Sheffield Submissions for WMT18 Multimodal Translation Shared Task

Chiraag Lala, Pranava Madhyastha, Carolina Scarton and **Lucia Specia**
Department of Computer Science, University of Sheffield, UK
{clala1, p.madhyastha, c.scarton, l.specia}@sheffield.ac.uk

Abstract

This paper describes the University of Sheffield's submissions to the WMT18 Multimodal Machine Translation shared task. We participated in both tasks 1 and 1b. For task 1, we build on a standard sequence to sequence attention-based neural machine translation system (NMT) and investigate the utility of multimodal re-ranking approaches. More specifically, n-best translation candidates from this system are re-ranked using novel multimodal cross-lingual word sense disambiguation models. For task 1b, we explore three approaches: (i) re-ranking based on cross-lingual word sense disambiguation (as for task 1), (ii) re-ranking based on consensus of NMT n-best lists from German-Czech, French-Czech and English-Czech systems, and (iii) data augmentation by generating English source data through machine translation from French to English and from German to English followed by hypothesis selection using a multimodal-reranker.

1 Introduction

This paper describes the University of Sheffield's submissions for both Tasks 1 and 1b of the third edition of the Multimodal Machine Translation shared task. Task 1 consists in translating source sentences in English that describe an image into German (DE) or French (FR) or Czech (CS), given the image. Task 1b consists in translating source sentences in English that describe an image into Czech, given the image and the French and German translations of the source sentence.

This task poses the challenging problem of building models that use both language and image modalities. The dataset for the shared task (Specia et al., 2016) has sentences with simple language constructions and it has been observed by earlier systems (Specia et al., 2016; Elliott et al., 2017) that standard text-only sequence to sequence neural machine translation models (NMT) with attention are able to obtain very high performance.

Building on this, for further inspection, we built our own standard NMT systems for EN-DE, EN-FR and EN-CS language directions and noticed that the translation hypotheses besides the 1-best output are also of high quality. We made our systems produce 20 translation hypotheses for English descriptions in the validation set and selected the hypothesis with the highest sentence-level METEOR (Denkowski and Lavie, 2014) score, called the Oracle, and compared this to the 1-best. In this experiment, we observed that the Oracle performs way better (11 to 13.5 METEOR points) than the 1-best output (See Table 1). This preliminary experiment motivated us to investigate re-ranking approaches.

Lang-Pair	1-best	Best of 20best (Oracle)	Scope/difference (Oracle - 1-best)
EN-DE	48.36	61.85	**+13.49**
EN-FR	64.91	76.87	**+11.96**
EN-CS	33.87	44.71	**+10.84**

Table 1: Motivation for re-ranking. In this preliminary experiment, we observe that re-ranking of the 20-best translation hypotheses generated by a standard NMT model has the potential of improving translation by upto 10.84 to 13.49 METEOR points for the three language pairs.

For a re-ranking strategy, we were inspired by how humans use images to translate image descriptions. We believe humans look at the image usually to disambiguate ambiguous words in the source sentence especially in those instances where the text alone is not sufficient. For example, translating '*A **sportsperson** is playing football*' into French requires us to know whether the sportsperson is a male or a female and accordingly

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 624–631
Belgium, Brussels, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64069

the translation is '*Une* ***sportif*** *joue au football*' (male) or '*Une* ***sportive*** *joue au football*' (female). In such cases, humans usually look at the image to disambiguate and select the correct translation which is what we try to mimic in our approach.

More specifically, in our systems we adopt a two-step pipeline approach. In the first step, we use an ensemble of text-only models initialized with different seeds to produce lists of 10-best translation hypotheses. In the second step, we re-rank the 10-best hypotheses using a novel multimodal cross-lingual Word Sense Disambiguation (WSD) approach. For control experiments, we also compare our results with monomodal cross-lingual WSD (Lefever and Hoste, 2013) and a system that performs re-ranking using the Most Frequent Sense (MFS) baseline (Section 3.1.2).

Our main goal is to investigate a multimodal, image-based, cross-lingual WSD that predicts the translation candidate which correctly disambiguates ambiguous words in the source sentence.

Our baseline NMT system is based on the attentive encoder-decoder (Bahdanau et al., 2015) approach with a Conditional GRU (CGRU) (Cho et al., 2014) decoder and is built using NMTPY toolkit (Caglayan et al., 2017b).

Our cross-lingual WSD models are based on neural sequence learning models for WSD (Raganato et al., 2017; Yuan et al., 2016; Kågebäck and Salomonsson, 2016) applied to the Multimodal Lexical Translation Dataset (Lala and Specia, 2018).

For task 1b, we explore three approaches. The first approach concatenates the 10-best translation hypotheses from DE-CS, FR-CS and EN-CS MT systems and then re-ranks them using the *image-aware* multimodal cross-lingual WSD mentioned earlier (the same way as in Task 1) (Section 3.1.2).

The second approach explores the consensus between the different 10-best lists. The best hypothesis is selected according to the number of times it appeared in the different 10-best lists. We followed the order of the n-best lists, meaning that the highest ranked hypothesis with the majority votes was selected.

The third approach uses data augmentation that hinges on the fact that the objective is to translate from English into Czech. Extra source data is generated by building systems that translate from German into English and French into English. With this extra data, we build an EN-CS system. We then obtain a 10-best list over training, development and test sets respectively. For selecting the best hypothesis from the 10-best list, we experiment with a classification-based approach. We calculate METEOR (Denkowski and Lavie, 2014) scores for each hypothesis in the 10-best list of the training set and threshold the scores to build classifiers to distinguish good from bad translations using a) word embeddings and image features with a Random Forest model and b) a multimodal Recurrent Neural Network (RNN) model.

In Section 3 we describe our systems in detail. We describe the data preprocessing in Section 2. The results are discussed in Section 4.

2 Data

2.1 Translation models

We use the Multi30K (Elliott et al., 2016) dataset provided by the organizers. Each image i contains one English description en_i taken from Flickr30K and human translations into German de_i, French fr_i and Czech cz_i. In other words, each instance is a 5-tuple of the form $(i, en_i, de_i, fr_i, cz_i)$. The dataset contains 29,000 training and 1,014 development instances.

For Task 1, the test sets of the previous two editions (2016 and 2017) have also been provided for validation purposes. These do not contain Czech translations. A new test set of 1,071 tuples containing an English description and its corresponding image is provided for evaluation.

For Task 1b, a test set of 1,000 tuples containing English, French, and German descriptions and their corresponding images is provided for evaluation. This test set corresponds to the unseen portion of the Czech Test 2017 data. The test set of 2016 is provided for validation purposes.

2.2 Cross-lingual WSD models

For the cross-lingual WSD models, we use the Multimodal Lexical Translation Dataset (MLTD) (Lala and Specia, 2018), which was extracted from the Multi30K (Elliott et al., 2016) dataset. MLTD consists of 4-tuples of the form (x, i, en_i, x_t) where x is an ambiguous[1] word in the English description en_i of the image i, and x_t is the lexical translation of x in a specified target language $t \in$

[1] We use the term 'ambiguous' for those words in the source language that have multiple translations in the target language in the training portion of the given parallel corpus, where these translations represent different 'senses' of the word in that corpus.

{German, French, Czech} that conforms with the image and the description. Only instances from the training portion of the Multi30K dataset are used to train the cross-lingual WSD models.

For English-German, MLTD consists of 745 ambiguous words in English with 4.09 different translations per word (on average) in German and 17.69 instances per translation (on average) totalling 53,868 MLTD instances.

For English-French, MLTD consists of 661 ambiguous words in English with 2.98 different translations per word (on average) in French and 22.73 instances per translation (on average) totalling 44,779 MLTD instances.

For English-Czech[2], MLTD consists of 3,217 ambiguous words in English with 5.15 different translations per word (on average) in Czech and 11.32 instances per translation (on average) totalling 187,495 MLTD instances.

2.3 Image features

We used the ResNet-50 image features provided by the task organizers. These are 2048-dimensional features extracted from *pool5* of a pretrained ResNet-50 (He et al., 2016) model which has been trained on the ImageNet dataset (Russakovsky et al., 2015).

3 System descriptions

In this section we describe the systems submitted for both tasks.

3.1 Task 1 systems

Our two-step pipeline consists in first obtaining high quality hypotheses from a NMT model, followed by a re-ranking step. We describe the setup of the NMT in Section 3.1.1. The cross-lingual WSD models used for re-ranking are described in Section 3.1.2 and the re-ranking formulation with examples is shown in Section 3.1.3.

3.1.1 Baseline NMT model setup

We make use of an ensemble of text only attention based NMT models (Bahdanau et al., 2015) with a conditional gated recurrent units (CGRU) (Cho et al., 2014) decoder. We build the system using the NMTPY toolkit (Caglayan et al., 2017b). Our models have a setting similar to Caglayan et al. (2016) with a bi-directional 256-dimensional recurrent GRU followed by a conditional GRU which is initialized with a non-linear transformation of the mean of encoder states. We use a simple feedforward network to compute the attention scores as described in Caglayan et al. (2016). We use Adam optimizer with a learning rate of $5e^{-}5$ and a batch size of 64. We set the embedding dimensionality of encoder and decoder to 128 and follow the default parametrization in (Caglayan et al., 2017a). Our final baseline model is an ensemble of different runs of the model with five different seeds.

3.1.2 Crosslingual WSD models

The goal of cross-lingual WSD (Lefever and Hoste, 2013) is to generate contextually correct translations of ambiguous words in the source language into the target language. For this, the sense inventory for the ambiguous words is created from the parallel corpus. MLTD (Lala and Specia, 2018) (Section 2.2) provides us with the data settings needed for this task.

As a baseline we have the **Most Frequent Sense** (MFS) model, which returns the most frequent translation of a given ambiguous word as seen in the training corpus. For example in the English-French MLTD, the ambiguous word *woods* appears 95 times in the training set. In 16 times the translation is *forêt* (forest), while in the remaining 79 times the translation is *bois* (timber/wood). In this case, the MFS model translates the word *woods* as *bois* irrespective of the context.

As a second baseline, we have a text-only **Lexical Translation** (LT) model. This is a single layer Bidirectional Long Short-Term Memory (BiLSTM) network (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) used as a sequence tagger as depicted in Figure 1.

For the LT model, we convert the classification task of cross-lingual WSD into a sequence tagging task as demonstrated in (Raganato et al., 2017). The 4-tuples of MLTD are transformed into a sequential tagged dataset. This consists of English sentences where each word is tagged to itself if it is unambiguous and tagged to the correct lexical translation in the target language if it is ambiguous.[3]

[2] This dataset has been extracted using the same procedure in Lala and Specia (2018) except the human filtering step and thus it contains noise: mainly, the multiple "senses" can sometimes correspond to morphological variants or synonym words.

[3] We tried a few more data settings - like each word tagged to 'NA' if it is unambiguous - but these did not result in any improvements.

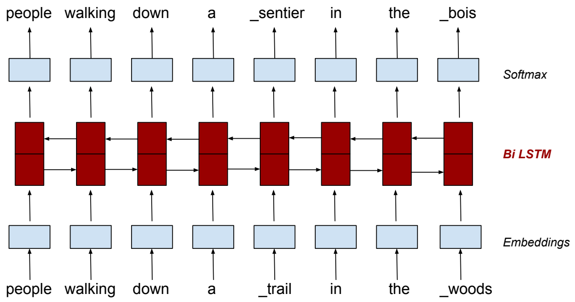

Figure 1: Lexical Translation (LT) model - A BiLSTM that tags each input word in the source sentence. The training is done such that an unambiguous word is tagged with itself, while an ambiguous word, like *trail* and *woods* in this example, is tagged with the corresponding lexical translation in the target language like *sentier* and *bois* respectively.

Our proposed model is a **Multimodal Lexical Translation** (MLT) model. It has the same architecture as the LT model except that the LSTM weights are initialized with the image features. [4] To avoid dimensionality mismatch, the image features (Section 2.3) undergo a dimensionality reduction via a fully connected layer, which is also trained.

Training: Both LT and MLT models are trained on only those sentences which have at least one ambiguous word as per MLTD. 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. These models are implemented and trained in the TensorFlow framework.

The performance of the models (Table 2)[5], measured in terms of percentage of correctly translated ambiguous words (accuracy), suggests that the image-aware MLT model is slightly better than the text-only LT and MFS models.

	test17flickr	test17coco	test16	train	val
EN-DE					
MFS	60.47	52.49	65.34	68.93	70.25
LT	**61.40**	57.22	69.61	79.71	67.77
MLT	59.68	**57.48**	**69.79**	**80.18**	**68.85**
EN-FR					
MFS	**77.29**	67.12	77.73	78.38	79.33
LT	76.83	70.52	80.35	88.05	**81.15**
MLT	75.20	**70.75**	**80.43**	**88.44**	80.87

Table 2: Performance of cross-lingual WSD models (Section 3.1.2) measured in terms of accuracy: proportion of correctly translated ambiguous words.

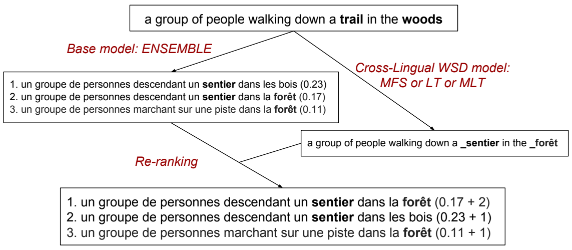
a group of people walking down a **trail** in the **woods**
Base model: ENSEMBLE
1. un groupe de personnes descendant un **sentier** dans les bois (0.23)
2. un groupe de personnes descendant un **sentier** dans la **forêt** (0.17)
3. un groupe de personnes marchant sur une piste dans la **forêt** (0.11)
Cross-Lingual WSD model: MFS or LT or MLT
a group of people walking down a **_sentier** in the **_forêt**
Re-ranking
1. un groupe de personnes descendant un **sentier** dans la **forêt** (0.17 + 2)
2. un groupe de personnes descendant un **sentier** dans les bois (0.23 + 1)
3. un groupe de personnes marchant sur une piste dans la **forêt** (0.11 + 1)

Figure 2: Task 1 system pipeline. The base model generates n-best translation candidates of the source sentence. The cross-lingual WSD model translates ambiguous words in the source sentence. The re-ranking step uses these lexical translations to re-score the translation candidates.

3.1.3 Re-ranking

Our re-ranking strategy is depicted in Figure 2. First, given an English source sentence, the base model generates an n-best list of translation candidates with a likelihood score. The idea is to select the translation candidate in the n-best translations which correctly disambiguates as many ambiguous words in the source sentence as possible.

The source sentence in our example (Figure 2) contains two ambiguous words *trail* and *woods* as per the English-French MLTD. We use a cross-lingual WSD model, MFS or LT or MLT, to predict the lexical translations of these words (the correct ones being *sentier* and *forêt* respectively in this example). Next, we match these to the words in the translation candidates and add the number of matching words to the original score[6] of the candidates. Then, the n-best translations are re-ranked using the new scores and the top candidate (which has the highest number of matches) is used in the evaluation.

[4] We tried a few other ways of using the image features - like concatenating it to word embeddings, using it as a separate word, etc. - but these did not result in any improvements.

[5] The performance of cross-lingual WSD models for EN-CS language direction could not be evaluated because the EN-CS Multimodal Lexical Translation Dataset was noisy. The clean 'filtered by human' versions of the EN-CS MLTD test sets were not ready at the time of submitting this paper.

[6] The likelihood score assigned to the candidate by the baseline NMT model

3.2 Task 1b systems

Three different approaches were explored in our submissions for Task 1b. The first approach follows the re-raking experiments using MLT for Task 1. The second approach exploits consensus-based selection and the third explores data augmentation and n-best selection through classification. We try two different types of classifiers - Random Forest and Recurrent Neural Network.

Re-ranking using MLT For the re-ranking approach, we first train three baseline EN-CS, DE-CS and FR-CS NMT models. Given a source sentence in the test set, we generate 10-best translation hypotheses using each of the three models. The three 10-best lists are concatenated to form a list of 30 translation hypotheses. We then use the trained EN-CS MLT model for cross-lingual WSD and perform re-ranking as mentioned in 3.1.2 and 3.1.3.

Consensus-based selection For the consensus-based selection approach, we again use the three 10-best translation hypotheses coming from the EN-CS, DE-CS and FR-CS systems. We then explore consensus between the different 10-best lists. The best hypothesis is selected according to the number of times it appears in the different lists. We follow the order of the EN-CS 10-best list: the highest ranked hypothesis in the EN-CS list with the majority of the votes (measured in terms of whether it occurs in the DE-CS and FR-CS 10-best lists) is selected.

Data augmentation We explore data augmentation by creating systems that first translate source sentences from French, German and Czech into English. This leads to variants of the source data that translate into the same Czech sentence. The augmented data is used to train an NMT system to translate test source sentences from English into Czech. We then obtain a 10-best list for the training, development and test sets. For the selection approach, we compute METEOR scores for each of the hypotheses in the 10-best list of the training set. To treat this as a binary classification task, we set a threshold such that the top four hypotheses are assumed to be the best translations and are chosen as positive samples, with the remaining six as bad examples.[7] This is then used to train two types of classifiers:

- Random Forest (RF) classifier: we use the image vectors concatenated with sentence embeddings from source and target sentences as features for training the classifier. For extracting sentence embeddings, we use the approach of Arora et al. (2016). Pre-trained embeddings for English and Czech from MUSE[8] (Conneau et al., 2018) are used.The RF algorithm in the scikit-learn framework (Pedregosa et al., 2011) is trained to distinguish between good and bad translations.
- RNN classifier: We use a simple RNN-based classifier where the last hidden state of the encoded sentence is concatenated with the image vector and used with a hinge loss to distinguish between good and bad translations.

4 Results

For both tasks, the initial evaluation was performed in terms of METEOR, BLEU (Papineni et al., 2002) and TER (Snover et al., 2006), with METEOR as the primary metric. Direct human assessments of translation adequacy will be used for the final evaluation by the task organizers.

For task 1, our submitted systems consisted of: a) SHEF_LT: re-ranking using LT model; b) SHEF_MLT: re-ranking using MLT model; c) SHEF_MFS: re-ranking using MFS model; and d) SHEF_Baseline: our baseline text-only ensemble NMT model

Table 3 shows the official evaluation results of our systems submitted to Task 1 and the baseline system provided by the organizers. For all language pairs, our systems outperform the official baseline for all metrics.

For EN-DE and EN-FR, the systems with LT and MLT are slightly better than the system with MFS. For EN-CS, however, the MFS system scores better than the LT and MLT variants. This is, perhaps, because the EN-CS MLTD (on which LT and MLT models are trained) is noisy, as previously mentioned. The dataset has been extracted using the same procedure in (Lala and Specia, 2018) except for the human filtering step, which is crucial for a clean dataset.

On further inspection, we observe that the cross-lingual WSD re-ranking affects only 127 to

[7] This threshold was empirically defined.

[8] https://github.com/facebookresearch/MUSE

	EN-DE			EN-FR			EN-CS		
	METEOR	BLEU	TER	METEOR	BLEU	TER	METEOR	BLEU	TER
SHEF_LT	50.7	30.5	53.0	59.8	38.8	41.5	29.1	28.3	51.7
SHEF_MLT	50.7	30.4	52.9	59.8	38.9	41.5	29.1	28.2	51.7
SHEF_Baseline	50.7	30.9	52.4	59.8	38.9	41.2	29.4	29.0	51.1
SHEF_MFS	50.7	30.3	53.1	59.7	38.8	41.6	29.2	27.8	52.4
Baseline	47.4	27.6	55.2	56.9	36.3	41.6	27.7	26.5	54.4

Table 3: Evaluation of our systems and the baseline for Task 1. We show METEOR, BLEU and TER scores.

	MFS	LT	MLT
EN-DE	189 (239)	149 (200)	148 (189)
EN-FR	163 (244)	127 (180)	129 (192)
EN-CS	484 (649)	100 (124)	124 (148)

Table 4: The effect of re-ranking approaches on the baseline NMT model outputs. The number outside the bracket shows the number of instances that are affected due to re-ranking in the 1071 test instances. The number inside the bracket '()' shows the number of words in the entire test set that are affected (deleted, added or replaced) due to re-ranking.

	METEOR	BLEU	TER
SHEF_CON	27.6	24.7	52.1
SHEF_MLT	27.5	24.5	52.5
SHEF_ARNN	27.5	25.2	53.9
SHEF_ARF	27.1	24.1	54.6
Baseline	26.8	23.6	54.2

Table 5: Evaluation of our systems and the baseline for Task 1b.

189 test instances (for EN-DE and EN-FR only[9]) out of the total 1,071 test instances (See Table 4). These usually result in changing only one or two words and as a result it affects only 180 to 244 words in the entire test set (See Table 4). In other words, only 1.4% words in the entire test set are affected by the re-ranking, which may explain why the performance of all the systems is so similar. It also suggests that automatics metrics like BLEU, METEOR and TER may not be sufficient to detect subtle changes in translation quality making it difficult to deduce insights from our re-ranking approaches. We hope to rely on Direct Human Assessment and other more sensitive metrics to help to better understand the affects.

For Task 1b, we submitted four models: a) SHEF_CON: consensus based model; b) SHEF_MLT: a re-ranking approach using MLT model; c) SHEF_ARNN: a data augmentation and hypothesis selection approach using an RNN classifier; and d) SHEF_ARF: data augmentation and hypothesis selection approach using an RF classifier.

Table 5 shows the automatic metric scores for our systems and the official baseline. Our systems outperform the baseline in terms of BLEU and METEOR. For TER, all systems are better than the baseline except for SHEF_ARF. Our best performing system is SHEF_CON.

5 Conclusions

We have described our submissions to the Multimodal Machine Translation shared task at WMT18. We explored novel multimodal n-best re-ranking approaches for task 1, and consensus-based approaches for task 1b using image information for re-ranking of an augmented n-best list with outputs from different translation models.

All our models perform better than the official baseline for all metrics and language pairs in task 1. However, we observe that SHEF_LT and SHEF_MLT, for the dataset and in the current setup, are not significantly different and their performance are nearly identical which indicates that the image information is not contributing significantly for this task and cross-lingual WSD is, perhaps, not very useful. On the other hand, it is worth emphasising that the corpora used may not show many ambiguous words and our model is not expected to be beneficial in this case.

For task 1b, our models also outperform the official baseline, with the best model being SHEF_CON. As for task 1, the use of image information do not lead to improvements when evaluated using automatic metrics METEOR, BLEU and TER.

[9] We ignore EN-CS in this observation because the EN-CS MLTD is noisy and thus the trained cross-lingual WSD models are not reliable for this language pair.

Acknowledgements

This work is supported by the MultiMT project (H2020 ERC Starting Grant No. 678017).

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. *In proceedings of International Conference on Learning Representations*.

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*.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost Van de Weijer. 2017a. Lium-cvc submissions for wmt17 multimodal translation task. *In proceedings of Conference on Machine Translation (WMT)*.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? *In proceedings of the First Conference on Machine Translation*.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017b. Nmtpy: A flexible toolkit for advanced neural machine translation systems. *In proceedings of The Prague Bulletin of Mathematical Linguistics*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation*, page 103.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *In proceedings of the International Conference on Learning Representations*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. *In proceedings of the Ninth Workshop on Statistical Machine Translation*.

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, Volume 2: Shared Task Papers*.

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

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*.

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

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)*.

Els Lefever and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. *In proceedings of the Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *In proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *In proceedings of the Journal of Machine Learning Research*.

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*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *In proceedings of International Journal of Computer Vision*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *In proceedings of the Seventh Biennial Conference of the Association for Machine Translation in the Americas*.

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*.

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*.