Human Evaluation of Multi-modal Neural Machine Translation:
a Case Study on E-commerce Listing Titles

Iacer Calixto[1], Daniel Stein[2], Evgeny Matusov[2], Sheila Castilho[1] and Andy Way[1]

[1]ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
[2]eBay Inc., Aachen, Germany
{iacer.calixto,sheila.castilho,andy.way}@adaptcentre.ie
{danstein,ematusov}@ebay.com

Abstract

In this paper, we study how humans perceive the use of images as an additional knowledge source to machine-translate user-generated product listings in an e-commerce company. We conduct a human evaluation where we assess how a multi-modal neural machine translation (NMT) model compares to two text-only approaches: a conventional state-of-the-art attention-based NMT and a phrase-based statistical machine translation (PBSMT) model. We evaluate translations obtained with different systems and also discuss the data set of user-generated product listings, which in our case comprises both product listings and associated images. We found that humans preferred translations obtained with a PBSMT system to both text-only and multi-modal NMT over 56% of the time. Nonetheless, human evaluators ranked translations from a multi-modal NMT model as better than those of a text-only NMT over 88% of the time, which suggests that images do help NMT in this use-case.

1 Introduction

In e-commerce, leveraging Machine Translation (MT) to make products accessible regardless of the customer's native language or country of origin is a very persuasive use-case. In this work, we study how humans perceive the machine translation of user-generated auction listings' titles as listed on the eBay main site[1]. Among the challenges for MT are the specialized language and grammar for listing titles, as well as a high percentage of user-generated content for non-business sellers, who are often not native speakers themselves. This is reflected on the data by means of extremely high trigram perplexities of product listings, which is in 4 digit numbers even for language models (LMs) trained on in-domain data, as we discuss in §3. This is not only a challenge for LMs but also for automatic evaluation metrics such as the n-gram precision-based BLEU metric (Papineni et al., 2002).

The majority of listings are accompanied by a product image, often (but not always) a user-generated shot. Moreover, images are known to bring useful complementary information to MT (Calixto et al., 2012; Hitschler et al., 2016; Huang et al., 2016; Calixto et al., 2017b). Therefore, in order to explore whether product images can benefit the machine translation of auction titles, we evaluate a multi-modal neural MT (NMT) system to eBay's production system, specifically a phrase-based statistical MT (PBSMT) one. We additionally train a text-only attention-based NMT baseline, so as to be able to measure eventual gains from the additional multi-modal data independently of the MT architecture.

According to a quantitative evaluation using a combination of four automatic MT evaluation metrics, a PBSMT system outperforms both text-only and multi-modal NMT models in the translation of product listings, contrary to recent findings (Bentivogli et al., 2016). We hypothesise that these automatic metrics were not created for the purpose of measuring the impact an image brings to an MT model, so we conduct a human evaluation of translations generated by three different systems: a PBSMT, a text-only attention-based NMT and a multi-modal NMT system. With that human evaluation we wish to see whether those findings corroborate the automatic scores or instead support results included in recent papers in the literature.

The remainder of the paper is structured as follows. In §2 we briefly describe the text-only and multi-modal MT models we evaluate in this work and in §3 the data sets we used, together with a discussion of interesting findings. In §4 we discuss how we structure our evaluation and in §5 we analyse and discuss our results. In §6 we discuss important related work and finally in §7 we draw conclusions and suggest avenues for future work.

2 MT Models evaluated in this work

We first introduce the two text-only baselines used in this work: a PBSMT model (§2.1) and a text-only attention-based NMT model (§2.2). We then briefly discuss the doubly-attentive multi-modal NMT model we use in our experiments (§2.3), which is comparable to the model evaluated by Calixto et al. (2016) and further detailed and analysed in Calixto et al. (2017a).
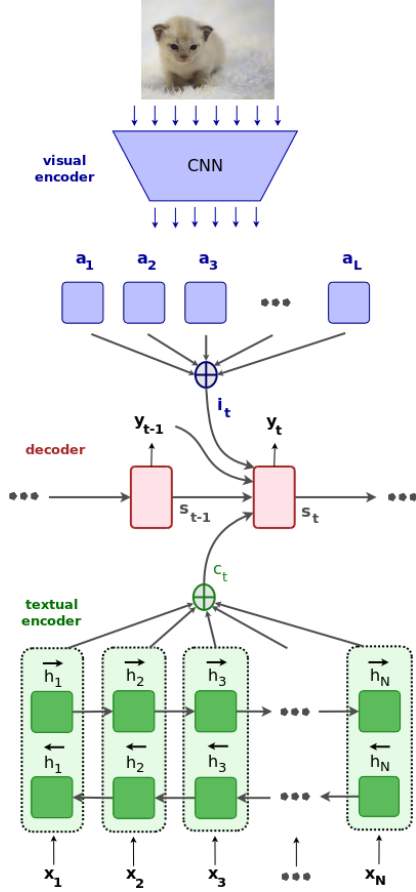
---

[1]http://www.ebay.com/

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 where 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 using BLEU as the objective function.

2.2 Text-only NMT (NMT$_t$)

We use the attention-based 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 $X = (x_1, x_2, \cdots, x_N)$, where $Y = (y_1, y_2, \cdots, y_M)$ is its target-language translation. A model is trained to maximise the log-likelihood of the target given the source.

The encoder is a bidirectional recurrent neural network (RNN) with GRU units (Cho et al., 2014). The annotation vector for a given source word x_i is the concatenation of forward and backward vectors $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$ obtained with forward and backward RNNs, respectively, and $C = (h_1, h_2, \cdots, h_N)$ is the set of source annotation vectors.

The decoder is also an RNN, more specifically a neural LM (Bengio et al., 2003) conditioned upon its past predictions via its previous hidden state s_{t-1} and the word emitted in the previous time step y_{t-1}, as well as the source sentence via an attention mechanism. The attention computes a context vector c_t for each time step t of the decoder where this vector is a weighted sum of the source annotation vectors C:

$$e_{t,i}^{\text{src}} = (v_a^{\text{src}})^T \tanh(U_a^{\text{src}} s_{t-1} + W_a^{\text{src}} h_i), \quad (1)$$

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

$$c_t = \sum_{i=1}^{N} \alpha_{t,i}^{\text{src}} h_i, \quad (3)$$

where $\alpha_{t,i}^{\text{src}}$ is the normalised alignment matrix between each source annotation vector h_i and the word to be emitted at time step t, and v_a^{src}, U_a^{src} and W_a^{src} are model parameters.

2.3 Multi-modal NMT (NMT$_m$)

We use a multi-modal NMT model similar to the one evaluated by Calixto et al. (2016) and further studied in Calixto et al. (2017a), 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 local 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. (2016) to extract convolutional image features $(a_1, \cdots, a_L)$ for all images in our dataset. These features are extracted from the res4f layer and consist of a 196 x 1024 dimensional matrix where each row, i.e. a 1024D vector, represents features from a specific area and so only encodes information about that specific area of the image.

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:

$$e_{t,l}^{\text{img}} = (v_a^{\text{img}})^T \tanh(U_a^{\text{img}} s_{t-1} + W_a^{\text{img}} a_l), \quad (4)$$

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

$$i_t = \sum_{l=1}^{L} \alpha_{t,l}^{\text{img}} a_l, \quad (6)$$

where $\alpha_{t,l}^{\text{img}}$ is the normalised alignment matrix between each image annotation vector a_l and the word to be emitted at time step t, and v_a^{img}, U_a^{img} and W_a^{img} are model parameters.

3 Data sets

We use the data set of product listings and images produced by eBay, henceforth referred to as eBay24k,

which consists of 23, 697 tuples of products each containing *(i)* a product listing in English, *(ii)* a product listing in German and *(iii)* a product image. In ~6k training tuples, the original user-generated product listing was given in English and was manually translated into German by in-house experts. The same holds for validation and test sets, which contain 480 and 444 triples, respectively. In the remaining training tuples (~18k), the original listing was given in German and manually translated into English. We also use the publicly available Multi30k dataset (Elliott et al., 2016), a multilingual expansion of the original Flickr30k (Young et al., 2014) with ~30k pictures from Flickr, each accompanied by one description in English and one human translation of the English description into German.

Although the curation of in-domain parallel product listings with an associated product image is costly and time-consuming, monolingual German listings with an image are far simpler to obtain. In order to increase the small amount of training data, we train the text-only model NMT$_t$ on the German–English eBay24k and Multi30k data sets (without images) and back-translate 83, 832 German in-domain product listings into English. We use the synthetic English, original German and original image as additional training tuples, henceforth eBay80k.

The translation of user-generated product titles raises particular challenges; they are often ungrammatical and can be difficult to interpret in isolation even by a native speaker of the language, as illustrated in Table 1. We note that the listings in both languages have many scattered keywords and/or phrases glued together, as well as few typos (e.g., English listing in the first example). Moreover, in the second example the product image has a white frame surrounding it. These are all complications that make the multi-modal MT of product listings a challenging task, where there are different difficulties derived from processing listings and images.

To further demonstrate these issues, we compute perplexity scores with LMs trained on one in-domain and one general-domain German corpus: the Multi30k (~ 29k sentences) and eBay's in-domain data (~ 99k sentences), respectively.[2] The LM trained on the Multi30k computes a perplexity of 25k on the eBay test set, and the LM trained on the in-domain eBay data produces a perplexity of 4.2k on the Multi30k test set. We note that the LM trained on eBay's in-domain data still computes a very high perplexity on eBay's test set (ppl = 1.8k). 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.

---

[2]These are 5-gram LMs trained with KenLM (Heafield et al., 2013) using modified Kneser-Ney smoothing on tokenized, lowercased data.

Image	Product Listing
	(en) apple macbook pro 13.3" laptop - dvd - rw drive / good screen / airport card keyboar
	(de) apple macbook pro laptop 13.3" - dvd - rw - laufwerk / gutes display / airport karte tastatur
	(en) modern napkin holder table top stainless steel weighted arm napkins paper towels
	(de) moderner tischserviettenhalter aus edelstahl mit beschwertem arm für servietten und papiertücher

Table 1: Examples of product listings accompanied by product images from the eBay test set.

Listing language	N	Difficulty		Adequacy
		listing only	listing+image	listing+image
English	20	2.50 ± 0.84	2.40 ± 0.84	2.45 ± 0.49
German	15	2.83 ± 0.75	$\mathbf{2.00} \pm 0.50$	2.39 ± 0.78

Table 2: Difficulty to understand product listings with and without images and adequacy of listings and images. N is the number of raters (Calixto et al., 2017b).

3.1 English and German product listings

Clearly, user-generated product listings are not very fluent in terms of grammar or even predictable word order. To better understand whether this has an impact on semantic intelligibility, Calixto et al. (2017b) have recently conducted experiments using eBay data to assess how challenging listings are to understand for a human reader. Specifically, they asked users how they perceive product listings with and without having the associated images available, under the hypothesis that images bring additional understanding to their corresponding listings.

In Table 2, we show results which suggest 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, in case of German, there is a statistically significant difference between the group who had access to the image and the product listing (M=2.00, SD=.50) and the group who only viewed the listing (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 (Calixto et al., 2017b).

Altogether, we have a strong indication that images can indeed help an MT model translate product listings, especially for translations into German.
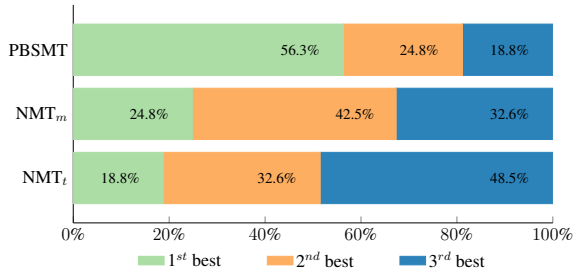
Figure 2: Models PBSMT, NMT$_t$ and NMT$_m$ ranked by humans from best to worst.

4 Experimental set-up

We use the eBay24k, the additional back-translated eBay80k and the Multi30k (Elliott et al., 2016) data sets to train all our models. In our experiments, we wish to contrast the human assessments of the adequacy of translations obtained with two text-only baselines, PBSMT and NMT$_t$, and one multi-modal model NMT$_m$, with scores computed with four automatic MT metrics: BLEU4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), TER (Snover et al., 2006), and chrF3 (Popović, 2015).[3] We report statistical significance with approximate randomisation for the first three metrics using the MultEval tool (Clark et al., 2011).

For our qualitative human evaluation, we ask bilingual native German speakers:

1. to assess the *multi-modal adequacy* of translations (number of participants $N = 18$), described in §4.1;

2. to *rank* translations generated by different models from best to worst (number of participants $N = 18$), described in §4.2.

On average, our evaluators' consisted of 72% women and 28% men. They were recruited from employees at eBay Inc., Aachen, Germany, as well as the student and staff body of Dublin City University, Dublin, Ireland.

4.1 Adequacy

Humans are presented with an English product listing, a product image and a translation generated by one of the models (without knowing which model). They are then asked how much of the meaning of the source is also expressed in the translation, taking the product image into consideration. They must then select from a four-level Likert scale where the answers range from *1 – All of it* to *4 – None of it*.

4.2 Ranking

We present humans with a product image and three translations obtained from different models for a particular English product listing (without identifying the

models) and ask them to rank translations from best to worst.

5 Results

In Table 3, we contrast the human assessments of the adequacy of translations obtained with two text-only baselines, PBSMT and NMT$_t$, and one multi-modal model NMT$_m$, with scores obtained computing four MT automatic metrics.

Both models NMT$_m$ and PBSMT improve on model NMT$_t$'s translations according to the first three automatic metrics ($p < 0.01$), and we also observe improvements in chrF3. Although a one-way anova did not show any statistically significant differences in adequacy between NMT$_m$ and NMT$_t$ (F$(2, 18) = 1.29$, $p > 0.05$), human evaluators ranked NMT$_m$ as better than NMT$_t$ over 88% of the time, a strong indication that images do help neural MT and bring important information that the multi-modal model NMT$_m$ can efficiently exploit.

If we compare models NMT$_m$ and PBSMT, the latter outperforms the former according to BLEU, METEOR and chrF3, but they are practically equal according to TER. Additionally, the adequacy scores for both these models are, on average, the same according to scores computed over $N = 18$ different human assessments. Nonetheless, even though both models NMT$_m$ and PBSMT are found to produce equally adequate output, translations obtained with PBSMT are ranked best by humans over 56.3% of the time, while translations obtained with the multi-modal model NMT$_m$ are ranked best 24.8% of the time, as can be seen in Figure 2.

We stress that the multi-modal model NMT$_m$ consistently outperforms the text-only model NMT$_t$, according to all four automatic metrics used in this work. Translations generated by model NMT$_m$ contain many neologisms, possibly due to training these models using sub-word tokens rather than just words (Sennrich et al., 2016). Some examples are: "sammlerset", "garagenskateboard", "kampffaltschlocker", "schneidsattel" and "oberreceiver". We argue that this generative quality of the NMT models and the data sets evaluated in this work could have made translations more confusing for native German speakers to understand, therefore the preference for the SMT translations.[4]

We note that the pairwise inter-annotator agreement for the ranking task shows a *fair* agreement among the annotators ($\kappa = 0.30$), computed using Cohen's kappa coefficient (Cohen, 1960). For all the other evaluations, according to Landis and Koch (1977) the pairwise inter-annotator agreement can be interpreted as *slight* ($\kappa = 0.15$ for the multi-modal translation adequacy). The lower agreement score seems plausible since our annotators were crowdsourced and so had limited guidelines and less training for the tasks that would have been ideal.

---

[3]We specifically compute character 6-gram F3.

[4]The SMT model was trained on words directly and therefore does not present these issues.

Model	BLEU4↑	METEOR↑	TER↓	chrF3↑	Adequacy↓
NMT$_t$	22.5	40.0	58.0	56.7	$2.71 \pm .48$
NMT$_m$	25.1†	42.6†	**55.5**†	58.6	**2.36** $\pm$ .47
PBSMT	**27.4**†‡	**45.8**†‡	**55.4**†	**61.6**	**2.36** $\pm$ .47

Table 3: Adequacy of translations and four automatic metrics on eBay's test set: BLEU, METEOR, TER and chrF3. For the first three metrics, results are significantly better than those of NMT$_t$ (†) or NMT$_m$ (‡) with $p < 0.01$.

6 Related work

Multi-modal MT has just recently been addressed by the MT community in a shared task (Specia et al., 2016), where many different groups proposed techniques for multi-modal translation using different combinations of NMT and SMT models (Caglayan et al., 2016; Calixto et al., 2016; Huang et al., 2016; Libovický et al., 2016; Shah et al., 2016). In the multi-modal translation task, participants are asked to train models to translate image descriptions from one natural language into another, while also taking the image itself into consideration. This effectively bridges the gap between two well-established tasks: image description generation (IDG) and MT.

There is an important body of research conducted in IDG. We highlight the work of Vinyals et al. (2015), who proposed an influential neural IDG model based on the sequence-to-sequence framework. They used global visual features to initialise an RNN LM decoder, used to generate the image descriptions in a target language, word by word. In contrast, Xu et al. (2015) were among the first to propose 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. In their model, local visual features were used instead. In both cases, as well as in this work and in most of the state-of-the-art models in the field, models transferred learning from CNNs pre-trained for image classification on ImageNet (Russakovsky et al., 2015).

In NMT, Bahdanau et al. (2015) was the first to propose to use an attention mechanism in the decoder. Their decoder learns to attend to the relevant source-language words as it generates a sentence in the target language, again word by word. Since then, many authors have proposed different ways to incorporate attention into MT. Luong et al. (2015) proposed among other things a local attention mechanism that was less costly than the original global attention; Firat et al. (2016) proposed a model to translate from many source and into many target languages, which involved a shared attention mechanism strategy; Tu et al. (2016) proposed an attention coverage strategy, so that

the model has explicit information from which source words are used to generate previous target words, and therefore addressed the problems of over- and under-translation.

Calixto et al. (2017b) has recently reported n-best list re-ranking experiments of e-commerce product listings using multi-modal eBay data. Whereas their focus is on improving translation quality with n-best list re-ranking experiments, in this work our focus is on the human evaluation of translations generated with the different text-only and multi-modal models. To the best of our knowledge, along with Calixto et al. (2017b) 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. Images bring important information to NMT models in this context; in fact, translations obtained with a multi-modal NMT model are preferred to ones obtained with a text-only model over 88% of the time. Nevertheless, humans still prefer phrase-based SMT over NMT output in this use-case. We attribute this to the nature of the task: listing titles have little syntactic structure and yet many rare words, which can produce many confusing neologisms especially if using subword units.

The core neural MT models still have to be improved significantly to address these challenges. However, in contrast to SMT, they already provide an effective way of improving MT quality with information contained in images. As future work, we will study the impact that additional back-translated data have on multi-modal NMT models.

Acknowledgements

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.*

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3:1137–1155.

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*, pages 257–267, Austin, Texas, USA.

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'12)*, 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.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017a. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. *CoRR*, abs/1702.01287.

Iacer Calixto, Daniel Stein, Evgeny Matusov, Pintu Lohar, Sheila Castilho, and Andy Way. 2017b. Using images to improve machine-translating e-commerce product listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017*, Valencia, Spain (Paper Accepted).

Kyunghyun Cho, Bart van Merrienboer, 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.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

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*, pages 376–380, Baltimore, Maryland, USA.

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. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 770–778, Las Vegas, NV, USA.

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.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 181–184, Detroit, Michigan.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

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*, 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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August.

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, WMT 2016*, pages 543–553, Berlin, Germany.

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, USA.

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.