Grounded Word Sense Translation

Chiraag Lala
University of Sheffield
clala1@sheffield.ac.uk

Pranava Madhyastha
Imperial College London
pranava@imperial.ac.uk

Lucia Specia
Imperial College London
l.specia@imperial.ac.uk

Abstract

Recent work on visually grounded language learning has focused on broader applications of grounded representations, such as visual question answering and multimodal machine translation. In this paper we consider grounded word sense translation, i.e. the task of correctly translating an ambiguous source word given the corresponding textual and visual context. Our main objective is to investigate the extent to which images help improve word-level (lexical) translation quality. We do so by first studying the dataset for this task to understand the scope and challenges of the task. We then explore different data settings, image features, and ways of grounding to investigate the gain from using images in each of the combinations. We find that grounding on the image is specially beneficial in weaker unidirectional recurrent translation models. We observe that adding structured image information leads to stronger gains in lexical translation accuracy.

1 Introduction

The multimodal machine translation (MMT) shared task has been conducted for the past three years (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018) with the main goal of investigating the effectiveness of information from images in machine translation (MT). However, as acknowledged in Barrault et al. (2018), it has been difficult to evaluate the impact of multimodality (images) on the sentence-level translation quality, since the changes incurred by having an additional modality can be quite subtle. The MMT shared task consists of translating English sentences that describe an image into a target language given the English sentence itself and the image that it describes.

Recently proposed, the multimodal lexical translation (MLT) (Lala and Specia, 2018) is a similar task but focused at the word level and only at ambiguous words. In MLT, the objective is to correctly translate each ambiguous word in the English source sentence into a corresponding word in the target language given the word itself, the English sentence in which it occurs and the image being described by that sentence. This is similar to the task of Visual Sense Disambiguation (Gella et al., 2016) where the objective is to disambiguate the ambiguous verbs using text and image contexts. The authors of MLT proposed to define a word in the source language to be ambiguous if it has multiple translations in the target language with different meanings in the dataset. However, they did not suggest any models for that.

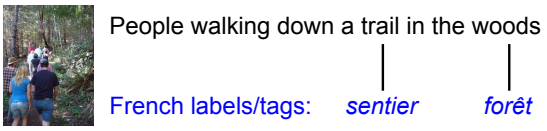

Figure 1: A labeled example from the dataset for multimodal lexical translation. Only ambiguous words in the sentence are labeled to their corresponding translation in the target language.

In this paper, we propose to treat MLT as a sequence labeling task, as depicted by the example in Figure 1, similar to part-of-speech tagging or named entity recognition. Our approach draws inspiration from neural sequence-based approaches to word sense disambiguation (Raganato et al., 2017; Yuan et al., 2016; Kågebäck and Salomonsson, 2016) and approaches to ground machine translation (Caglayan et al., 2017). More specifically, we propose and empirically evaluate grounded translation disambiguation models based on recurrent sequential units for the task of MLT. Our primary contributions are:

- An investigation of the MLT dataset to understand the scope and challenges of the task:

Proceedings of the Second Workshop on Shortcomings in Vision and Language, pages 78–85
Minneapolis, USA, June 6, 2019.

	Train	Val	Test
Sentences	29,000	1,014	1,000
Labels EnDe	49,626	1,775	1,708
Labels EnFr	41,191	1,427	1,298

Table 1: Data splits of the dataset for multimodal lexical translation, where EnDe indicates English-German, and EnFr, English-French.

we find the task is challenging because of the skewed distribution of translation candidates in the training set and that the scope of improvements from images is about 7.8% for English-German and 8.6% for English-French.

- An investigation into data settings for the task: we find that models trained to tag all words, irrespective of their ambiguity level, perform better than other settings.
- A study on the effect of visual representations for grounded recurrent models: we find that simple unidirectional recurrent models gain more with conditioning of visual information than stronger bidirectional recurrent models.
- An investigation on different visual representations for the task: we find that structured image information (in the form of objects) perform better than the popularly used ResNet `pool5` image features.

2 Dataset for MLT

Lala and Specia (2018) extract the MLT dataset from the Multi30K (Elliott et al., 2016, 2017). MLT was also used to compute Lexical Translation Accuracy for systems submitted to the WMT18 multimodal translation shared task (Barrault et al., 2018).

The dataset consists of 31,014 images with one English description per image, where the ambiguous words in the description, if any, are labeled to their corresponding lexical translations in the target language conforming to the given context (see Figure 1). The dataset is split into training, validation and test sets in the same way as in the WMT's MMT task in 2016 (see Table 1).

2.1 Skewed Distributions of Translations

Statistics about the dataset for MLT are shown in Table 2. We emphasize that a key aspect of the dataset worth noting is the skewed distribution over the lexical translation candidates. For instance, the English word *woods* has two possible lexical translations in French in this dataset - *forêt* and *bois*. Ideally, we would want both these lexical translations to occur equal number of times (uniform distribution) but in reality the distribution is skewed - *bois* occurs 79 times (we call it the Most Frequent Translation (MFT)) while *forêt* occurs 16 times.

Language Pair	UA	APS	APHW	TCPA	SR	WSR
EnDe	745	1.68	15.0	4.1	1.8	1.5
EnFR	661	1.39	12.5	3.0	1.6	1.3

Table 2: Some key statistics of the original dataset for MLT. UA: Unique Ambiguous words. APS: Ambiguous words Per Sentence. APHW: Ambiguous words Per Hundred Words. TCPA: Translation Candidates Per Ambiguous word. SR: Skewness Ratio as described in Section 2.1. WSR: Weigthed average of SRs.

For a better understanding of the skewness of the distributions, we define a Skewness Ratio (SR) of a word as the ratio of count of the word to the count of its most frequent translation. For example, SR(*woods*) = count(*woods*)/count(*bois*) = 1.2. For the whole dataset, we simply average the SRs over all the ambiguous words[1]. The averaged SR will be a number between 1 and the TCPA (the averaged Translation Candidates Per Ambiguous word). If it is closer to 1 this means that, in the dataset, the distribution over lexical translations is skewed. If it is closer to TCPA, then the distribution is more uniform.

We note, our definition of Skewness Ratio is similar to the inverse of 'Average Time-anchored Relative Frequency of Usage' metric defined in Ilievski et al. (2016) which is used to assess potential bias of meaning dominance with respect to its temporal popularity.

The averaged Skewness Ratios for both language pairs, mentioned in Table 2, are much closer to 1 than to their corresponding TCPAs. This implies that the distributions over the translations are highly skewed and suggests that it will be extremely challenging to demonstrate improvements over the MFT because of bias to MFT as indicated in Postma et al. (2016).

[1]We also compute the weighted average of SRs, called WSR in table 2, weighted by the frequency of the ambiguous word in the corpus

2.2 When Humans Find Images Useful

We extended the dataset for MLT to include the 2018 test set of MMT shared task by manually labeling the examples. In the process, human annotators were further instructed to inform whenever the image was useful in performing lexical translation.

2.2.1 Setup

The 2018 test set of the MMT shared task was made available, consisting of 1071 images and one English description per image. The ambiguous words from the original MLT dataset were searched in this test set using string matching to identify ambiguous test instances. From these test instances, the English description together with the ambiguous word and the set of all lexical translation candidates of the ambiguous word were provided to human annotators who are bilingual speakers of both English and the target language (German or French) under consideration. The corresponding images were also provided but not explicitly shown to the annotators; they had the option to look at the image if they have to and specify when they used the image.

The objective for the annotators was to select those translation candidates they thought conformed both the English description and the corresponding image; or in other words, they had to filter out the translation candidates that did not conform either the English description or the image, while having the option to look at the image (if they thought the visual context was needed to make a decision) or ignore it completely (if they thought the visual context was not needed). If they selected all available options (i.e. they did not filter out any single option) then those examples were removed from the study.

2.2.2 Results and Discussion

The human annotations of this experiment can be found together with the MLT dataset on https://github.com/sheffieldnlp/mlt. The results are shown in Table 3 and discussed below.

For English-German, the extension consists of 358 instances of ambiguous words. In 111 (or 31%) of these instances the annotators opted to look at the image. In 83 of these 111 image-aware instances the annotator selected the lexical translation candidate which happened to be the most frequent translation. The annotators did not know which translation candidate was the most frequent for the given ambiguous word in the corpus. This leaves us with 28 instances, which is 7.8% of all the instances, where the annotators looked at the image and chose to filter out the most frequent translation. Although the sample size is small, these numbers help us understand the scope of image at word-level translation task (7.8% for EnDe and 8.6% for EnFr; i.e. around 8% on average).

Language Pair	Ins	Img	Img-MFT	Img-MFT / Ins (Scope)
EnDe	358	111	28	7.8%
EnFr	407	72	35	8.6%

Table 3: Results of the Human Experiment. Ins: Instances with ambiguous words. Img: the Ins instances where the Image was used. Img-MFT: the Img instances where the Most Frequent Translation was not selected (filtered out) by the annotators. Img-MFT / Ins (Scope): the ratio of Img-MFT to Ins expressed in percentage; and as discussed in Section 2.2.2 this reflects the Scope of improvement at Lexical Translation using Images.

Ambiguous words where humans opted to look at the image include *pool, hat, coat, field, wall, etc.*, suggesting textual context is not sufficient for such words. Ambiguous words where humans ignored the image include *area, fall, watch, walk, etc.*, suggesting the textual context is often sufficient to identify the correct translation.

3 Lexical Translation Models

We explore two neural sequence labeling architectures following Graves (2012), using long short-term memory networks (LSTMs)[2]:

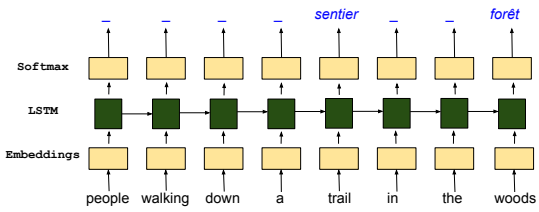

Figure 2: Unidirectional long short-term memory network used as a tagger for lexical translation of ambiguous words. The remaining unambiguous words are tagged to a common label (an underscore '_' in this case).

[2]We also experimented with sequence-to-sequence approaches (Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015) and their application to word sense disambiguation by Raganato et al. (2017), but these performed worse.

ULSTM: This is a single layer unidirectional LSTM network (Hochreiter and Schmidhuber, 1997). A similar setting is used in Yuan et al. (2016) as a classifier for word sense disambiguation. In our setup we use the LSTM as a tagger (see Figure 2).

BLSTM: This is a single layer bidirectional LSTM network (Graves and Schmidhuber, 2005) used as a tagger. BLSTMs are used in (Kågebäck and Salomonsson, 2016) as a classifier for word sense disambiguation and have shown promising results. Recent work also suggests that BLSTM-based tagging models give state of the art performance on multilingual sequence tagging (Plank et al., 2016).

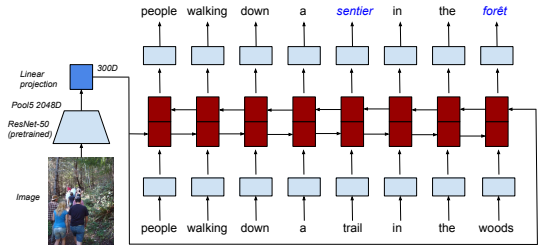

Figure 3: Multimodal-BLSTM for lexical translation of ambiguous words. Unambiguous words are tagged to self.

We extend these architectures to make them multimodal, as follows:

Multimodal Tagger: Following previous work in grounded machine translation and image captioning (Caglayan et al., 2017; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015), we propose multimodal models that are identical to the text-only ULSTM and BLSTM models but are conditioned with image information. Specifically, the hidden states of the LSTMs are initialized with the image features. We used the ResNet-50 (He et al., 2016) based image features and extract 2048-dimensional features extracted from the `pool5` layer of a pre-trained ResNet-50 model. To match the dimensions of the hidden states of the LSTM, we learn a linear projection. A multimodal BLSTM architecture, trained on a data setting where we also label the unambiguous words to itself, is depicted in Figure 2.

Object-based Grounding: Given that the ambiguities are associated with content words, we assume that these correspond to objects and propose a model that uses objects in the image associated to the ambiguous words. We experiment with two ways of incorporating object information - a) Initializing and b) Prepending.

The **Initializing** approach is identical to the multimodal tagger above where instead of the 2048-dimensional ResNet-50 image features we initialize the ULSTM and BLSTM with a binary vector representing the presence or absence of objects in the image corresponding to its ambiguous words. In the **Prepending** approach, motivated by recent work in neural machine translation (Johnson et al., 2017), we prepend the word that represents the object category (e.g. 'person') associated with the ambiguous word to the source sentence.

We extract object category information from the images using annotations on Plummer et al. (2015). These consist of a set of 16 object categories that abstractly depict the objects present in the image.

3.1 Data Settings

A significant proportion of sentences in the training (16% for EnDe and 21% for EnFr) dataset do not have any ambiguous word. Therefore at training time we experiment in two ways a) to ignore such sentences ('**ambiguous sentences**' setting); or b) train on all sentences ('**all sentences**' setting). Secondly, for unambiguous words (i.e. tokens that are not labelled), we experiment in two settings – a) leave it unlabelled ('**ambiguous word**' setting) or b) to label it to itself ('**all words**' setting). These choices amount to four different data settings for training.

3.2 Training and Baselines

Training and Evaluation: For optimization, we use the Adam (Kingma and Ba, 2014) algorithm with a learning rate = 0.001 and batch size = 32. The LSTM hidden state dimensions and the word embedding dimensions are set to 300 and the dropout rate is set to 0.3. Training is stopped early if model accuracy over the validation set does not improve for 30 epochs and then the best performing model over the validation set is selected. These models are implemented and trained in the TensorFlow framework.

As the focus of the task is on translating ambiguous words only, we measure the performance of all the models in terms of accuracy of correctly translating ambiguous words, ignoring the label-

ing accuracy on other words[3]. We also measure gains from the image, i.e. the difference (Δ) between the performance of multimodal and corresponding text-only baseline models.

Frequency Baselines: We consider baselines that completely disregard the visual and the textual contexts. The Random baseline translates an ambiguous word by selecting a translation candidate at random. The MFT baseline selects the most frequent translation of the ambiguous word as seen in the training data. As noted earlier, the most frequent translation is expected to be difficult to outperform because of the skewed distribution of translation candidates in the dataset (Postma et al., 2016).

Text-only and Image-only Baselines: The text-only baselines are the ULSTM and BLSTM that do not consider the visual contexts. The image-only baselines are the multimodal tagger conditioned on the image (either image features or object vector) except that they do not read textual context but only the ambiguous words in the sentence, i.e. all unambiguous words are removed.

4 Results and Discussion

Results of the two text-only (ULSTM and BLSTM) and two multimodal models (ULSTM+image and BLSTM+image) in the four different data settings on the test set are shown in Table 4.

We observe that all models perform better than Random baseline and most models perform better than MFT. We see that the BLSTM models always perform better than the corresponding ULSTM models, as expected.

With ResNet-50 `pool5` global image features, the multimodal ULSTM+image models perform better than the corresponding text-only ULSTM models in all data settings (See Table 4). This shows ULSTM models benefit from the ResNet-50 image features. The same cannot be said for BLSTM. Also, ULSTM tends to gain more from the image as compared to the BLSTM. We posit the lack of sufficient contextual information in ULSTMs as the reason. The visual information seems to compensate for the incomplete textual context. We provide examples in Figure 4.

[3]As a sanity check we note that, for all the models we experimented with, the labeling/tagging accuracy on all words (both ambiguous and unambiguous combined) ranges between 85% and 94% on the validation set and 85% and 91% on the test set.

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**

Table 4: Comparing multimodal models with their text-only counterparts in different data settings. We observe ULSTM benefits more from the ResNet-50 global image feature as compared to BLSTM.

Figure 4: Examples showing ULSTM tends to benefit more from the ResNet-50 `pool5` image features as compared to BLSTM.

Further, we observe that models perform better in **all words** data settings compared to **ambiguous words** setting. This is surprising for sequence

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**

Table 5: Comparing object-based grounding BLSTM models with other BLSTM models in different data settings.

labeling since in such data settings the number of labels are larger than the source language vocabulary. Nevertheless, we observe that this data setting outperforms others. We hypothesize that a possible reason is that it forces the models to capture better context. We also note that the gains Δ from the image are larger in the **ambiguous words** data setting, especially for ULSTM. This suggests that the image information assists the model to learn better context representations. Models tend to perform slightly better in the **ambiguous sentences** setting as compared to **all sentences**. This hints that more data is not necessarily better as the unambiguous sentences are not always relevant to the task. This is in line with observations in Postma et al. (2016).

Results of our proposed object-based structured grounding models (BLSTM+object and BLSTM+objct-prepend) together with other BLSTM models are shown in Table 5. The object-based structured grounding models outperform the multimodal models that use ResNet-50 image features in most cases. More specifically, grounding via prepending performs the best in all data settings with gains over the corresponding text-only baselines ranging from 1.41% to 2.52% for EnDe and 1.62% to 4.00% for EnFr across different data settings. The best multimodal model is BLSTM+object-prepend trained in the **ambiguous sentences** and **all words** data settings and it outperforms the best performing text-only baseline model by 1.41% for EnDe and 2.24% for EnFr. This suggests that region-specific information in terms of explicit objects corresponding to the ambiguous words in the sentences are highly beneficial. We observe a similar trend when comparing the ObjectOnly baseline vs ImageOnly baseline, i.e. object information is better than ResNet-50 global image features in absence of textual context too.

5 Conclusions

We studied the MLT dataset and found that the distribution of translation candidates is very skewed making the word-level translation task challenging. In a human study, we found the scope of improvement gains from images is about 7.8% for EnDe and 8.6% for EnFr in this task on this dataset. We proposed grounded models for the task of word-level translation. We found the 'ambiguous sentences' and 'all words' data setting is most suitable for the task. Also, we found the ULSTM tends to benefit more from the image as compared to the BLSTM and posit that this is because the image compensates for the weak textual information for the ULSTM. We found that object-based grounded models, i.e. models that have explicit information about the objects associated with the ambiguities, outperform other models including ones which use the popularly used ResNet-50 `pool5` global image features. Also, we found that grounding by prepending performs better than initializing.

Acknowledgements

We thank Josiah Wang for his comments. This is supported by MultiMT (H2020 ERC Starting Grant No. 678017) and MMVC (Newton Fund Institutional Links Grant, ID: 3523433575) projects.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Alex Graves. 2012. Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks*. Springer.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. In *Proceedings of Neural Networks*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. In *Proceedings of Neural Computation*.

Filip Ilievski, Marten Postma, and Piek Vossen. 2016. Semantic overfitting: what 'world' do we consider when evaluating disambiguation of text? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Googles multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5.

Mikael Kågebäck and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional lstm. In *Proceedings of International Conference on Computational Linguistics*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations*.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *The 54th Annual Meeting of the Association for Computational Linguistics*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*.

Marten Postma, Ruben Izquierdo Bevia, and Piek Vossen. 2016. More is not always better: balancing sense distributions for all-words word sense disambiguation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*.