Interactive-Predictive Neural Multimodal Systems

Álvaro Peris(✉) and Francisco Casacuberta(✉)

Pattern Recognition and Human Language Technology Research Center,
Universitat Politècnica de València, Valencia, Spain
{lvapeab,fcn}@prhlt.upv.es

Abstract. Despite the advances achieved by neural models in sequence to sequence learning, exploited in a variety of tasks, they still make errors. In many use cases, these are corrected by a human expert in a posterior revision process. The interactive-predictive framework aims to minimize the human effort spent on this process by considering partial corrections for iteratively refining the hypothesis. In this work, we generalize the interactive-predictive approach, typically applied in to machine translation field, to tackle other multimodal problems namely, image and video captioning. We study the application of this framework to multimodal neural sequence to sequence models. We show that, following this framework, we approximately halve the effort spent for correcting the outputs generated by the automatic systems. Moreover, we deploy our systems in a publicly accessible demonstration, that allows to better understand the behavior of the interactive-predictive framework.

Keywords: Interactive-predictive pattern recognition · Multimodal sequence to sequence learning · Deep learning

1 Introduction

The automatic prediction of structured objects is an extensively studied topic within the pattern recognition field. Many tasks involve the generation of a structured output, given an input object. As structure we understand a dependency across the elements of the object. Typical structured objects include sequences, trees or graphs. The application of neural networks to these problems has recently brought impressive advances. If both input and output objects are sequences, this problem is referred as sequence to sequence learning [9]. Many tasks can be posed as a sequence to sequence problem: machine translation [30],

The research leading to these results has received funding from MINECO under grant IDIFEDER/2018/025 "Sistemas de fabricación inteligentes para la industria 4.0", action co-funded by the European Regional Development Fund 2014–2020 (FEDER), and from the European Commission under grant H2020, reference 825111 (Deep-Health). We also acknowledge NVIDIA Corporation for the donation of GPUs used in this work.

A. Morales et al. (Eds.): IbPRIA 2019, LNCS 11867, pp. 16–28, 2019.
https://doi.org/10.1007/978-3-030-31332-6_2

speech recognition [5] or the automatic description of visual content, known as captioning [37,38].

Notwithstanding the important breakthroughs achieved in the last years, these automatic systems are far from being error-free [17]. However, they are useful for providing initial predictions, which are revised and corrected by a human expert. In some industries, such as machine translation, this revision procedure is widely used, as it increases the productivity with respect to performing the task from scratch [13]. This process is known as translation post-editing.

Nevertheless, this correction process can be improved in several ways. Aiming to increase the productivity of the system and seeking for a symbiotic human–computer collaboration, the so-called interactive-predictive pattern recognition was developed [3,8]. Under this paradigm, the user introduces a correction to the system prediction. Next, the system reacts to this feedback, offering a new prediction, expected to be better than the previous one, as the system has more information.

This interactive-predictive paradigm, initially devised for machine translation, can be extended to several tasks and technologies. In this work, we explore the application of this framework to several scenarios, which include data source from multiple modalities. In a nutshell, our main contributions are:

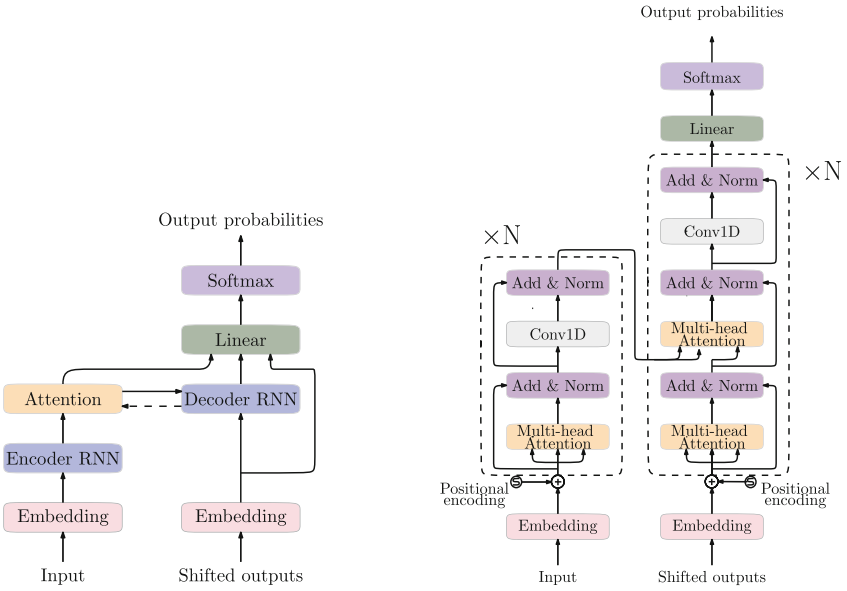
- We successfully apply the interactive-predictive protocol to the automatic captioning of image and videos and to the machine translation post-editing, using neural sequence to sequence models. To the best of our knowledge, this is the first work that delves into this topic.
- We conduct experiments on several datasets, using two common neural architectures: a recurrent neural network (RNN) with attention and a Transformer model.
- We deploy our system in a freely accessible demonstration website.
- We release all the code developed in this work, fostering the research on this topic.

The rest of the manuscript is structured as follows: in Sect. 2 we introduce the neural sequence to sequence modeling. Moreover, we describe the interactive-predictive pattern recognition framework and its implementation with neural models. Next, Sect. 2.2 details the experimental setup followed for assessing our systems. The evaluation and discussion of such systems are shown in Sect. 3. Section 4 reviews the related work. Finally, in Sect. 5 we extract conclusions and set the basis of future works.

2 Interactive-Predictive Multimodal Pattern Recognition

The pattern recognition discipline consists in automatically obtaining a prediction $\hat{\mathbf{y}}$, given an input object $\mathbf{x}$. A common approach to pattern recognition is based its statistical formalization. Following this probabilistic framework, the goal is to obtain the most likely prediction, given the input object:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \Pr(\mathbf{y} \mid \mathbf{x}) \tag{1}$$



(a) RNN-based encoder–decoder with attention.

(b) Transformer model, as illustrated by [34].

Fig. 1. Different architectures for sequence to sequence learning: RNN-based (left) and Transformer models (right). Both models have the same inputs and outputs and differ on the mechanisms applied for learning their representations. In the first case, the input sequence is analyzed by an encoder RNN. The output sequence is generated, word by word, by another RNN. Both RNNs are connected through an attention mechanism. In the case of the Transformer model, the encoder and the decoder are stacks of multi-head attention mechanisms that compute different representations of the inputs. Both models have a vocabulary-sized output layer with a softmax activation, that computes a probability distribution over the output vocabulary.

Since the true probability distribution is unknown, it is approximated by a model with parameters $\boldsymbol{\Theta}$. Therefore, the prediction is given according to this model:

$$\hat{\mathbf{y}} \approx \arg\max_{\mathbf{y}} \; \mathrm{p}(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\Theta}) \tag{2}$$

As aforementioned in the previous section, we are interested in the case in which both $\mathbf{x}$ and $\mathbf{y}$ are sequences. In the last years, and framed into the resurgence of neural networks, $\boldsymbol{\Theta}$ has been frequently implemented as a (deep) neural network, yielding the so-called neural sequence to sequence modeling. This neural network is usually trained on an end-to-end manner on large datasets, via stochastic gradient descent. Moreover, since performing a complete search is prohibitively expensive, the arg max is solved by applying a heuristic search method, typically, beam search [30].

2.1 Neural Architectures for Multimodal Sequence to Sequence Learning

Most neural models for sequence to sequence learning rely on the encoder–decoder paradigm: first, a neural encoder computes a representation of the input sequence. Next, a neural decoder takes this representation is then generates, element by element, the output sequence. Alternative architectures for encoder and decoder have been proposed in the literature. The most popular among them are those based on RNNs with attention [2] or those based solely on attention mechanisms [34] (the so-called Transformer models). Figure 1 depicts a schematic view of these systems. However, providing an in-depth review of these models is out of the scope of this paper. Hence, we refer the reader to the original works for a detailed explanation of these architectures.

This encoder–decoder paradigm can be applied to sequences from arbitrary sources. The only requirement is that we need to encode the input object into a low dimensional, real-valued representation. In this work, we focus on objects from three different sources: text, images and video. Hence, before being introduced to the encoder–decoder system, we need to compute an adequate representation of them. In the computer vision field, this process is known as feature extraction. Depending on the modality of the input object, we thus apply a different feature extractor:

Text: each word is mapped to a continuous representation by using an embedding matrix [30]. Hence, the sequence of input words is converted to a sequence of word embeddings. The embedding matrix is usually estimated with the rest of the parameters of the model.

Images: convolutional neural networks (ConvNets, [20]) excel in several computer vision tasks [18]. These models are also powerful feature extractors. We process the image with a ConvNet and use as features the final representation computed by the ConvNet that preserves positional information. A complete image is thus seen as a sequence of image crops. Hence, we can directly apply the sequence to sequence framework, as done by Xu et al. [37].

Videos: A video is a sequence of images. Therefore, we also rely on the usage of ConvNet for extracting the features from the each video frame. For alleviating the computational overload, we compute global features for each video image. In addition, we subsample the frames introduced to the system [38], also for reducing the computational load.

2.2 Interactive-Predictive Pattern Recognition

As discussed in the previous section, in an interactive-predictive scenario, the user introduces corrections to the predictions generated by a pattern recognition system. This correction is introduced as a feedback signal f. The systems reacts then to the introduction of the feedback, producing an alternative

hypothesis, compatible with f. Considering this, the interactive-predictive framework rewrites Eq. (2) for also taking into account the user feedback signal:

$$\tilde{\mathbf{y}} = \underset{\mathbf{y} \text{ compatible with } f}{\arg\max} \; \mathrm{p}(\mathbf{y} \mid \mathbf{x}, f; \boldsymbol{\Theta}) \tag{3}$$

Hence, the goal of an interactive-predictive system is to generate the most likely prediction that is compatible with the feedback provided by the user. Depending on the meaning conveyed by f, alternative interactive protocols can be defined. In this work, we follow the prefix-based interactive protocol. We also assume that the user introduces the corrections using a keyboard and a mouse.

The prefix-based protocol arguably is the most natural way of work. In this protocol, the user searches, from the left to the right, for the first error in the prediction given by the system and introduces the correct character. This feedback signal conveys a two-fold meaning: on the one hand, it states a correct character at a given position. On the other hand, it also validates the hypothesis up to this position. Taking this into account, a prefix-based interactive-predictive system must generate the most likely suffix, to a prefix validated by the user [3].

The implementation of this protocol in neural sequence to sequence systems requires to constrain the search [26]: the system applies a forced decoding of the feedback provided by the user. The suffix is obtained then by applying a regular search. For introducing corrections at a character level, we apply a vocabulary mask as described by [25], which ensures that the next word generated complies with the user feedback.

We evaluate our interactive-predictive framework in six different scenarios, involving three tasks and two different datasets per task. The main figures of the datasets are shown in Table 1. The tasks under study are:

Machine translation: translation of English sentences to French, on two datasets[1]: UFAL and Europarl. The first one belongs to a medical domain and the latter refers to the translation of the proceedings from the European parliament.

Image captioning: we tackled two common datasets: Flickr8k [12] and Flickr30k [27]. The goal is to generate descriptions of pictures crawled from Flickr users.

Video captioning: we tested our systems on the popular Microsoft Research Video Description (MSVD) dataset [6], a general task, relating the description of YouTube videos from multiple domains. In addition, we apply our methods to the EDUB-SegDesc dataset [4], a collection of egocentric videos and first person captions.

[1] Datasets available at: http://statmt.org/wmt18.

Table 1. Figures of the different datasets. M denotes millions of elements. The column #References indicates the number of different references per sample. * denotes a variable number of references. In this case, we report the average references per sample.

Task	Dataset	#Samples			#References
		Training	Validation	Test	
Machine translation	UFAL	$2.8M$	$1,000$	$1,000$	1
	Europarl	$2.0M$	$3,003$	$3,000$	1
Image captioning	Flickr8k	$30,000$	$1,000$	$1,000$	5
	Flickr30k	$145,000$	$1,014$	$1,000$	5
Video captioning	MSVD	$48,779$	100	670	$41^\star$
	EDUB-SegDesc	$2,652$	204	246	3

2.3 Evaluation Metrics

We evaluate two main aspects of our systems. On the one hand, we measure the quality of the initial predictions provided by the system. This is the most common scenario in the literature. This evaluation is carried on by comparing the predictions with the ground-truth references from each dataset. The final goal of these metrics is to correlate with the human perception of prediction quality. The metrics range from 0 (worst quality) to 100 (best quality):

BLEU [22]**:** Computes the geometric mean of the n-gram precision of prediction and references. In includes n-grams from order 1 to 4. It also includes a penalty for short predictions.
METEOR [19]**:** Computes the F1 score of precision and recall of matches between prediction and references words. To this end, it applies linguistic resources such as stemmers, paraphrase and synonym dictionaries.

On the other hand, under an interactive-predictive framework, our objective is to reduced the amount of effort spent by the user during the correction process. We follow the literature and estimate this effort as the number of keystrokes and mouse actions performed by the user during the correction process. To this end, we rely on two metrics:

CharacTER [36]**:** Translation edit rate computed at a character level: minimum number of character edit operations (insertion, substitution, deletion and swapping) that must be made in order to transform the hypothesis into the reference. The number of edit operations is normalized by the number of characters.
KSMR [3]**:** accounts for the number of keystrokes plus mouse actions involved in the interactive correction process, divided by the number of characters of the final prediction obtained.

CharacTER and KSMR are error-based metrics, hence the lower, the better. Following Zaidan et al. [39], CharacTER is an estimate of the effort of static post-edition; while the effort of interactive-predictive systems can be assessed via KSMR [3].

2.4 Usage of the System and User Simulation

Using an interactive-predictive system requires to follow the procedure described in Sect. 2: the process starts with an automatic prediction given by the system to an input object. The user then reviews the prediction, starting and the interactive-predictive process: the user searches in this hypothesis the first error, and introduces a correction. The system then reacts, providing an alternative hypothesis, considering the user feedback. This protocol is repeated until the user finds satisfactory the hypothesis given by the system. We implemented a live demonstration of this system[2].

Properly assessing interactive-predictive systems involves the experimentation with human users, which is prohibitively expensive. Hence, during the development of such systems, it is common to rely on simulated users [3,26]. We used the ground-truth samples from the different datasets as the desired outputs by our simulated users. The simulation is done by correcting the leftmost wrong character of each hypothesis from the interactive-predictive system, until reaching the desired output.

2.5 Description of the Systems

Our neural sequence to sequence systems[3] were developed with NMT-Keras [24]. This library is built upon Keras[4] and works for the Theano and Tensorflow backends. For each task and dataset, we built two models: one using RNNs with attention and another one using a Transformer architecture.

The RNN-based systems had long short-term memory units [11]. Encoder and decoder were bridged together through an additive attention mechanism [2]. We set all model dimensions to the same value. In the case of machine translation, all layers had a dimension of 512. In the case of image and video captioning, we reduced the model size to 256, since we are dealing with smaller datasets.

In the case of the Transformer models, we set two stacks of 6 layers for the encoder and the decoder. In the case of machine translation, all model dimensions were 512 and the number of attention heads was 8. This configuration is the same as the *base* model described by Vaswani et al. [34]. For the captioning tasks, we reduced again our model, to 256 dimensions on each layer.

Machine translation and image captioning systems were trained using Adam [15], with a learning rate of 0.0002. In the case of video captioning, we obtained better performance using Adadelta [40], in both datasets. In all cases, the batch

[2] Accessible at http://casmacat.prhlt.upv.es/interactive-seq2seq.
[3] Source code: https://github.com/lvapeab/interactive-keras-captioning.
[4] https://keras.io.

Table 2. Prediction quality for the different tasks, datasets and models. The RNN column denotes RNN-based system (Fig. 1a) and the Trans. column indicates a Transformer model (Fig. 1b)

Task	Dataset	BLEU [↑]		METEOR [↑]	
		RNN	Trans.	RNN	Trans.
Machine translation	UFAL	37.2	37.8	59.6	60.4
	Europarl	24.6	26.6	45.7	47.9
Image captioning	Flickr8k	22.1	19.6	20.8	19.8
	Flickr30k	22.2	19.3	20.0	18.5
Video captioning	MSVD	49.6	45.7	33.4	30.7
	EDUB-SegDesc	30.4	25.8	21.9	20.3

size was 64. During training, we applied an early-stopping strategy, watching the BLEU on the development set. At decoding time, we used a beam size of 6.

In the case of machine translation, the word embeddings were randomly initialized and learned together with the rest of the parameters of the system. In the case of image captioning, we extracted image features using a NASNet architecture [41], trained on the ImageNet dataset [7]. The video features were extracted with an Inception v4 network [31], also trained on the ImageNet dataset. Following Yao et al. [38], we subsampled the frames from a video, selecting 26 images per clip. Image and video feature remained static along the training process of the sequence to sequence model.

3 Results and Discussion

We show and discuss now the results obtained by our systems. First, we will assess the systems quantitatively, in terms of prediction quality and effort required during the correction stage. Next, in order to gain some insights into the behavior of the system, we analyze an image captioning example.

3.1 Quantitative Evaluation

We start by evaluating the systems in a traditional way, assessing their prediction quality. Table 2 shows the BLEU and METEOR results of the different systems for all tasks. These results are similar to those reported in the literature for each task and dataset [4,25,37,38].

It is worth to note that the Transformer model only outperformed the RNN-based systems in the case of machine translation. This model is more data-eager than RNN systems. Many of the recent advances yielded with this architecture leverage huge data collections (e.g. Radford et al. [29]). We also contrasted this fact in our experimentation: the machine translation datasets were way larger than the captioning ones (see Table 1. Hence, the Transformer model only was fully exploited in the machine translation case.

Next, we evaluate the performance of the interactive-predictive systems. To that end, we estimate the effort required for correcting the output of a static system (using CharacTER) and the effort needed by a interactive system (using KSMR). These results are shown in Table 3. The results obtained in machine translation are similar to the literature [25]. Due to the novelty of this scenario, we lack from references in the literature, regarding the other tasks.

Table 3. Effort required for correcting the outputs of static (St.) and interactive-predictive (Int.) systems, using RNN and Transformer (Trans.) models. The effort of static systems is measured in terms of CharacTER while the effort required by interactive-predictive systems is evaluated in terms of KSMR

Task	Dataset	CharacTER [↓]		KSMR [↓]	
		St. RNN	St. Trans.	Int. RNN	Int. Trans.
Machine translation	UFAL	35.7	36.5	19.0	15.9
	Europarl	53.6	51.2	30.1	29.4
Image captioning	Flickr8k	77.8	79.6	36.6	36.9
	Flickr30k	81.7	86.1	36.0	40.0
Video captioning	MSVD	58.1	64.1	36.4	40.5
	EDUB-SegDesc	72.3	71.4	40.0	38.0

Interactive-predictive systems approximately halved the amount of corrections required for correcting their outputs, with respect to traditional, static systems. The results were consistent across all tasks and for all models. Hence, these results indicate that the interactive protocol effectively achieved its goal of reducing the correction effort.

Moreover, a crucial aspect of the usability of interactive systems is their response time. Hence, it is important to keep it in adequate values. The average response time of our systems was always below 0.2 s. This provides the user of a feeling of almost instant reactivity [21].

Finally, we are aware that properly assessing the usability and effort reduction brought by these system requires a human evaluation on its usage. In this paper, we set the first step toward future developments on multimodal neural interactive-predictive pattern recognition, with positive initial results.

3.2 Qualitative Analysis and Discussion

We show and analyze an image captioning example. Other examples for the machine translation and video captioning tasks are alike. The example is taken from our multimodal showcase and shown in Fig. 2.

We can see that the caption generated by the system (at iteration 0) has an error. The user wants to indicate that the people are sitting on a bench. Hence, the feedback introduced is the character “b”. The system is able to properly

complete the word "bench", with this single interaction. The same happens when the user wants to introduce the clause "under a". With only typing the character "u", the system generates this clause. Finally, it is interesting to observe the behavior of the last interaction. The user introduced the character "n" to the word "a". Hence, the next word must start with a vowel. The system is able to properly account for this concordance and generates the word "umbrella". We observe that the systems also handle correctly other concordances, such as singular/plural clauses.

4 Related Work

Neural sequence to sequence learning has been a widely studied topic since its reintroduction, framed to the deep learning era [9,30]. As stated above, neural

Iter 0	**System**	A group of people sit on a ramp.
Iter 1	**User**	*A group of people sit on a* **b**ramp.
	System	*A group of people sit on a b*ench.
Iter 2	**User**	*A group of people sit on a bench* **u**.
	System	*A group of people sit on a bench u*nder a building.
Iter 3	**User**	*A group of people sit on a bench under a* **n**building.
	System	*A group of people sit on a bench under an* umbrella.
Iter 4	**User**	*A group of people sit on a bench under an umbrella.*

Fig. 2. Interactive-predictive session example, for correcting the caption generated for the image. At each iteration, the user introduces a character correction (boxed). The system modifies its hypothesis, taking into account this feedback: keeping the correct prefix (green) and generating a compatible suffix. Post-editing this sample in a static way, would have required the deletion of 4 characters and the addition of 23 characters (Color figure online)

machine translation [2,34] has meant a revolution in the field. Nowadays, these systems are standard in research and industry. In addition to machine translation, different tasks have been tackled following this approach: speech recognition [5], speech translation [14], syntactic parsing [35], or the already discussed image and video captioning [23,37,38].

Regarding the interactive-predictive pattern recognition framework, it has been mainly applied to machine translation. The addition of interactive protocols for fostering the productivity of translation environments have been studied for long time, for phrase-based models [3,10] and neural machine translation systems [16,26].

The interactive-predictive approach has been also previously generalized for tackling other tasks, involving multimodal signals. This is the case of the interactive transcription of handwritten text documents [32], layout detection [28], among others [33]. None of these works however, involved fully end-to-end neural multimodal systems.

5 Conclusions and Future Work

In this work, we empirically demonstrated the capabilities of the interactive-predictive framework applied to multimodal, neural sequence-to-sequence systems. We tackled a variety of tasks, using two state-of-the-art models and, in all cases, the interactive-predictive systems were able to decrease the human effort required for correcting the outputs of the system. We obtained savings of approximately a 50%. We also analyzed these systems through an online demo website. We released all source code developed.

These encouraging results open several avenues for future research. The construction of multimodal, interactive-predictive systems allow the application of this framework to other structured prediction tasks, e.g. tables to text. More precisely, this framework is directly applicable to the automatic report of medical images or to the automatic generation of life-loggers. In addition to an end application, these tools can be used by human annotators, for creating datasets on a more efficient way.

Moreover, we experimented with multimodal inputs. In a future, we want to explore the inclusion of multimodal feedback signals. This was already done for statistical models [1] and we think that neural models are able to exploit this very effectively. In addition, we used a different system for each task. In a future, we would like to explore the construction of a single multitask, multimodal system. The recent advances achieved in multitask learning [29] heavily support this research direction. Finally, for properly assessing the efficiency of this framework, we should conduct and experimentation involving human users.

References

1. Alabau, V., Sanchis, A., Casacuberta, F.: Improving on-line handwritten recognition in interactive machine translation. Pattern Recognit. **47**(3), 1217–1228 (2014)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2015). arXiv:1409.0473
3. Barrachina, S., et al.: Statistical approaches to computer-assisted translation. Comput. Linguist. **35**(1), 3–28 (2009)
4. Bolaños, M., Peris, Á., Casacuberta, F., Soler, S., Radeva, P.: Egocentric video description based on temporally-linked sequences. J. Vis. Commun. Image Represent. **50**, 205–216 (2018)
5. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: Proceedings of the ICASSP, pp. 4960–4964 (2016)
6. Chen, D.L., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the ACL, pp. 190–200 (2011)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the CVPR, pp. 248–255 (2009)
8. Foster, G., Isabelle, P., Plamondon, P.: Target-text mediated interactive machine translation. Mach. Transl. **12**, 175–194 (1997)
9. Graves, A.: Sequence transduction with recurrent neural networks (2012). arXiv:1211.3711
10. Green, S., Chuang, J., Heer, J., Manning, C.D.: Predictive translation memory: a mixed-initiative system for human language translation. In: Proceedings of the ACM UIST, pp. 177–187 (2014)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
12. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. J. Artif. Intell. Res. **47**, 853–899 (2013)
13. Hu, K., Cadwell, P.: A comparative study of post-editing guidelines. In: Proceedings of the EAMT, pp. 34206–353 (2016)
14. Jia, Y., et al.: Direct speech-to-speech translation with a sequence-to-sequence model (2019). arXiv:1904.06037
15. Kingma, D., Ba, J.: Adam: a method for stochastic optimization (2014). arXiv:1412.6980
16. Knowles, R., Koehn, P.: Neural interactive translation prediction. In: Proceedings of the AMTA, pp. 107–120 (2016)
17. Koehn, P., Knowles, R.: Six challenges for neural machine translation. In: Proceedings of the First Workshop on NMT, pp. 28–39 (2017)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of NIPS, pp. 1097–1105 (2012)
19. Lavie, A., Denkowski, M.J.: The METEOR metric for automatic evaluation of machine translation. Mach. Transl. **23**(2–3), 105–115 (2009)
20. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
21. Nielsen, J.: Usability Engineering. Morgan Kaufmann Publishers Inc., Burlington (1993)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the ACL, pp. 311–318 (2002)

23. Peris, Á., Bolaños, M., Radeva, P., Casacuberta, F.: Video description using bidirectional recurrent neural networks. In: Proceedings of the ICANN, pp. 3–11 (2016)
24. Peris, A., Casacuberta, F.: NMT-Keras: a very flexible toolkit with a focus on interactive NMT and online learning. Prague Bull. Math. Linguist. **111**, 113–124 (2018)
25. Peris, Á., Casacuberta, F.: Online learning for effort reduction in interactive neural machine translation. Comput. Speech Lang. **58**, 98–126 (2019)
26. Peris, Á., Domingo, M., Casacuberta, F.: Interactive neural machine translation. Comput. Speech Lang. **45**, 201–220 (2017)
27. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the ICCV, pp. 2641–2649 (2015)
28. Quirós, L., Martínez-Hinarejos, C.-D., Toselli, A.H., Vidal, E.: Interactive layout detection. In: Alexandre, L.A., Salvador Sánchez, J., Rodrigues, J.M.F. (eds.) IbPRIA 2017. LNCS, vol. 10255, pp. 161–168. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58838-4_18
29. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. Technical report, Open-AI (2019)
30. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the NIPS, vol. 27, pp. 3104–3112 (2014)
31. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the CVPR, pp. 2818–2826 (2016)
32. Toselli, A., Romero, V., Rodríguez, L., Vidal, E.: Computer assisted transcription of handwritten text images. In: Proceedings of the ICDAR, vol. 2, pp. 944–948 (2007)
33. Toselli, A.H., Vidal, E., Casacuberta, F.: Multimodal Interactive Pattern Recognition and Applications. Springer, Heidelberg (2011). https://doi.org/10.1007/978-0-85729-479-1
34. Vaswani, A., et al.: Attention is all you need. In: Proceedings of NIPS, pp. 5998–6008 (2017)
35. Vinyals, O., Kaiser, Ł., Koo, T., Petrov, S., Sutskever, I., Hinton, G.: Grammar as a foreign language. In: Proceedings of NIPS, pp. 2755–2763 (2015)
36. Wang, W., Peter, J.T., Rosendahl, H., Ney, H.: CharacTer: translation edit rate on character level. In: Proceedings of the WMT, vol. 2, pp. 505–510 (2016)
37. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the ICML, pp. 2048–2057 (2015)
38. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Describing videos by exploiting temporal structure. In: Proceedings of the ICCV, pp. 4507–4515 (2015)
39. Zaidan, O.F., Callison-Burch, C.: Predicting human-targeted translation edit rate via untrained human annotators. In: Proceedings of the NAACL, pp. 369–372 (2010)
40. Zeiler, M.D.: ADADELTA: an adaptive learning rate method (2012). arXiv:1212.5701
41. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceedings of the CVPR, pp. 8697–8710 (2018)