Using visual context(s) in neural REG

Simeon Schüz

CLAUSE Group
Bielefeld University

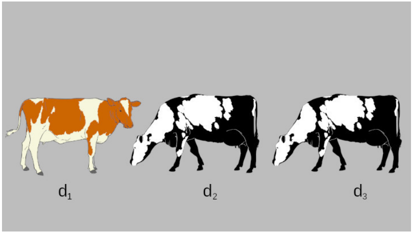
07 June 2022

Outline

- Background
  - Classical & end-to-end formulations of the REG task
  - Context in REG
- Using different kinds of visual context to...
  - Discriminate
  - Identify
  - Facilitate Understanding
  - Guide Perspectives
- Conclusion
- Questions for Subsequent Research

The REG task

- **Referring Expression Generation (REG)**: Generating expressions for entities, which allow their identification in context



Referring Expression for d_1:
"The brown / left cow"

Classical REG formulation

- Classical REG formulation: Dale and Reiter (1995)
  - Focus on **Content selection** $\rightarrow$ selecting discriminative properties
  - Entities in a domain: Represented in terms of pairs of **symbolic attributes and values**
- Subsequent research: Various extensions, e.g. accounting for multiple targets, relational descriptions, discourse context, ...

	Object	*Type*	*Color*	*Location*
$\rightarrow$	d_1	cow	brown	left
	d_2	cow	black & white	middle
	d_3	cow	black & white	right

Discriminative properties for d_1: $\{\langle color, brown\rangle\}$ or $\{\langle location, left\rangle\}$

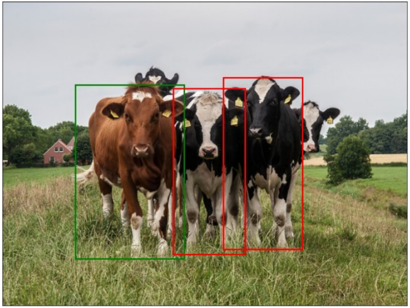
Context in classical REG

- (Extralinguistic) context: Other entities in the current domain
- Mainly important as **Distractors** during Content Selection
  - needed to determine distinguishing properties of the target
  - **used for discrimination**, i.e. to ensure pragmatic informativity (in line with the overarching goal in REG)

	Object	*Type*	*Color*	*Location*
$\rightarrow$	d_1	cow	brown	left
	d_2	cow	black & white	middle
	d_3	cow	black & white	right

End-to-end REG modelling

- Neural REG (e.g. Mao et al. 2016; Yu et al. 2016)
  - looking at **more realistic settings**, i.e. objects in photographs (e.g. the RefCOCO dataset, Kazemzadeh et al. 2014)
  - **Encoder/Decoder** models, allowing for **end-to-end** generation
  - Closely related to image captioning, but for individual objects and with a defined pragmatic objective (i.e. refering *unambiguously*)



Context in neural REG

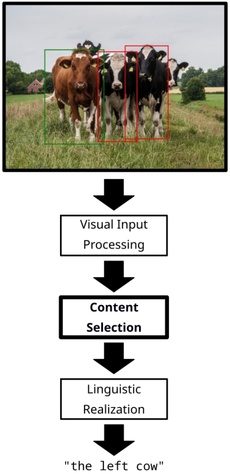
- Context in neural REG: Visual information surrounding the target
- Integration in REG models: Often as a *global feature vector*
- In general: Visual context no major concern in neural REG
  - if discussed explicitly: Mostly as *distractor* context
  - further ways to include visual context are seldom discussed (with some exceptions)

Classical vs. neural REG

- In general: Same objective, i.e. allowing the identification of the referent
- Fundamental differences in other regards:
  1. **Input Representations**:
     Symbolic vs. raw visual information
  2. **Outputs**:
     Sets of distinguishing properties vs. Full descriptions of referential targets
- Neural REG: Broader **scope of the task**
  - Classical REG task: Modelling a **specific processing stage** of a more general NLG pipeline (cf. Reiter 2000)
  - Neural REG task: Modelling a **complete (end-to-end) processing pipeline**, spanning from from perception to formulation

Classical vs. neural REG

- neural REG pipeline (conceptually) includes Content Selection, amongst other challenges:
  1. identify entities in the raw visual input
  2. select informative properties (Content Selection)
  3. formulate an utterance
     - 3.1 decide on appropriate lexical terms for depicted entities
     - 3.2 ensure that expression is easy to understand
- more sub-problems $\rightarrow$ more possibilities to integrate context!



Context in neural REG

- we argue: additional challenges in the end-to-end pipeline
  $\rightarrow$ further possibilities to integrate visual context
- we could use visual context for:
  1. **discrimination**
  2. **identification**
  3. **understandability**
  4. **choosing categorizations / conceptual perspectives**

Using Context to Discriminate

- **Distractors**: other (similar) objects in the same scene we have to distinguish the target from $\rightarrow$ at the core of the REG task itself
- Relevant for Content Selection; contextual information helps in determining distinguishing features
  - satisfying the Gricean maxim of Quantity (Grice, 1975)

Annotations:

(a) “white horse”

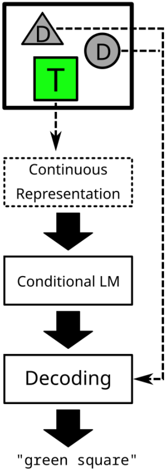
(b) “white horse”

(c) “white house”

(Example from RefCOCO)

Using Context to Discriminate

- relates to our previous work on *discriminative decoding* (Schüz, Han, and Zarrieß, 2021; Schüz and Zarrieß, 2021; Zarrieß et al., 2021):
  - integrating distractors during decoding, e.g. using the RSA model for pragmatic reasoning
  - findings: pragmatic informativity increases when distractors are given more weight
  - (as well as linguistic diversity, which was another topic of interest for us)

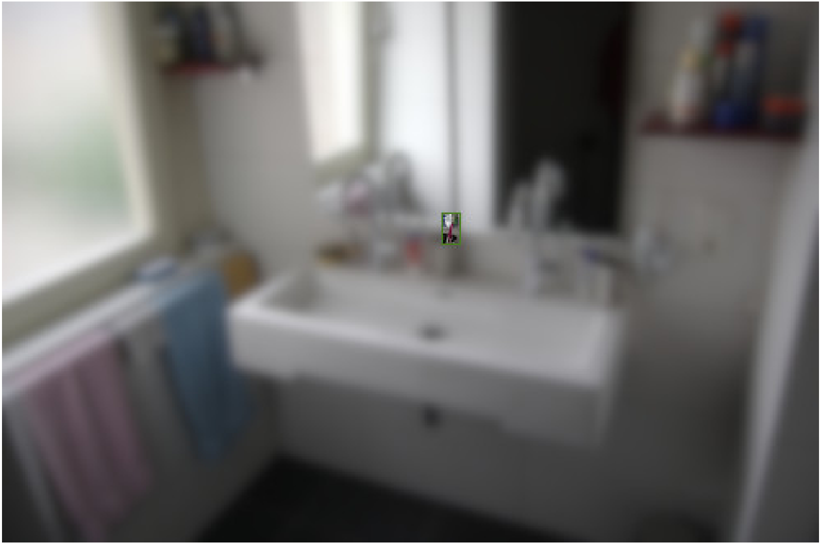


Using Context to Identify

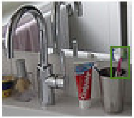
- Raw visual input makes identification a non-trivial task (e.g. Zarrieß and Schlangen 2016)
- Humans use **scene context** to facilitate identification (and other cognitive tasks, e.g. Võ 2021)
  - local features (e.g. *anchors*)
  - global features (e.g. the *gist* of a scene)
- Relevant for identification as an early stage in the neural REG pipeline
  - Especially useful for deficient visual representations (difficult lighting conditions, occlusion, small size, ...)
  - relates to specifying the *type* attribute in classical REG (Dale and Reiter, 1995)
  - satisfying the Gricean Maxim of Quality

Using Context to Identify

Using Context to Identify



Using Context to Identify



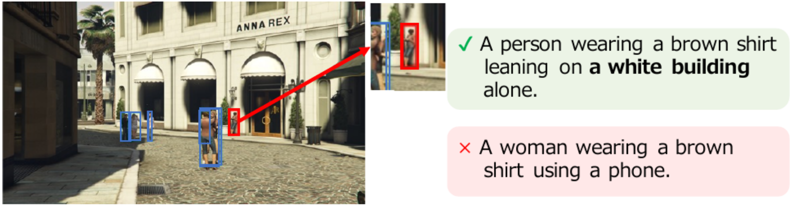
Using Context to Identify

(Example from VisualGenome)

Using Context to Facilitate Understanding

- targets with low saliency might be hard to identify, even if described unambiguously
- possible solution: guiding visual search via **landmarks**
  - i.e. describing targets via their relations to more salient objects in their surroundings
  - related to generating relational descriptions in REG (e.g. Krahmer and Theune 2002; Krahmer, Erk, and Verleg 2003)
  - investigated for neural REG as well, but without looking at visual context in detail (Tanaka et al., 2019; Kim, Ko, and Wu, 2020)
  - satisfying the Gricean Maxim of Manner

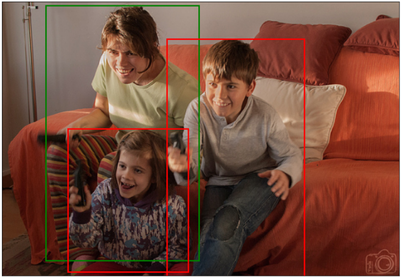
Using Context to Facilitate Understanding



Example from Tanaka et al. (2019)

Using Context to Guide Perspectives

- there are often multiple ways to **categorize and name** a referential target
- choosing a categorization $\rightarrow$ choosing a **conceptual perspective** (Gatt and Deemter, 2006; Gatt and Deemter, 2007)
- assumption: visual context can guide conceptual perspective, e.g. if surrounding objects or depicted actions suggest a certain role
  - e.g. conceptualizing a woman as *mother*, if shown as interacting with children



Annotations:
(a) “mother”
(b) “mom on left”
(c) “parent”
(Example from RefCOCO)

Conclusion

- The neural REG task became more complex: from content selection to modeling a generation pipeline
- Visual Context could play a role at different stages of this pipeline
- For neural REG, this has not been investigated in detail yet

Questions for subsequent research

- How can we assess the influence of different kinds of visual context in human annotations?
- How should visual context be represented in neural REG? (Effectiveness of global feature vectors is sometimes questioned, e.g. Yu et al. 2016; Zarrieß and Schlangen 2018)
- Can we support different stages of the neural REG pipeline by integrating different kinds of context?

Citations I

Dale, Robert and Ehud Reiter (Apr. 1995). “Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions”. In: *Cognitive Science* 19.2, pp. 233–263. DOI: 10.1207/s15516709cog1902_3.

Gatt, Albert and Kees van Deemter (July 2006). “Conceptual Coherence in the Generation of Referring Expressions”. In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Sydney, Australia: Association for Computational Linguistics, pp. 255–262. URL: https://aclanthology.org/P06-2033.

— (Sept. 2007). “Lexical Choice and Conceptual Perspective in the Generation of Plural Referring Expressions”. In: *Journal of Logic, Language and Information* 16.4, pp. 423–443. DOI: 10.1007/s10849-007-9047-0.

Citations II

- Grice, H. P. (1975). "Logic and Conversation". In: *Syntax and Semantics: Vol. 3: Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. New York: Academic Press, pp. 41–58. URL: `http://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf`.
- Kazemzadeh, Sahar et al. (Oct. 2014). "ReferItGame: Referring to Objects in Photographs of Natural Scenes". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 787–798. DOI: `10.3115/v1/D14-1086`. URL: `https://www.aclweb.org/anthology/D14-1086`.

Citations III

Kim, Jungjun, Hanbin Ko, and Jialin Wu (Dec. 2020). "CoNAN: A Complementary Neighboring-based Attention Network for Referring Expression Generation". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 1952–1962. DOI: 10.18653/v1/2020.coling-main.177. URL: https://aclanthology.org/2020.coling-main.177.

Krahmer, Emiel, Sebastiaan van Erk, and André Verleg (2003). "Graph-based generation of referring expressions". In: *Computational Linguistics* 29.1, pp. 53–72.

Citations IV

- Krahmer, Emiel and Mariet Theune (2002). "Efficient context-sensitive generation of referring expressions". Undefined. In: *Information Sharing: Reference and Presupposition in Language Generation and Interpretation, Center for the Study of Language and Information-Lecture Notes*. Ed. by K. van Deemter and R. Kibble. Lecture Notes 143. CSLI Publications, pp. 223–264. ISBN: 1-57586-404-5.
- Mao, Junhua et al. (2016). "Generation and Comprehension of Unambiguous Object Descriptions". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20.
- Reiter, Ehud (2000). *Building natural language generation systems*. Casmbridge, U.K. New York: Cambridge University Press. ISBN: 9780511519857.

Citations V

- Schüz, Simeon, Ting Han, and Sina Zarrieß (July 2021). "Diversity as a By-Product: Goal-oriented Language Generation Leads to Linguistic Variation". In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Singapore and Online: Association for Computational Linguistics, pp. 411–422. URL: `https://aclanthology.org/2021.sigdial-1.43`.
- Schüz, Simeon and Sina Zarrieß (Oct. 2021). "Decoupling Pragmatics: Discriminative Decoding for Referring Expression Generation". In: *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 47–52. URL: `https://aclanthology.org/2021.reinact-1.7`.
- Tanaka, M. et al. (2019). "Generating Easy-to-Understand Referring Expressions for Target Identifications". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5793–5802.

Citations VI

- Võ, Melissa Le-Hoa (Apr. 2021). "The meaning and structure of scenes". In: *Vision Research* 181, pp. 10–20. DOI: 10.1016/j.visres.2020.11.003.
- Yu, Licheng et al. (2016). "Modeling Context in Referring Expressions". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, pp. 69–85. ISBN: 978-3-319-46475-6.
- Zarrieß, Sina and David Schlangen (Nov. 2018). "Decoding Strategies for Neural Referring Expression Generation". In: *Proceedings of the 11th International Conference on Natural Language Generation*. Tilburg University, The Netherlands: Association for Computational Linguistics, pp. 503–512. DOI: 10.18653/v1/W18-6563. URL: https://www.aclweb.org/anthology/W18-6563.

Citations VII

- Zarrieß, Sina et al. (Aug. 2021). "Decoding, Fast and Slow: A Case Study on Balancing Trade-Offs in Incremental, Character-level Pragmatic Reasoning". In: *Proceedings of the 14th International Conference on Natural Language Generation*. Aberdeen, Scotland, UK: Association for Computational Linguistics, pp. 371–376. URL: https://aclanthology.org/2021.inlg-1.41.
- Zarrieß, Sina and David Schlangen (Aug. 2016). "Easy Things First: Installments Improve Referring Expression Generation for Objects in Photographs". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 610–620. DOI: 10.18653/v1/P16-1058. URL: https://www.aclweb.org/anthology/P16-1058.