h_m : LSTM hidden state
x_m: m-th input token to the LSTM, $m \in 1, ..., t$
y_m: m-th ouput token of the LSTM
for image i in $[1, n]$ **do**
Speaker agent:
target image embedding $u_i \leftarrow$ ResNet-50
while $m + 1 <$ max caption length or $y_{m+1} \neq$ STOP **do**
for caption token x_m in $[1, t]$ of image i **do**
maximize $logP(x_m|x_0...x_{m-1}, u_i) = \sum_{m=0}^{m-1} logP(x_{m-1}|x_0...x_{m-1}, u_i)$
[Unclear: $h_0 \leftarrow u$]
[Train: $h_{m+1} = f(h_m, [x_m, u_i])$; 3 gates;] [Unclear: concatenated x_m and u_i]
[Train: $y_{m+1} = Softmax(h_m)$] over all words in vocabulary
end for
end while
Loss speaker:
$L_S = \lambda_f L^{functional} + \lambda_s L^{structural} = -r^L(caption, u, i) \sum_{m=0}^{t} logP_{\theta_S^{LSTM}}(caption^m|caption^{<m}, [u_i, u_d]) + \lambda_s \sum_{m=0}^{t} logP_{\theta_S^{LSTM}}(caption^m|caption^{<m}, u_i) +$ `KL-regularisation` where t is longest caption in the dataset, updating $L^{functional}$ via `reinforce`

Listener agent:
target image embedding $u_i \leftarrow$ ResNet-50
distractor image embedding $u_d \leftarrow$ ResNet-50 (sampled either at random or within-category of the target)
for token y_m in $[1, t]$ of received message y **do**
[Train: $v_m = f(y_m)$]
end for
[Unclear: caption embedding: $v = maxpool(v_1...v_t)$ [Conneau et al., 2017]]
target = $max_{i,d}(v \times u_i, v \times u_d)$ where $\times$ is the dot product similarity
Loss listener:
$L_L = -(target\ log(P(i)) + (1 - target) log(1 - P(i)))$

Loss:
if target identified correcly **then**
Reward $r^L = 1$
else
Reward $r^L = -1$
end if
[Unclear: $L = L_S + L_L$ (cf. [Lee et al., 2019]). Only LSTM and projection components are trainable, although e.g., [Vinyals et al., 2015] also update the top CNN layer.]
end for

Algorithm 2 Set up of models required for drift metrics

Semantic drift: Pretrained language model
Compute perplexity of the captions under a pretrained model, e.g., GPT-2 provided by the `huggingface` library

Structural drift: Image captioning model
[Train: LSTM trained with caption log likelihood objective only, as proposed by [Vinyals et al., 2015]], such that the architecture maximally closely matches experimental architecture in Algorithm 1. The same training split would be used.
Loss: $L = -\sum_{m=0}^{t} logP_m(x_m)$ for a caption of length t. [Unclear: The same components of the model should be trainable as in Alg. 1. The availability of a pretrained model will also be investigated]

Functional drift: [Unclear: Text-to-image]
[Train: The pretrained model by `https://github.com/ShanHaoYu/Text2Image` is implemented in TensorFlow, which would allow to easily include it in the pipeline. However, it is unclear on which dataset it is trained. It might potentially need finetuning, since the original implementation by [Zhang et al., 2018] also trains on different datasets than MS Coco.]