Report on Doctoral thesis: Multimodality in Machine Translation by Jindrich Libovicky

This thesis reports novel work in the area of multimodal neural machine translation. The manuscript starts with three chapters of background. Chapter 2 summarizes deep learning and its applications to computer vision and natural language processing. It describes the most commonly used architectures and their use in the two areas. Chapter 3 covers representative work combining vision and language, including grounded representations and image captioning. Chapter 4 narrows the background down to the topic of the thesis: multimodal machine translation. It provides a good (albeit very short) summary of existing architectures, as well as Multi30K, the dataset used by most existing work and in shared tasks on the topic. Alternative architectures to the problem are covered in Section 4.3.

The core contributions of the thesis are given in Chapters 5 and 6. Chapter 5 describes work done by the candidate in collaboration with colleagues, which revolve around three topics: a novel recurrent neural network model architecture and a novel self-attentive model architecture where multiple sources (including multiple modalities) are encoded and combined through a hierarchical attention mechanism, which is proposed as a way to combine two modalities (or more generally, sources). This constitutes solid work, as evidenced by an ACL publication (most competitive conference in natural language processing) and a strong system (and paper) in the WMT 2018 shared task on multimodal machine translation. The submission builds on the state-of-the-art approach, where multi-task learning is used to model the visual and textual modalities more independently (while still sharing parameters). This allowed the use of more data (text-only parallel data or text-image monolingual data), which significantly improved results. This architecture is also applied to automatic post-editing, with modest results. Finally, the thesis shows an interesting linguistic analysis which correlates several characteristics of sentences and images versus translation quality (measured automatically). This is a great addition to the thesis.

The thesis certainly offers enough novel scientific contributions to knowledge and proves the ability of the candidate for creative scientific work. While the manuscript is almost too concise (esp. Chapters 5-6), this does not compromise the quality of the report and – more generally – the work ignificantly. Corrections are not strictly necessary, but in what follows I list some recommendations/questions for clarification:

- Clarify individual contribution to wok done collaboratively, especially work described in Chapter 5. A footnote describing the role of the candidate in the paper published should suffice.
- Detail the process of corpus creation in section 4.2.2 – how was the inter-annotator agreement done? 1% sentences of 29,000? What is the agreement metric? Kappa?

- Chapter 5 claims that the proposed attention mechanism makes the model more interpretable. In which way? Is it because it allows inspecting the two attention mechanisms independently? How is this different from other previous work using two attention mechanisms?
- neuralmonkey could have been added as a contribution of the thesis. Why was it not?
- There's a very large difference in scores between tables 5.1 and 5.2 – is it all down to GRU vs LSTM? Please discuss this more extensively.
- It would be good to have more discussion on the results – in general they are just listed but not discussed. For example, Table 5.3: Czech does not seem to benefit from images, and the use of incongruent images does harm the results much. Why is this the case for this language pair but not the others?
- The experiments in Section 5.3.1 are not very clear: the number of selected sentences seems arbitrary and very different for different languages/datasets. The decisions could be better explained.
- In Chapter 6, the analysis could have been done with gold annotations for objects in Multi30K, instead of the automatic object detections. Was this not considered? What is the impact of errors in object detection in the analysis?

Lucia Specia

Professor of Natural Language Processing

Imperial College London