

# Using Images to Improve Machine-Translating E-Commerce Product Listings

Iacer Calixto<sup>1</sup>, Daniel Stein<sup>2</sup>, Evgeny Matusov<sup>2</sup>,  
Pintu Lohar<sup>1</sup>, Sheila Castilho<sup>1</sup> and Andy Way<sup>1</sup>

<sup>1</sup>ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland

<sup>2</sup>eBay Inc., Aachen, Germany

{iacer.calixto, pintu.lohar, sheila.castilho, andy.way}@adaptcentre.ie

{danstein, ematusov}@ebay.com

## Abstract

In this paper we study the impact of using images to machine-translate user-generated e-commerce product listings. We study how a multi-modal Neural Machine Translation (NMT) model compares to two text-only approaches: a conventional state-of-the-art attentional NMT and a Statistical Machine Translation (SMT) model. User-generated product listings often do not constitute grammatical or well-formed sentences. More often than not, they consist of the juxtaposition of short phrases or keywords. We train our models end-to-end as well as use text-only and multi-modal NMT models for re-ranking  $n$ -best lists generated by an SMT model. We qualitatively evaluate our user-generated training data also analyse how adding synthetic data impacts the results. We evaluate our models quantitatively using BLEU and TER and find that (i) additional synthetic data has a general positive impact on text-only and multi-modal NMT models, and that (ii) using a multi-modal NMT model for re-ranking  $n$ -best lists improves TER significantly across different  $n$ -best list sizes.

## 1 Introduction

In e-commerce, there is a strong requirement to make products accessible regardless of the customer’s native language and home country, by leveraging the gains available from machine translation (MT). Among the challenges in automatic processing are the specialized language and grammar for listing titles, as well as a high percentage of user-generated content for non-business sellers, who often are not native speakers themselves.

We investigate the nature of user-generated auction listings’ titles as listed on the eBay main site<sup>1</sup>. Product listings contain extremely high trigram perplexities even if trained (and applied) on in-domain data, which is a challenge not only for proper language models but also for automatic evaluation metrics such as the  $n$ -gram precision-based BLEU (Papineni et al., 2002)

<sup>1</sup><http://www.ebay.com/>

metric. Nevertheless, when presenting humans with images of the product which come along with the auction titles, the listings are perceived as somewhere between “easy” and “neutral” to understand.

Images can bring useful complementary information to MT (Calixto et al., 2012; Hitschler et al., 2016; Huang et al., 2016). Therefore, we explore the potential of multi-modal, multilingual MT of auction listings’ titles and product images from English into German. To that end, we compare eBay’s production system, due to service-level agreements a classic phrase-based statistical MT (PBSMT) system, with two neural MT (NMT) systems. One of the NMT models is a text-only attentional NMT and the other is a multi-modal attentional NMT model trained using the product images as additional data.

PBSMT still outperforms both text-only and multi-modal NMT models in the translation of product listings, contrary to recent findings (Bentivogli et al., 2016). Under the hypothesis that the amount of training data could be the culprit and since curated multilingual, multi-modal in-domain data is very expensive to obtain, we back-translate monolingual listings and incorporate them as additional synthetic training data. Utilising synthetic data leads to big gains in performance and ultimately brings NMT models closer to bridging the gap with an optimized PBSMT system. We also use multi-modal NMT models to rescore the output of a PBSMT system and show significant improvements in TER (Snover et al., 2006).

This paper is structured as follows. In §2 we describe the text-only and multi-modal MT models we evaluate and in §3 the data sets we used, also introducing and discussing interesting findings. In §4 we discuss how we structure our quantitative evaluation, and in §5 we analyse and discuss our results. In §6 we discuss some relevant related work and in §7 we draw conclusions and devise future work.

## 2 Model

We first briefly introduce the two text-only baselines used in this work: a PBSMT model (§2.1) and a text-only attentive NMT model (§2.2). We then discuss the doubly-attentive multi-modal NMT model that we use in our experiments (§2.3), which is comparable to the model introduced by Calixto et al. (2016).

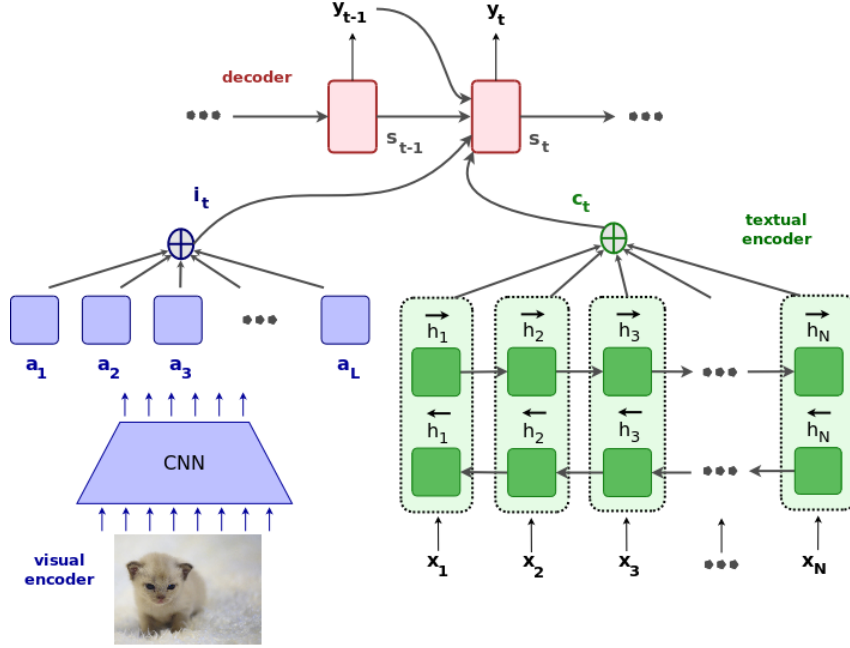


Figure 1: Decoder RNN with attention over source sentence and image features. This decoder learns to independently attend to image patches and source-language words when generating translations.

## 2.1 Statistical Machine Translation (SMT)

We use a PBSMT model built with the Moses SMT Toolkit (Koehn et al., 2007). The language model (LM) is a 5-gram LM with modified Kneser-Ney smoothing (Kneser and Ney, 1995). We use minimum error rate training (Och, 2003) for tuning the model parameters for BLEU scores.

## 2.2 Text-only Neural Machine Translation (NMT<sub>t</sub>)

We use the attentive NMT model introduced by Bahdanau et al. (2015) as our text-only NMT baseline. It is based on the encoder–decoder framework and it implements an attention mechanism over the source-sentence words. Being  $X = (x_1, x_2, \dots, x_N)$  and  $Y = (y_1, y_2, \dots, y_M)$  a one-hot representation of a sentence in a source language and its translation into a target language, respectively, the model is trained to maximise the log-likelihood of the target given the source.

The encoder is a bidirectional recurrent neural network (Schuster and Paliwal, 1997) with GRU units (Cho et al., 2014). The annotation vector for a given source word  $x_i$ ,  $i \in [1, N]$  is the concatenation of forward and backward vectors  $\mathbf{h}_i = [\vec{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i]$  obtained with forward and backward RNNs, respectively, and  $\mathbf{C} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N)$  is the set of source annotation vectors.

The decoder is also a recurrent neural network, more specifically a neural LM (Bengio et al., 2003) conditioned upon its past predictions via its previous hidden state  $\mathbf{s}_{t-1}$  and the word emitted in the previous time step  $y_{t-1}$ , as well as the source sentence via an *atten-*

*tion mechanism*. The attention mechanism computes a context vector  $\mathbf{c}_t$  for each time step  $t$  of the decoder where this vector is a weighted sum of the source annotation vectors  $\mathbf{C}$ :

$$\mathbf{e}_{t,i}^{\text{src}} = (\mathbf{v}_a^{\text{src}})^T \tanh(\mathbf{U}_a^{\text{src}} \mathbf{s}_{t-1} + \mathbf{W}_a^{\text{src}} \mathbf{h}_i), \quad (1)$$

$$\alpha_{t,i}^{\text{src}} = \frac{\exp(\mathbf{e}_{t,i}^{\text{src}})}{\sum_{j=1}^N \exp(\mathbf{e}_{t,j}^{\text{src}})}, \quad (2)$$

$$\mathbf{c}_t = \sum_{i=1}^N \alpha_{t,i}^{\text{src}} \mathbf{h}_i, \quad (3)$$

where  $\alpha_{t,i}^{\text{src}}$  is the normalised alignment matrix between each source annotation vector  $\mathbf{h}_i$  and the word to be emitted at time step  $t$ , and  $\mathbf{v}_a^{\text{src}}$ ,  $\mathbf{U}_a^{\text{src}}$  and  $\mathbf{W}_a^{\text{src}}$  are model parameters.

## 2.3 Multi-modal Neural Machine Translation (NMT<sub>m</sub>)

We use a multi-modal NMT model similar to the one introduced by Calixto et al. (2016), illustrated in Figure 1. It can be seen as an expansion of the attentive NMT framework described in §2.2 with the addition of a *visual component* to incorporate visual features.

We use a publicly available pre-trained Convolutional Neural Network (CNN), namely the 50-layer Residual network (ResNet-50) of He et al. (2015) to extract convolutional image features  $(\mathbf{a}_1, \dots, \mathbf{a}_L)$  for all images in our dataset. These features are extracted from the `res4f` layer and consist of a  $196 \times 1024$  dimensional matrix where each row (i.e., a 1024D vector) represents features from a specific area and therefore only encodes information about that specific area

of the image. In our NMT experiments, the ResNet-50 network is fixed during training, and there is no fine-tuning done for the translation task.

The visual attention mechanism computes a context vector  $i_t$  for each time step  $t$  of the decoder similarly to the textual attention mechanism described in §2.2:

$$e_{t,l}^{\text{img}} = (\mathbf{v}_a^{\text{img}})^T \tanh(\mathbf{U}_a^{\text{img}} \mathbf{s}_{t-1} + \mathbf{W}_a^{\text{img}} \mathbf{a}_l), \quad (4)$$

$$\alpha_{t,l}^{\text{img}} = \frac{\exp(e_{t,l}^{\text{img}})}{\sum_{j=1}^L \exp(e_{t,j}^{\text{img}})}, \quad (5)$$

$$\mathbf{i}_t = \sum_{l=1}^L \alpha_{t,l}^{\text{img}} \mathbf{a}_l, \quad (6)$$

where  $\alpha_{t,l}^{\text{img}}$  is the normalised alignment matrix between each image annotation vector  $\mathbf{a}_l$  and the word to be emitted at time step  $t$ , and  $\mathbf{v}_a^{\text{img}}$ ,  $\mathbf{U}_a^{\text{img}}$  and  $\mathbf{W}_a^{\text{img}}$  are model parameters.

### 3 Data sets

The multi-modal NMT model we evaluate uses parallel sentences and an image as input. Thus, we use the data set of product listings and images produced by eBay. They consist of 23,697 triples of products, henceforth *original*, containing each (i) a listing in English, (ii) its translation into German and (iii) a product image. Validation and test sets used in our experiments consist of 480 and 444 triples, respectively.

The curation of parallel product listings with an accompanying product image is costly and time-consuming, so the in-domain data is rather small. More easily accessible are monolingual German listings accompanied by the product image where the source text input can be emulated by back-translating the target listing. For this set of experiments, we use 83,832 tuples, henceforth *mono*. Finally, we also use the publicly available Multi30k dataset (Elliott et al., 2016), a multilingual expansion of the original Flickr30k (Young et al., 2014) with  $\sim 30\text{k}$  pictures from Flickr, one description in English and one human translation of the English description into German.

Translating user-generated product listings has particular challenges; they are often ungrammatical and can be difficult to interpret in isolation even by a native speaker of the language, as can be seen in the examples in Table 1. To further demonstrate this issue, in Table 2 we show the number of running words as well as the perplexity scores obtained with LMs trained on three sets of different German corpora: the Multi30k, eBay’s in-domain data and a concatenation of the WMT 2015<sup>2</sup> Europarl (Koehn, 2005), Common Crawl and News Commentary corpora (Bojar et al., 2015).<sup>3</sup>

<sup>2</sup>We use the German side of the English–German parallel WMT 2015 corpora.

<sup>3</sup>These are 5-gram LMs trained with KenLM (Heafield et al., 2013) using modified Kneser-Ney smoothing (Kneser and Ney, 1995) on tokenized, lowercased data.



Image	Product Listing
	(en) just rewired original mission 774 fluid damped low mass tonearm , very good cond . (de) vor kurzem neu verkabelter flüssigkeitsgedämpfter leichter original - mission 774 - tonarm , sehr guter zustand
	(en) mary kay cheek color mineral pick citrus bloom shy blush bold berry + more (de) mary kay mineral cheek colour farbauswahl citrus bloom shy blush bold berry + mehr

Table 1: Examples of product listings and their accompanying image.

LM training corpus	#words (×1000)	Perplexity (×1000)	
		eBay	Multi30k
WMT’15	4310.0	60.1	<b>0.5</b>
Multi30k	29.0	25.2	<b>0.05</b>
eBay	99.0	<b>1.8</b>	4.2

Table 2: Perplexity on eBay and Multi30k’s test sets for LMs trained on different corpora. WMT’15 is the concatenation of the Europarl, Common Crawl and News Commentary corpora (the German side of the parallel English–German corpora).

We see that different LM perplexities on eBay’s test set are high even for an LM trained on eBay in-domain data. LMs trained on mixed-domain corpora such as the WMT 2015 corpora or the Multi30k have perplexities below 500 on the Multi30k test set, which is expected. However, when applied to eBay’s test data, perplexities computed can be over 60k. Conversely, an LM trained on eBay in-domain data, when applied to the Multi30k test set, also computes very high perplexity scores. These perplexity scores indicate that *fluency* might not be a good metric to use in our study, i.e. we should not expect a fluent machine-translated output of a model trained on poorly fluent training data.

Clearly, translating user-generated product listings is very challenging; for that reason, we decided to check with humans how they perceive that data with and without having the associated images available. We hypothesise that images bring additional understanding to their corresponding listings.

#### 3.1 Source (target) product title–image assessment

A human evaluator is presented with the English (German) product listing. Half of them are also shown the product image, whereas the other half is not. For the first group, we ask two questions: (i) in the context of the product image, how easy it is to understand the English (German) product listing and (ii) how well does the English (German) product listing describe the

Listing language	$N$	Difficulty		Adequacy
		listing only	listing+image	listing+image
English	20	$2.50 \pm 0.84$	$2.40 \pm 0.84$	$2.45 \pm 0.49$
German	15	$2.83 \pm 0.75$	<b><math>2.00 \pm 0.50</math></b>	$2.39 \pm 0.78$

Table 3: Difficulty to understand product listings with and without images and adequacy of product listings and images.  $N$  is the number of raters.

product image. For the second group, we just ask (i) how easy it is to understand the English (German) product listing. In all cases humans must select from a five-level Likert scale where in (i) answers range from 1–*Very easy* to 5–*Very difficult* and in (ii) from 1–*Very well* to 5–*Very poorly*.

Table 3 suggests that the intelligibility of both the English and German product listings are perceived to be somewhere between “easy” and “neutral” when images are also available. It is notable that, for German, there is a statistically significant difference between the group who had access to the image and the product listings ( $M=2.00$ ,  $SD=.50$ ) and the group who only viewed the listings ( $M=2.83$ ,  $ST=.30$ ), where  $F(1,13) = 6.72$ ,  $p < 0.05$ . Furthermore, humans find that product listings describe the associated image somewhere between “well” and “neutral” with no statistically significant differences between the adequacy of product listings and images in different languages.

Altogether, we have a strong indication that images can indeed help an MT model translate product listings, especially for translations into German.

## 4 Experimental setup

The PBSMT model we use as a baseline is trained on 120k in-domain parallel sentences (§2.1).

To measure how well multi-modal and text-only NMT models perform when trained on exactly the same data with and without images, respectively, we trained them only on the *original* and the Multi30k (Elliott et al., 2016) data sets. We also did not use any additional parallel, but out-of-domain data that had been used to train eBay’s PBSMT production system (see Section 5). Training our text-only NMT<sub>t</sub> baseline on this large corpus would not help shed more light on how multi-modality helps MT, since it has no images available and thus cannot be used to train the multi-modal model NMT<sub>m</sub>. Rather, we report results of re-ranking experiments using  $n$ -best lists generated by eBay’s best-performing PBSMT production system.

In order to measure the impact of the training data size on MT quality, we follow Sennrich et al. (2016) and back-translate the *mono* German product listings using our baseline NMT<sub>t</sub> model trained on the *original* 23, 697 German→English corpus (- images). These additional synthetic data (including images) are added to the *original*’s 23, 697 triples and used in our translation experiments. We do not include the back-translated data set when training NMT models for re-ranking  $n$ -

Model	Training data	BLEU	TER
PBSMT	original + Multi30k	26.1	<b>54.9</b>
	+ backtranslated	<b>27.4</b> $\uparrow 1.3$	55.4 $\uparrow 0.5$
NMT <sub>t</sub>	original + Multi30k	21.1	60.0
	+ backtranslated	22.5 $\uparrow 1.4$	58.0 $\downarrow 2.0$
NMT <sub>m</sub>	original + Multi30k	17.8	62.2
	+ backtranslated	<b>25.1</b> $\uparrow 7.3$	<b>55.5</b> $\downarrow 6.7$
Improvements			
NMT <sub>m</sub> vs. NMT <sub>t</sub>		$\uparrow 2.3$	$\downarrow 2.5$
NMT <sub>m</sub> vs. SMT <sub>t</sub>		$\downarrow 2.3$	$\uparrow 0.6$

Table 4: Comparative results with PBSMT, NMT<sub>t</sub> and multi-modal models NMT<sub>m</sub> evaluated on eBay’s test set. Best PBSMT and NMT results in bold.

best lists to be able to evaluate these two scenarios independently.

We evaluate our models quantitatively using BLEU4 (Papineni et al., 2002) and TER (Snover et al., 2006) and report statistical significance computed using approximate randomisation with the Multeval toolkit (Clark et al., 2011).

## 5 Results

In Table 4 we present quantitative results obtained with the two text-only baselines SMT and NMT<sub>t</sub> and one multi-modal model NMT<sub>m</sub>.

It is clear that the gains from adding more data are much more apparent to the multi-modal NMT<sub>m</sub> model than to the two text-only ones. This can be attributed to the fact that this model has access to more data, i.e. image features, and consequently can learn better representations derived from them. The PBSMT model’s improvements are inconsistent; its TER score even deteriorates by 0.5 with the additional data. The same does not happen with the NMT models, which both (text-only and multi-modal) benefit from the additional data. Model NMT<sub>m</sub>’s gains are more than 3× larger than that of models NMT<sub>t</sub> and SMT, indicating that they can properly exploit the additional data. Nevertheless, even with the added back-translated data, model NMT<sub>m</sub> still falls behind the PBSMT model both in terms of BLEU and TER, although it seems to be catching up as the data size increases.

In Table 5, we show results for re-ranking 10- and 100-best lists generated by eBay’s PBSMT production system. This system was trained with additional data sampled from out-of-domain corpora and also includes extra features and optimizations. Its BLEU score on the eBay test set is 29.0. Nevertheless, we still observe improvements in rescoring of  $n$ -best lists from this system using our “weaker” NMT models. When  $n = 10$ , both models NMT<sub>t</sub> and NMT<sub>m</sub> significantly improve the baseline in terms of TER, with model NMT<sub>m</sub> performing slightly better. With larger lists ( $n = 100$ ), it seems that both neural models have more difficulty to re-rank. Nonetheless, in this scenario model NMT<sub>m</sub> still sig-

Model	Training data	N	BLEU	oracle	TER	oracle	Translation length
baseline		—	29.0	—	53.0	—	13.60 ± 2.59
NMT <sub>t</sub>	100k in-domain	10	29.3 ↑ 0.3	35.4	52.4 † ↓ 0.6	46.4	13.48 ± 2.59
NMT <sub>m</sub>	orig. + Multi30k	10	<b>29.4</b> ↑ 0.4	35.4	<b>52.1</b> † ↓ 0.9	46.4	13.41 ± 2.58
NMT <sub>t</sub>	100k in-domain	100	<u>28.9</u> ↓ 0.1	42.2	53.6 ↑ 0.6	41.0	13.80 ± 2.67
NMT <sub>m</sub>	orig. + Multi30k	100	<u>28.9</u> ↓ 0.1	42.2	<u>52.4</u> † ↓ 0.6	41.0	13.50 ± 2.59

Table 5: Results for re-ranking  $n$ -best lists generated for eBay’s test set with text-only and multi-modal NMT models. †Difference is statistically significant ( $p \leq 0.05$ ). Best individual results are underscored, best overall results in bold. We also show the translation length for re-ranked  $n$ -best lists.

nificantly improves the MT quality in terms of TER, while model NMT<sub>t</sub> shows differences in BLEU and TER which are not statistically significant ( $p \leq 0.05$ ). We note that model NMT<sub>m</sub>’s improvements in TER are consistent across different  $n$ -best list sizes; model NMT<sub>t</sub>’s improvements are not.

The best BLEU (= 29.4) and TER (= 52.1) scores were achieved by model NMT<sub>m</sub> when applied to re-rank 10-best lists, although model NMT<sub>m</sub> still improves in terms of TER when  $n = 100$ . This suggests that model NMT<sub>m</sub> can efficiently exploit the additional multi-modal signals.

In order to check whether improvements observed in TER could be due to a preference of text-only and multi-modal NMT models for shorter sentences (Table 5), we also computed the average length of translations for  $n$ -best lists re-ranked with each of our models, and note that there is no significant difference between the length of translations for the baseline and the re-ranked models.

## 6 Related work

NMT has been successfully tackled by different groups using the sequence-to-sequence framework (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014). However, multi-modal MT has just recently been addressed by the MT community in a shared task (Specia et al., 2016). In NMT, Bahdanau et al. (2015) first proposed to use an *attention mechanism* in the decoder. Their decoder learns to attend to the relevant source-language words as it generates each word of the target sentence. Since then, many authors have proposed different ways to incorporate attention into MT (Luong et al., 2015; Firat et al., 2016; Tu et al., 2016).

In the context of image description generation (IDG), Vinyals et al. (2015) proposed an influential neural IDG model based on the sequence-to-sequence framework and trained end-to-end. Elliott et al. (2015) put forward a model to generate multilingual descriptions of images by learning and transferring features between two independent, non-attentive neural image description models. Finally, Xu et al. (2015) proposed an attention-based model where a model learns to attend to specific areas of an image representation as it

generates its description in natural language with a soft-attention mechanism.

Although no purely neural multi-modal model to date has significantly improved on both text-only NMT and SMT models on the Multi30k data set (Specia et al., 2016), different research groups have proposed to include images in re-ranking  $n$ -best lists generated by an SMT system or directly in a NMT framework with some success (Caglayan et al., 2016; Calixto et al., 2016; Huang et al., 2016; Libovický et al., 2016; Shah et al., 2016).

To the best of our knowledge, we are the first to study multi-modal NMT applied to the translation of product listings, i.e. for the e-commerce domain.

## 7 Conclusions and Future work

In this paper, we investigate the potential impact of multi-modal NMT in the context of e-commerce product listings. With only a limited amount of multi-modal and multilingual training data available, both text-only and multi-modal NMT models still fail to outperform a productive SMT system, contrary to recent findings (Bentivogli et al., 2016). However, the introduction of back-translated data leads to substantial improvements, especially to a multi-modal NMT model. This seems to be an interesting approach that we will continue to explore in future work.

We also found that NMT models trained on small in-domain data sets can still be successfully used to rescore a standard PBSMT system with significant improvements in TER. Since we know from our experiments with LM perplexities that these are very high for e-commerce data. i.e. fluency is quite low, it seems fitting that BLEU scores do not improve as much. In future work, we will also conduct a human evaluation of the translations generated by the various systems.

## Acknowledgements

The ADAPT Centre for Digital Content Technology ([www.adaptcentre.ie](http://www.adaptcentre.ie)) at Dublin City University is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.



## References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations. ICLR 2015*, San Diego, CA.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3:1137–1155, March.
- Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 257–267, Austin, Texas.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.
- 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*, pages 627–633, Berlin, Germany.
- Iacer Calixto, Teofilo de Campos, and Lucia Specia. 2012. Images as context in Statistical Machine Translation. In *The 2nd Annual Meeting of the EPSRC Network on Vision & Language (VL<sup>2</sup>)*, Sheffield, UK. EPSRC Vision and Language Network.
- Iacer Calixto, Desmond Elliott, and Stella Frank. 2016. DCU-UvA Multimodal MT System Report. In *Proceedings of the First Conference on Machine Translation*, pages 634–638, Berlin, Germany.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar.
- Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT '11*, pages 176–181, Portland, Oregon.
- Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-Language Image Description with Neural Sequence Models. *CoRR*, abs/1510.04709.
- Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Workshop on Vision and Language at ACL '16*, Berlin, Germany.
- Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.
- Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.
- Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal Pivots for Image Caption Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2399–2409, Berlin, Germany.
- Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based Multimodal Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 639–645, Berlin, Germany.
- Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1700–1709, Seattle, USA.
- Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 181–184, Detroit, Michigan.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07*, pages 177–180, Prague, Czech Republic.
- Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand.

- Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. CUNI System for WMT16 Automatic Post-Editing and Multimodal Translation Tasks. In *Proceedings of the First Conference on Machine Translation*, pages 646–654, Berlin, Germany.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, Lisbon, Portugal.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03*, pages 160–167, Sapporo, Japan.
- 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 on Association for Computational Linguistics, ACL '02*, pages 311–318, Philadelphia, Pennsylvania.
- M. Schuster and K.K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.
- Kashif Shah, Josiah Wang, and Lucia Specia. 2016. SHEF-Multimodal: Grounding Machine Translation on Images. In *Proceedings of the First Conference on Machine Translation*, pages 660–665, Berlin, Germany.
- 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 Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA.
- 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*, pages 543–553, Berlin, Germany.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems, NIPS*, pages 3104–3112, Montréal, Canada.
- Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling Coverage for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 3156–3164, Boston, Massachusetts.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2048–2057, Lille, France.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.