# Multi-modal Context Modelling for Machine Translation

**Lucia Specia**
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, S1 4DP
Sheffield, UK
l.specia@sheffield.ac.uk

## Abstract

MultiMT is an European Research Council Starting Grant whose aim is to devise data, methods and algorithms to exploit multi-modal information (images, audio, metadata) for context modelling in machine translation and other cross-lingual tasks. The project draws upon different research fields including natural language processing, computer vision, speech processing and machine learning.

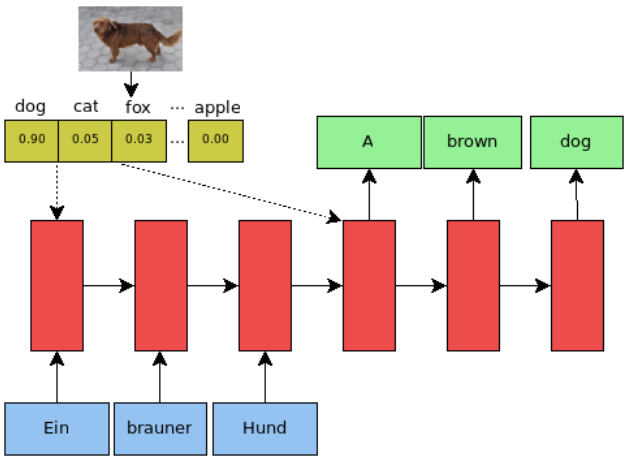
## 1 Description

Human translators have access to a number of contextual cues beyond the actual segment to translate when performing translation, for example, images associated with the text. Machine translation approaches, however, have historically disregarded any form of non-textual context and make little or no reference to wider surrounding textual content. This results in translations that miss relevant information or convey incorrect meaning. Such issues drastically affect reading comprehension and may render translations less useful. One example is the word 'seal' in the sentence 'The man is holding a seal'. When translating to German, this sentence can become 'Ein Mann hält ein Siegel' or 'Ein Mann hält einen Seehund'. Pictures such as the ones below could help in making this decision:



With an emphasis on images as additional modality and drawing parallels to work on image captioning, thus far the project has mainly targeted four main lines of research: data acquisition, representations, models and evaluation. For data acquisition, we have been following two approaches: (i) making multimodal data multilingual, where English image description datasets is extended to include translations of the descriptions in multiple languages, and (ii) making multilingual data multimodal, where parallel data is complemented by visual representations.

For representations, we have been exploiting high-level, abstract representations, such as the presence and frequency of objects in images, rather than relying on low-level, dense representations. We show that these representations are effective in both image captioning and machine translation. Our models are extensions of sequence-to-sequence neural models where different modalities can complement parallel text in different ways:




Project website: https://multimt.github.io/

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
*Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, p. 363
Alacant, Spain, May 2018.