Short Term Scientific Mission
iV&L COST Action IC1307

Logical words in Vision and Language

April 26, 2017

Practical details

- Applicant: Raffaella Bernardi, University of Trento (Italy)
- Host: Raquel Fernández, University of Amsterdam (The Netherlands), iV&L Net MC member
- Title: Visually-grounded logical words enter
- Period: 21st-28th of March 2017
- Attachment: Confirmation by the host institution of the successful execution of the STSM

1 Purpose of the STSM

This STSM has allowed the applicant (PhD awarded at Utrecht University in June 2002, currently Assistant Professor at the University of Trento) with expertise in Computational Semantics and Computer Vision to visit the Dialogue Modeling Group lead by Raquel Fernández. The research themes explored during the scientific mission are related to the iV&L Net working group "Integrated Modeling of Language and Vision" (WG1). We had planned to make progress towards the following research objectives:

1. To discuss current results on visually-grounded reasoning skills in Vision and Language Models.
2. To select reasoning skills linked to certain logical words that are visually-grounded and have been studied in detail in the fields of Computational Semantics and Dialogue.
3. To design empirical experiments with human participants to investigate literal and pragmatic intepretations of the selected visually-grounded logical words.

4. To design experiments to evaluate current state-of-the-art computational systems against human literal vs. pragmatic interpretaions of logical words.

2 Description of the work carried out during the STSM

During the first part of the visit, the applicant has been acquinted with the work carried out at ILLC on the interface between Language and Vision or on work that could be further investigated in this direction. She has met with Elliot Desmond, Dieuwke Hupkes, Marco del Tredici and other researchers.

Quantifiers and Vision The applicant has given a talk on the work on Quantifiers and Vision just posted on ArXiv [Sorodoc et al.,]. Interesting questions for future work have been addressed by people in the audience, like Willem Zuidema and Jakub Szymanik, Maria Aloni working on quantifiers from a cognitive and/or formal semantics perspective.

Scalar Adjectives The applicant has had daily meetings with Laura Aina and Raquel Fernández to start study scalar adjectives in language and vision. Some adjectives are claimed to live in a scale (e.g., "small, medium, big, huge"). Several studies have been carried out on how negation shifts their meaning within such scale. We would like to investigate whether the claims made in theoretical linguistics are met in the data and to bring into the discussion the role of visual features. To this end during the visit we have carried out a preliminary feasibility study. First of all, we have identified the relevant datasets among the available ones; secondly, we have created linguistic semantic spaces and started observing the behaviour of the adjectives and their negated counterparts in them. Finally, we have discussed possible models of negation and evaluation measures to attest the models' performance.

Interactive Visual Question Answering (IVQA) The applicant has had daily meetings with Elia Bruni and Raquel Fernàndez. We have identified a possible intersection between the work the two research groups have been carried out lately. In particular, the italian unit has recently published a work on finding one mistmatch between language caption and vision [Shanker et al., 2017] and shown that current State of the Art Language and Vision models fail in identify the mismatch. On the other hand, the dutch unit has been working on "Autonomous Learning Agents" using Generative Adversarial Networks (GAN). We conjectured that multimodal conversational agents could help tackling the FOIL task by learning to ask questions that lead to find the mismatch between the image and the caption. As a first step in this direction, we have carried out a first literature overview and identified the relevant papers [Das et al., 2017a, Das et al., 2017b, Huang et al., 2016, Mostafazadeh et al., 2016, Huang et al., 2017].

3 Description of the main results obtained

Scalar Adjectives The scale of adjectives seem to be reflected into linguistic spaces. The observed negated adjectives seem to live in a different region of the semantic spaces. This observation raises the question of whether within the new region the original scale is reversed or modified. Before addressing such a question however, we need to better explore the results obtain so far, in particular to verify whether frequency issues are involved in the current picture.

IVQA The literature overview we have carried out reveals the lack of studies, within the LaVi community, on "question generation". The only two papers we have found in this direction are [Mostafazadeh et al., 2016, Huang et al., 2017]. Much more is known in the Dialogue community regarding this human ability. Aiming to address the FOIL task with an IVQA approach might help bringing the LaVi community to address the need to focus on "question generation". As a start, within the identified task, we are currently discussing the role of "clarification questions".

4 Future collaboration with the host institution

Scalar Adjectives Laura Aina has chosen to continue working on scalar adjectives for her MSc Thesis under the joint supervision of Raffaella Bernardi and Raquel Fernández. We expect the integration of visual features to be left as future work after the thesis. The work of the thesis will however be crucial to then investigate the integration of language and vision in scalar adjectives.

Quantifiers and Vision The applicant has been invited to give a talk at the kick-off workshop of the ERC project Cognitive Semantic and Quantities to be held in Amsterdam on September 28th and 29th 2017 (PI: Jakub Szymanik). To start the collaboration on quantifier and vision, Sandro Pezzelle, PhD student supervised by Raffaella Bernardi at the Center of Mind and Brain, University of Trento, is planning to attend the aforementioned workshop and visit the ILLC again during the Fall 2017.

IVQA In the month of April 2017, we have had reading groups on IVQA via skype to help us better shape our project on this direction. To continue the collaboration on the IVQA work, Ravi Shekhar, PhD student supervised by Raffaella Bernardi at the Department of Information engineering and computer science (DISI), University of Trento, is planning to carry out a 3 month internship at the ILLC during the Fall 2017.

These planned further collaboration will help address the objectives 3 and 4 of the original plan that have been left unanswered during the current visit.

5 Foreseen publications/articles resulting from the STSM

We are aiming to publish the work on IVQA to one of the major conferences in either Computer Vision (e.g. CVPR 2018 – submission deadline November 2017) or Computational Linguistics (e.g. ACL 2018).

References

[Das et al., 2017a] Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. (2017a). Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Das et al., 2017b] Das, A., Kottur, S., Moura, J. M., Lee, S., and Batra, D. (2017b). Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*.

[Huang et al., 2017] Huang, J.-H., Alfadly, M., and Ghanem, B. (2017). Vqabq: Visual question answering by basic questions. arXiv:1703.06492.

[Huang et al., 2016] Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., and Mitchell, M. (2016). Visual storytelling.

[Mostafazadeh et al., 2016] Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., and Vanderwende, L. (2016). Generating natural questions about an image. In *Proceedings of ACL 2016*.

[Shanker et al., 2017] Shanker, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., and Bernardi, R. (2017). Foil it! find one mismatch between image and language caption. In *Proceedings of ACL 2017*.

[Sorodoc et al.,] Sorodoc, I., Pezzelle, S., Herbelot, A., Dimiccoli, M., and Bernardi, R. Pay attention to those sets! learning quantification from images. arXiv:1704.02923.