

CUNI System for WMT16 Automatic Post-Editing and Multimodal Translation Tasks

Jindřich Libovický Jindřich Helcl
Marek Tlustý Ondřej Bojar Pavel Pecina

Charles University in Prague
Malostranské náměstí 25, 112 00 Prague, Czech Republic
{libovicky, helcl, tlusty, bojar, pecina}@ufal.mff.cuni.cz

Abstract

Neural sequence to sequence learning recently became a very promising paradigm in machine translation, achieving competitive results with statistical phrase-based systems. In this system description paper, we attempt to utilize several recently published methods used for neural sequential learning in order to build systems for WMT 2016 shared tasks of Automatic Post-Editing and Multimodal Machine Translation.

1 Introduction

Neural sequence to sequence models are currently used for variety of tasks in Natural Language Processing including machine translation (Sutskever et al., 2014; Bahdanau et al., 2014), text summarization (Rush et al., 2015), natural language generation (Wen et al., 2015), and others. This was enabled by the capability of recurrent neural networks to model temporal structure in data, including the long-distance dependencies in case of gated networks (Hochreiter and Schmidhuber, 1997; Cho et al., 2014).

The deep learning models' ability to learn a dense representation of the input in the form of a real-valued vector recently allowed researchers to combine machine vision and natural language processing into tasks believed to be extremely difficult only few years ago. The distributed representations of words, sentences and images can be understood as a kind of common data type for language and images within the models. This is then used in tasks like automatic image captioning (Vinyals et al., 2015; Xu et al., 2015), visual question answering (Antol et al., 2015) or in attempts to ground lexical semantics in vision (Kiela and Clark, 2015).

In this system description paper, we bring a summary of the Recurrent Neural Network (RNN)-based system we have submitted to the automatic post-editing task and to the multimodal translation task. Section 2 describes the architecture of the networks we have used. Section 3 summarizes related work on the task of automatic post-editing of machine translation output and describes our submission to the Workshop of Machine Translation (WMT) competition. In a similar fashion, Section 4 refers to the task of multimodal translation. Conclusions and ideas for further work are given in Section 5.

2 Model Description

We use the neural translation model with attention (Bahdanau et al., 2014) and extend it to include multiple encoders, see Figure 1 for an illustration. Each input sentence enters the system simultaneously in several representations \mathbf{x}_i . An encoder used for the i -th representation $\mathbf{X}_i = (x_i^1, \dots, x_i^k)$ of k words, each stored as a one-hot vector x_i^j , is a bidirectional RNN implementing a function

$$f(\mathbf{X}_i) = \mathbf{H}_i = (h_i^1, \dots, h_i^k) \quad (1)$$

where the states h_i^j are concatenations of the outputs of the forward and backward networks after processing the j -th token in the respective order.

The initial state of the decoder is computed as a weighted combination of the encoders' final states.

The decoder is an RNN which receives an embedding of the previously produced word as an input in every time step together with the hidden state from the previous time step. The RNN's output is then used to compute the attention and the next word distribution.

The attention is computed over each encoder separately as described by Bahdanau et al. (2014). The attention vector a_i^m of the i -th encoder in the

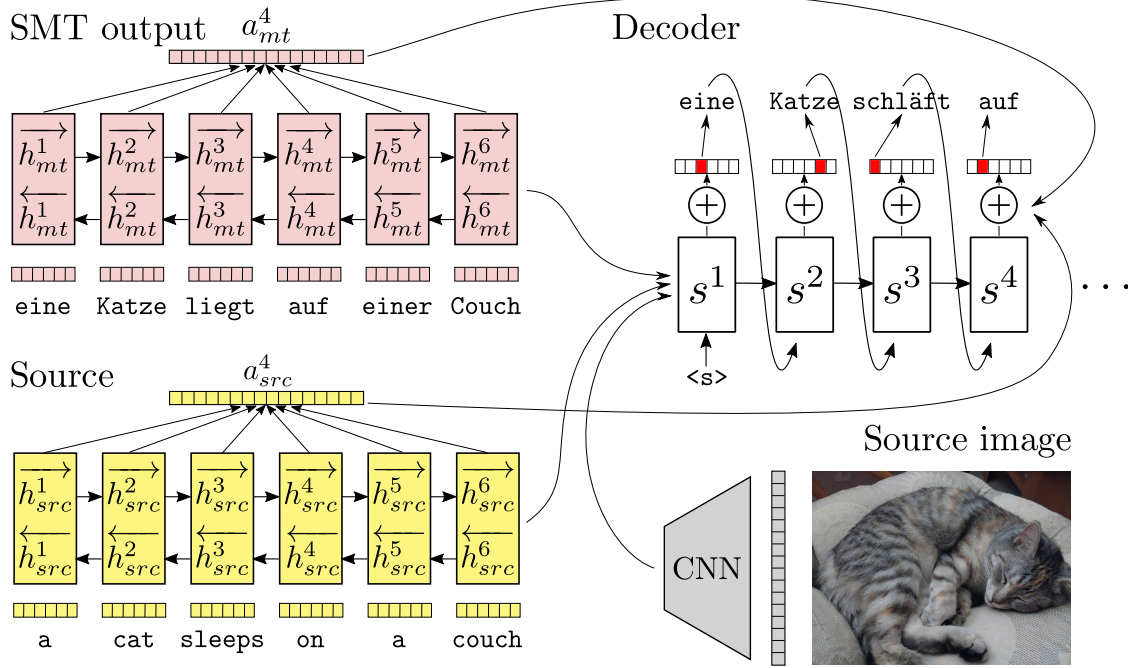


Figure 1: Multi-encoder architecture used for the multimodal translation.

m -th step of the decoder is

$$a_i^m = \sum_{h_i^k \text{ in } \mathbf{H}_i} h_i^k \alpha_i^{k,m} \quad (2)$$

where the weights α_i^m is a distribution estimated as

$$\alpha_i^m = \text{softmax}(v^T \cdot \tanh(s^m + W_{\mathbf{H}_i} \mathbf{H}_i)) \quad (3)$$

with s^m being the hidden state of the decoder in time m . Vector v and matrix $W_{\mathbf{H}_i}$ are learned parameters for projecting the encoder states.

The probability of the decoder emitting the word y_m in the j -th step, denoted as $P(y_m | \mathbf{H}_1, \dots, \mathbf{H}_n, \mathbf{Y}_{0..m-1})$, is proportional to

$$\exp\left(W_o s^j + \sum_{i=1}^n W_{a_i} a_i^j\right) \quad (4)$$

where \mathbf{H}_i are hidden states from the i -th encoder and $\mathbf{Y}_{0..m-1}$ is the already decoded target sentence (represented as matrix, one-hot vector for each produced word). Matrices W_o and W_{a_i} are learned parameters; W_o determines the recurrent dependence on the decoder's state and W_{a_i} determine the dependence on the (attention-weighted) encoders' states.

For image captioning, we do not use the attention model because of its high computational demands and rely on the basic model by Vinyals

et al. (2015) instead. We use Gated Recurrent Units (Cho et al., 2014) and apply the dropout of 0.5 on the inputs and the outputs of the recurrent layers (Zaremba et al., 2014) and L2 regularization of 10^{-8} on all parameters. The decoding is done using a beam search of width 10. Both the decoders and encoders have hidden states of 500 neurons, word embeddings have the dimension of 300. The model is optimized using the Adam optimizer (Kingma and Ba, 2014) with learning rate of 10^{-3} .

We experimented with recently published improvements of neural sequence to sequence learning: scheduled sampling (Bengio et al., 2015), noisy activation function (Gülçehre et al., 2016), linguistic coverage model (Tu et al., 2016). None of them were able to improve the systems' performance, so we do not include them in our submissions.

Since the target language for both the task was German, we also did language dependent pre- and post-processing of the text. For the training we split the contracted prepositions and articles ($am \leftrightarrow an \text{ dem}$, $zur \leftrightarrow zu \text{ der}$, ...) and separated some pronouns from their case ending ($keinem \leftrightarrow kein \text{-em}$, $unserer \leftrightarrow unser \text{-er}$, ...). We also tried splitting compound nouns into smaller units, but on the relatively small data sets we have worked with, it did not bring any improvement.

3 Automatic Post-Editing

The task of automatic post-editing (APE) aims at improving the quality of a machine translation system treated as black box. The input of an APE system is a pair of sentences – the original input sentence in the source language and the translation generated by the machine translation (MT) system. This scheme allows to use any MT system without any prior knowledge of the system itself. The goal of this task is to perform automatic corrections on the translated sentences and generate a better translation (using the source sentence as an additional source of information).

For the APE task, the organizers provided tokenized data from the IT domain (Turchi et al., 2016). The training data consist of 12,000 triplets of the source sentence, its automatic translation and a reference sentence. The reference sentences are manually post-edited automatic translations. Additional 1,000 sentences were provided for validation, and another 2,000 sentences for final evaluation. Throughout the paper, we report scores on the validation set; reference sentences for final evaluation were not released for obvious reasons.

The performance of the systems is measured using Translation Error Rate (Snover et al., 2006) from the manually post-edited sentences. We thus call the score HTER. This means that the goal of the task is more to simulate manual post-editing, rather than to reconstruct the original unknown reference sentence.

3.1 Related Work

In the previous year’s competition (Bojar et al., 2015), most of the systems were based on the phrase-base statistical machine translation (SMT) in a monolingual setting (Simard et al., 2007).

There were also several rule-based post-editing systems benefiting from the fact that errors introduced by statistical and rule-based systems are of a different type (Rosa, 2014; Mohaghegh et al., 2013).

Although the use of neural sequential model is very straightforward in this case, to the best of our knowledge, there have not been experiments with RNNs for this task.

3.2 Experiments & Results

The input sentence is fed to our system in a form of multiple input sequences without explicitly telling which sentence is the source one and which one

method	HTER	BLEU
baseline	.2481	62.29
edit operations	.2438	62.70
edit operations+	.2436	62.62

Table 1: Results of experiments on the APE task on the validation data. The ‘+’ sign indicates the additional regular-expression rules – the system that has been submitted.

is the MT output. It is up to the network to discover their best use when producing the (single) target sequence. The initial experiments showed that the network struggles to learn that one of the source sequences is almost correct (even if it shares the vocabulary and word embeddings with the expected target sequence). Instead, the network seemed to learn to paraphrase the input.

To make the network focus more on editing of the source sentence instead of preserving the meaning of the sentences, we represented the target sentence as a minimum-length sequence of edit operations needed to turn the machine-translated sentence into the reference post-edit. We extended the vocabulary by two special tokens *keep* and *delete* and then encoded the reference as a sequence of *keep*, *delete* and *insert* operations with the insert operation defined by the placing the word itself. See Figure 2 for an example.

After applying the generated edit operations on the machine-translated sentences in the test phase, we perform a few rule-based orthographic fixes for punctuation. The performance of the system is given in Table 1. The system was able to slightly improve upon the baseline (keeping the translation as it is) in both the HTER and BLEU score. The system was able to deal very well with the frequent error of keeping a word from the source in the translated sentence. Although neural sequential models usually learn the basic output structure very quickly, in this case it made a lot of errors in pairing parentheses correctly. We ascribe this to the edit-operation notation which obfuscated the basic orthographic patterns in the target sentences.

4 Multimodal Translation

The goal of the multimodal translation task is to generate an image caption in a target language (German) given the image itself and one or more captions in the source language (English).

Source	Choose Uncached Refresh from the Histogram panel menu.
MT	Wählen ₁ Sie ₂ Uncached ₃ ” ₄ Aktualisieren ₅ ” ₆ aus ₇ dem ₈ Menü ₉ des ₁₀ Histogrammbedienfeldes ₁₁ . ₁₅
Reference	Wählen ₁ Sie ₂ ” ₄ Nicht ₁₂ gespeicherte ₁₃ aktualisieren ₁₃ ” ₆ aus ₇ dem ₈ Menü ₉ des ₁₀ Histogrammbedienfeldes ₁₁ . ₁₅
Edit ops.	keep ₁ keep ₂ delete ₃ keep ₄ Nicht ₁₂ gespeicherte ₁₃ aktualisieren ₁₃ delete ₅ keep ₆ keep ₇ keep ₈ keep ₉ keep ₁₀ keep ₁₁ keep ₁₅

Figure 2: An example of the sequence of edit operations that our system should learn to produce when given the candidate MT translation. The colors and subscripts denote the alignment between the edit operations and the machine-translated and post-edited sentence.

Recent experiments of Elliott et al. (2015) showed that including the information from the images can help disambiguate the source-language captions.

The participants were provided with the Multi30k dataset (Elliott et al., 2016) which is an extension of the Flickr30k dataset (Plummer et al., 2015). In the original dataset, 31,014 images were taken from the users collections on the image hosting service Flickr. Each of the images were given five independent crowd-sourced captions in English. For the Multi30k dataset, one of the English captions for each image was translated into German and five other independent German captions were provided. The data are split into a training set of 29,000 images, a validation set of 1,014 images and a test set with 1,000 images.

The two ways in which the image annotation were collected also lead to two sub-tasks. The first one is called Multimodal Translation and its goal is to generate a translation of an image caption to the target language given the caption in source language and the image itself. The second task is the Cross-Lingual Image Captioning. In this setting, the system is provided five captions in the source language and it should generate one caption in target language given both source-language captions and the image itself. Both tasks are evaluated using the BLEU (Papineni et al., 2002) score and METEOR score (Denkowski and Lavie, 2011). The translation task is evaluated against a single reference sentence which is the direct human translation of the source sentence. The cross-lingual captioning task is evaluated against the five reference captions in the target language created independently of the source captions.

4.1 Related Work

The state-of-the-art image caption generators use a remarkable property of the Convolutional Neural Network (CNN) models originally designed for ImageNet classification to capture the semantic features of the images. Although the images in ImageNet (Deng et al., 2009; Russakovsky et al., 2015) always contain a single object to classify, the networks manage to learn a representation that is usable in many other cases including image captioning which usually concerns multiple objects in the image and also needs to describe complex actions and spacial and temporal relations within the image.

Prior to CNN models, image classification used to be based on finding some visual primitives in the image and transcribing automatically estimated relations between the primitives. Soon after Kiros et al. (2014) showed that the CNN features could be used in a neural language model, Vinyals et al. (2015) developed a model that used an RNN decoder known from neural MT for generating captions from the image features instead of the vector encoding the source sentence. Xu et al. (2015) later even improved the model by adapting the soft alignment model (Bahdanau et al., 2014) nowadays known as the attention model. Since then, these models have become a benchmark for works trying to improve neural sequence to sequence models (Bengio et al., 2015; Gülçehre et al., 2016; Ranzato et al., 2015).

4.2 Phrase-Based System

For the translation task, we trained Moses SMT (Koehn et al., 2007) with additional language models based on coarse bitoken classes. We follow the approach of Stewart et al. (2014): Based on the word alignment, each target word

system	Multimodal translation		Cross-lingual captioning	
	BLEU	METEOR	BLEU	METEOR
Moses baseline	32.2	54.4	11.3	33.8
MM baseline		27.2		32.6
tuned Moses	36.8	57.4	12.3	35.0
NMT	37.1	54.6	13.6	34.6
NMT + Moses	36.5	54.3	13.7	35.1
NMT + image	34.0	51.6	13.3	34.4
NMT + Moses + image	37.3	55.2	13.6	34.9
— ” —, submitted	31.9	49.6	13.0	33.5
captioning only			9.1	25.3
5 en captions			22.7	38.5
5 en captions + image			24.6	39.3
— ” —, submitted			14.0	31.6

Table 2: Results of experiments with the multimodal translation task on the validation data. At the time of the submission, the models were not tuned as well as our final models. The first six system are targeted for the translation task. They were trained against one reference – a German translation of one English caption. The last four systems are target to the cross-lingual captioning task. They were trained with 5 independent German captions (5 times bigger data).

is concatenated with its aligned source word into one bitoken (e.g. “Katze-cat”). For unaligned target words, we create a bitoken with NULL as the source word (e.g. “wird-NULL”). Unaligned source words are dropped. For more than one-to-one alignments, we join all aligned word pairs into one bitoken (e.g. “hat-had+gehabt-had”). These word-level bitokens are afterwards clustered into coarse classes (Brown et al., 1992) and a standard n -gram language model is trained on these classes. Following the notation of Stewart et al. (2014), “400bi” indicates a LM trained on 400 bitoken classes, “200bi” stands for 200 bitoken classes, etc. Besides bitokens based on aligned words, we also use class-level bitokens. For example “(200,400)” means that we clustered source words into 200 classes and target words into 400 classes and only then used the alignment to extract bitokens of these coarser words. The last type is “100bi(200,400)”, a combination of both independent clustering in the source and target “(200,400)” and the bitoken clustering “100bi”.

Altogether, we tried 26 configurations combining various coarse language models. The best three were “200bi” (a single bitoken LM), “200bi&(1600,200)&100tgt” (three LMs, each with its own weight, where 100tgt means a language model over 100 word classes trained on the target side only) and “200bi&100tgt”.

Manual inspection of these three best configurations reveals almost no differences; often the outputs are identical. Comparing to the baseline (a single word-based LM), it is evident that coarse models prefer to ensure agreement and are much more likely to allow for a different word or preposition choice to satisfy the agreement.

4.3 Neural System

For the multimodal translation task, we combine the RNN encoders with image features. The image features are extracted from the 4096-dimensional penultimate layer ($fc7$) of the VGG-16 Imagenet network Simonyan and Zisserman (2014) before applying non-linearity. We keep the weights of the convolutional network fixed during the training. We do not use attention over the image features, so the image information is fed to the network only via the initial state.

We also try a system combination and add an encoder for the phrase-based output. The SMT encoder shares the vocabulary and word embeddings with the decoder. For the combination with SMT output, we experimented with the CopyNet architecture (Gu et al., 2016) and with encoding the sequence the way as in the APE task (see Section 3.2). Since neither of these variations seems to have any effect on the performance, we report only the results of the simple encoder combina-



Source	A group of men are loading cotton onto a truck	
Reference	Eine Gruppe von Männern lädt Baumwolle auf einen Lastwagen	
Moses	eine Gruppe von Männern lädt <u>cotton</u> auf einen <u>Lkw</u>	
<i>2 Errors:</i>	<i>untranslated "cotton" and capitalization of "LKW"</i>	
MMMT	Eine Gruppe von Männern lädt <u>etwas</u> auf einem <u>Lkw</u> .	
<i>Gloss:</i>	<i>A group of men are loading something onto a truck.</i>	
CLC	Mehrere Personen stehen an einem LKW.	
<i>Gloss:</i>	<i>More persons stand on a truck.</i>	
Source	A man sleeping in a green room on a couch.	
Reference	Ein Mann schläft in einem grünen Raum auf einem Sofa.	
Moses	Ein Mann schläft in einem grünen Raum auf einem Sofa.	
MMMT	Ein Mann schläft in einem grünen Raum auf einer Couch.	
	<i>No error, a correctly used synonym for "couch".</i>	
CLC	Eine Frau schläft auf einer Couch.	
	<i>A man ("Mann") is mistaken for a woman ("Frau").</i>	

Figure 3: Sample outputs of our multimodal translation (MMMT) system and cross-lingual captioning (CLC) system in comparison with phrase-based MT and the reference. The *MMMT* system refers to the ‘NMT + Moses + image’ row and *CLC* system to the ‘5 captions + image’ row in Table 2.

tion.

Systems targeted for the multimodal translation task have a single English caption (and eventually its SMT and the image representation) on its input and produce a single sentence which is a translation of the original caption. Every input appears exactly once in the training data paired with exactly one target sentence. On the other hand, systems targeted for the cross-lingual captioning use all five reference sentences as a target, i.e. every input is present five times in the training data with five different target sentences, which are all independent captions in German. In case of the cross-lingual captioning, we use five parallel encoders sharing all weights combined with the image features in the initial state.

Results of the experiments with different input combinations are summarized in the next section.

4.4 Results

The results of both the tasks are given in Table 2. Our system significantly improved since the competition submission, therefore we report both the performance of the current system and of the submitted systems. Examples of the system output can be found in Figure 3.

The best performance has been achieved by the neural system that combined all available input both for the multimodal translation and cross-lingual captioning. Although, using the image as the only source of information led to poor results, adding the image information helped to improve

the performance in both tasks. This supports the hypothesis that for the translation of an image caption, knowing the image can add substantial piece of information.

The system for cross-lingual captioning tended to generate very short descriptions, which were usually true statements about the images, but the sentences were often too general or missing important information. We also needed to truncate the vocabulary which brought out-of-vocabulary tokens to the system output. Unlike the translation task where the vocabulary size was around 20,000 different forms for both languages, having 5 source and 5 reference sentences increased the vocabulary size more than twice.

Similarly to the automatic postediting task, we were not able to come up with a setting where the combination with the phrase-based system would improve over the very strong Moses system with bitoken-classes language model. We can therefore hypothesize that the weakest point of the models is the weighted combination of the inputs for the initial state of the decoder. The difficulty of learning relatively big combination weighting matrices which are used just once during the model execution (unlike the recurrent connections having approximately the same number of parameters) probably over-weighted the benefits of having more information on the input. In case of system combination, more careful exploration of explicit copy mechanism as CopyNet (Gu et al., 2016) may be useful.

5 Conclusion

We applied state-of-the-art neural machine translation models to two WMT shared tasks. We showed that neural sequential models could be successfully applied to the APE task. We also showed that information from the image can significantly help while producing a translation of an image caption. Still, with the limited amount of data provided, the neural system performed comparably to a very well tuned SMT system.

There is still a big room for improvement of the performance using model ensembles or recently introduced techniques for neural sequence to sequence learning. An extensive hyper-parameter testing could be also helpful.

Acknowledgment

We would like to thank Tomáš Musil, Milan Straka and Ondřej Dušek for discussing the problem with us and countless tips they gave us during our work.

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 645 452 (QT21) and no. 644 753 (KConnect) and the Czech Science Foundation (grant n. P103/12/G084). Computational resources were provided by the CESNET LM2015042, provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures.”

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.
- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam M. Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems, NIPS*.
- 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*. Association for Computational Linguistics, Lisbon, Portugal, pages 1–46.
- Peter F. Brown, Peter V. de Souza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18(4):467–479.
- 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-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, Doha, Qatar, pages 103–111.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pages 248–255.
- Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, UK, pages 85–91.
- D. Elliott, S. Frank, K. Sima’an, and L. Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. *ArXiv e-prints*.
- Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR* abs/1510.04709.
- Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *CoRR* abs/1603.06393.
- Çağlar Gülçehre, Marcin Moczulski, Misha Denil, and Yoshua Bengio. 2016. Noisy activation functions. *CoRR* abs/1603.00391.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997.

- Long short-term memory. *Neural computation* 9(8):1735–1780.
- Douwe Kiela and Stephen Clark. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*. Lisbon, Portugal, pages 2461–2470.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.
- Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. Beijing, China, pages 595–603.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, Prague, Czech Republic, pages 177–180.
- Mahsa Mohaghegh, Abdolhossein Sarrafzadeh, and Mehdi Mohammadi. 2013. A three-layer architecture for automatic post editing system using rule-based paradigm. In *Proceedings of the 4th Workshop on South and Southeast Asian Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya Congress Center, Nagoya, Japan, pages 17–24.
- 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*. Association for Computational Linguistics, Philadelphia, PA, USA, pages 311–318.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*. pages 2641–2649.
- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *CoRR* abs/1511.06732.
- Rudolf Rosa. 2014. Depfix, a tool for automatic rule-based post-editing of SMT. *The Prague Bulletin of Mathematical Linguistics* 102:47–56.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 379–389.
- 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. *International Journal of Computer Vision* 115(3):211–252.
- Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, Czech Republic, pages 203–206.
- Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings AMTA*. pages 223–231.
- Darlene Stewart, Roland Kuhn, Eric Joanis, and George Foster. 2014. Coarse split and lump bilingual language models for richer source information in SMT. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas (AMTA)*. Vancouver, BC, Canada, volume 1, pages 28–41.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neu-

- ral networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 3104–3112.
- Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Coverage-based neural machine translation. *CoRR* abs/1601.04811.
- Marco Turchi, Rajen Chatterjee, and Matteo Negri. 2016. WMT16 APE shared task data. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.
- 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, Boston, MA, USA, June 7-12, 2015*. pages 3156–3164.
- Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1711–1721.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. 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 2015*. Lille, France, pages 2048–2057.
- Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *CoRR* abs/1409.2329.