Grounded Word Sense Translation

Chiraag Lala, Pranava Madhyastha and Lucia Specia
clala1@sheffield.ac.uk
github.com/sheffieldnlp/mlt
github.com/ImperialNLP/mltcode

Motivation

We hypothesise that images are useful to disambiguate ambiguous words in the source sentence to produce the correct translation

*"A **sportsperson** is playing football"*

*"Une **sportive** joue au football"*

*"Une **sportif** joue au football"*

The Task and its Dataset

From the Multi30K dataset we identified words in the source language (En) with multiple translations in the target languages (De, Fr) with different meanings/senses

Task: Given an image and its description, identify the ambiguous words and tag these to the correct sense translations

People walking down a trail in the woods

French labels/tags: *sentier* *forêt*

Skewed Distribution over Translation Candidates

EnDe: 4.1 Translation candidates per ambiguous word (TCPA) and the Most Frequent Translation (MFT) occurs in 65% of the samples

EnFr: 3.0 TCPA and the MFT occurs in 77% of the samples

Human Experiment

We asked humans to perform the task by manually labelling the 2018 test set of the WMT Multimodal Machine Translation shared task. The annotators find image useful in only 7.8% of the samples for EnDe, and 8.6% for EnFr

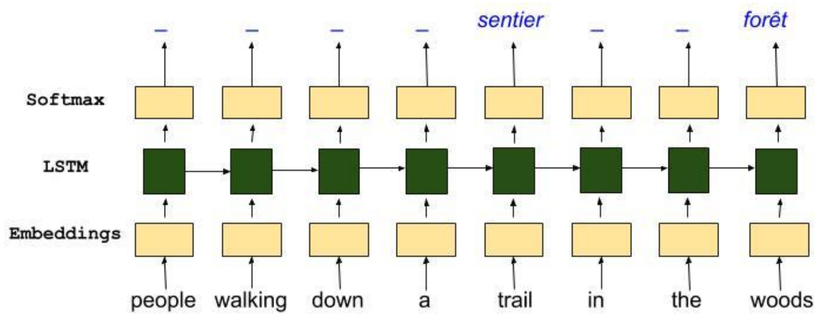
Words like *player*, *hat* and *coat* require the image as text alone is not sufficient to disambiguate

Computational Models, Data Settings and Results

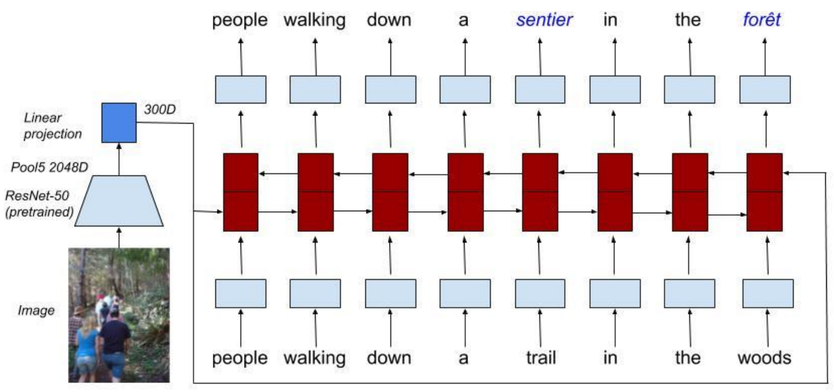
ULSTM benefits from the ResNet-50 `pool5` global image features (+image) as compared to BLSTM, especially in the 'ambiguous words' data setting where only ambiguous words are tagged

The 'ambiguous sentences' and 'all words' data setting is most suitable for word sense translation

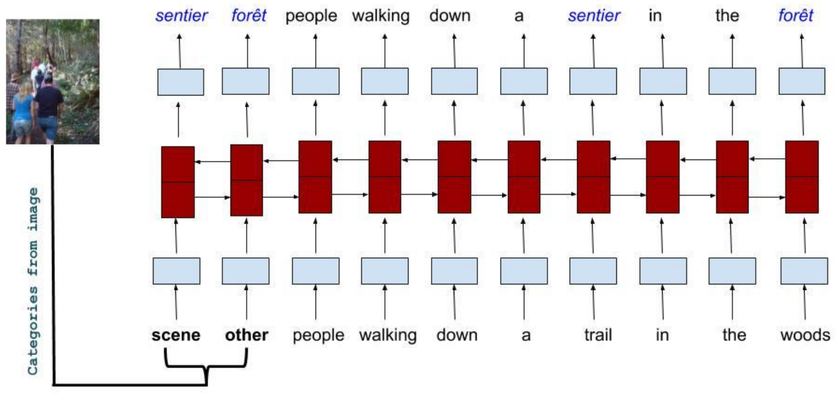
Architectures	EnDe	Δ	EnFr	Δ
Random	24.4	-	33.6	-
MFT	65.34	-	77.73	-
all sentences + ambiguous words				
ULSTM	63.99	-	73.65	
ULSTM+image	66.10	**2.11**	75.58	**1.93**
BLSTM	67.56	-	76.89	
BLSTM+image	**68.44**	0.88	**77.66**	0.77
ambiguous sentences + ambiguous words				
ULSTM	63.58	-	74.42	
ULSTM+image	66.33	**2.75**	76.89	**2.47**
BLSTM	68.15	-	78.58	
BLSTM+image	**68.62**	0.47	**79.12**	0.54
all sentences + all words				
ULSTM	66.63	-	76.50	
ULSTM+image	66.86	**0.23**	77.12	**0.62**
BLSTM	**69.03**	-	78.35	
BLSTM+image	68.74	-0.29	**78.97**	**0.62**
ambiguous sentences + all words				
ULSTM	67.27	-	78.20	
ULSTM+image	67.56	**0.29**	78.27	0.07
BLSTM	69.61	-	80.35	
BLSTM+images	**69.79**	0.18	**80.43**	**0.08**



Uni-directional Long Short-Term Memory network (ULSTM) with 'ambiguous words' data setting



Bi-directional LSTM with hidden states initialized with the ResNet-50 `pool5` global image features (BLSTM+image) with the 'all words' data setting



Sixteen object categories were extracted from the images and aligned to the words **(Oracle)**. Those corresponding to ambiguous words were prepended to the source sentence and then a BLSTM model was trained (BLSTM+object-prepend)

BLSTM models trained by pre-pending object categories outperform all the other models

Architectures	EnDe	Δ	EnFr	Δ
Random	24.4	-	33.6	-
MFT	65.34	-	77.73	-
all sentences + ambiguous words				
ImageOnly	67.56	-	77.20	
ObjectOnly	68.33	-	78.89	
BLSTM	67.56	-	76.89	
BLSTM+image	68.44	0.88	77.66	0.77
BLSTM+object	67.80	0.24	79.28	2.39
BLSTM+object-prepend	**70.08**	**2.52**	**80.89**	**4.00**
ambiguous sentences + ambiguous words				
ImageOnly	67.92	-	78.35	
ObjectOnly	68.15	-	79.74	
BLSTM	68.15	-	78.58	
BLSTM+image	68.62	0.47	79.12	0.54
BLSTM+object	69.03	0.88	79.43	0.85
BLSTM+object-prepend	**70.44**	**2.29**	**80.20**	**1.62**
all sentences + all words				
ImageOnly	67.56	-	77.20	
ObjectOnly	68.33	-	78.89	
BLSTM	69.03	-	78.35	
BLSTM+image	68.74	-0.29	78.97	0.62
BLSTM+object	69.85	0.82	79.89	1.54
BLSTM+object-prepend	**70.90**	**1.87**	**81.97**	**3.62**
ambiguous sentences + all words				
ImageOnly	67.92	-	78.35	
ObjectOnly	68.15	-	79.74	
BLSTM	69.61	-	80.35	
BLSTM+images	69.79	0.18	80.43	0.08
BLSTM+object	69.79	0.18	81.28	0.93
BLSTM+object-prepend	**71.02**	**1.41**	**82.59**	**2.24**