Supplementary Material for "Reliability and Learnability of Human Bandit Feedback for Sequence-to-Sequence Reinforcement Learning"

Julia Kreutzer[1] and **Joshua Uyheng**[3]* and **Stefan Riezler**[1,2]
[1]Computational Linguistics & [2]IWR, Heidelberg University, Germany
{kreutzer,riezler}@cl.uni-heidelberg.de
[3]Departments of Psychology & Mathematics, Ateneo de Manila University, Philippines
juyheng@ateneo.edu

A Rating Task

A.1 Rating Instructions

Participants for the 5-star rating task were given the following instructions: "You will be presented with a German statement and a translation of this statement in English. You must assign a rating from 1 (Very Bad) to 5 (Very Good) to each translation."

Participants for the pairwise task were given the following instructions: "You will be presented with a German statement and two translations of this statement in English. You must decide which of the two translations you prefer, or whether you have no preference."

A.2 Example Ratings

Table 1 lists low- and high-variance items for 5-star ratings, Table 2 for pairwise ratings. From the annotations in the tables, the reader may get an impression which translations are "easier" to judge than others.

B Reward Estimation

B.1 Auxiliary Data for Reward Estimation

In order to augment the small collection of 1,000 rated translations, we leverage the available out-of-domain bitext as auxiliary training data: 10k source sentences of WMT (out-of-domain) are translated by the out-of-domain model. Translations from 9 beam search ranks are compared to their references to compute sBLEU rewards. This auxiliary data hence provides 90k out-of-domain training samples with sBLEU reward. For pairwise rewards, sBLEU scores for two translations for the same source are compared. Each mini-batch during training is sampled from the auxiliary data with probability p_{aux}, from the original training data with probability $1 - p_{aux}$. Adding this auxiliary data as a regularization through multi-task learning prevents the model from overfitting to the small set of human ratings. In our experiments, $p_{aux} = 0.8$ worked best.

B.2 Reward Estimation Architecture

Input source and target sequence are split into the BPE subwords used for NMT training, padded up to a maximum length of 100 tokens, and represented as 500-dimensional subword embeddings. Subword embeddings are pre-trained on the WMT bitext with `word2vec` (Mikolov et al., 2013), normalized to unit length and held constant during further training. Additional 10-dimensional BPE-feature embeddings are appended to the subword embeddings, where a binary indicator encodes whether each subword contains the subword prefix marker "@@". BPE-prefix features are useful information for the model since bad translations can arise from "illegal" compositions of subword tokens. The embeddings are then fed to a ssource-side and a target-side bidirectional LSTM (biLSTM) (Hochreiter and Schmidhuber, 1997), respectively. The biLSTM outputs are concatenated for each time step and fed to a 1-D convolutional layer with 50 filters each for filter sizes from 2 to 15. The convolution is followed by max-over-time pooling, producing 700 input features for a fully-connected output layer with leaky ReLU (Maas et al., 2013) activation function. Dropout (Srivastava et al., 2014) with $p = 0.5$ is applied before the final layer. This architecture can be seen as a biLSTM-enhanced bilingual extension to the convolutional model for sentence classification proposed by Kim (2014).

*The work for this paper was done while the second author was an intern in Heidelberg.

#1	source	Diese könnten Kurierdienste sein, oder Techniker zum Beispiel, nur um sicherzustellen, dass der gemeldete AED sich immer noch an seiner Stelle befindet.
	target	These could be courier services, or technicians like, for example, just to make sure that the abalone aed is still in its place.
	rating	$\sigma = 0.46, \varnothing = -0.30$
#2	source	Es muss für mich im Hier und Jetzt stimmig sein, sonst kann ich mein Publikum nicht davon überzeugen, dass das mein Anliegen ist.
	target	It must be for me here and now, otherwise i cannot convince my audience that my concern is.
	rating	$\sigma = 0.46, \varnothing = -0.70$
#3	source	Aber wenn Sie biologischen Evolution akzeptieren, bedenken Sie folgendes: ist es nur über die Vergangenheit, oder geht es auch um die Zukunft?
	target	But if you accept biological evolution, consider this: Is it just about the past, or is it about the future?
	rating	$\sigma = 0.48, \varnothing = 1.12$
#4	source	Finden Sie heraus, wie Sie überleben würden. Die meisten unserer Spieler haben die im Spiel gelernten Gewohnheiten beibehalten.
	target	Find out how you would survive.
	rating	$\sigma = 1.31, \varnothing = -0.79$
#5	source	Sie können das googlen, aber es ist keine Infektion des Rachens sondern der oberen Atemwege und verursacht den Verschluss der Atemwege.
	target	You can googlen, but it's not an infection of the rag, but the upper respiratory pathway, and it causes respiratory traction.
	rating	$\sigma = 1.31, \varnothing = -0.52$
#6	source	Nun, es scheint mir, dieses Thema wird, oder sollte wenigstens die interessanteste politische Debatte zum Verfolgen sein über die nächsten paar Jahre.
	target	Well, it seems to me that this issue is going to be, or should be at least the most interesting political debate about the next few years.
	rating	$\sigma = 1.25, \varnothing = -0.93$

Table 1: Items with lowest (top) and highest (bottom) deviation in 5-star ratings. Mean normalized rating and standard deviation are reported. Problematic parts of source and target are underlined, namely hallucinated or inadequate target words (#1, #5, #6), over-literal translations (#2), ungrammatical source (#3, #6) and omissions (#4).

#1	source	Zu diesem Zeitpunkt haben wir mehrzellige Gemeinschaften, Gemeinschaften von vielen verschiedlichen Zellentypen, welche zusammen als einzelner Organismus fungieren.
	target1	At this time we have multi-tent communities, communities of many different cell types, which act together as individual organism.
	target2	At this point, we have multicellular communities, communities of many different cell types, which act together as individual organism.
	rating	$\sigma = 0.0, \varnothing = 1.0$
#2	source	Wir durchgehen dieselben Stufen, welche Mehrzellerorganismen durchgemacht haben – Die Abstraktion unserer Methoden, wie wir Daten festhalten, präsentieren, verarbeiten.
	target1	We pass the same steps that have passed through multi-cell organisms to process the abstraction of our methods, how we record data.
	target2	We go through the same steps that multicellular organisms have gone through – the abstraction of our methods of holding data, representing, processing.
	rating	$\sigma = 0.0, \varnothing = 1.0$
#3	source	Ich hielt meinen üblichen Vortrag, und danach sah sie mich an und sagte: "Mhmm. Mhmm. Mhmm."
	target1	I thought my usual talk, and then she looked at me and said: mhmm.
	target2	I gave my usual talk, and then she looked at me and said, "mhmm. Mhmm. Mhmm."
	rating	$\sigma = 0.0, \varnothing = 1.0$
#4	source	So in diesen Plänen, wir hatten ungefähr 657 Plänen die den Menschen irgendetwas zwischen zwei bis 59 verschiedenen Fonds anboten.
	target1	So in these plans, we had about 657 plans that offered the people something between two to 59 different funds.
	target2	So in these plans, we had about 657 plans that offered people anything between two to 59 different funds.
	rating	$\sigma = 0.99, \varnothing = 0.14$
#5	source	Wir fingen dann an, über Musik zu sprechen, angefangen von Bach über Beethoven, Brahms, Bruckner und all die anderen Bs, von Bartók bis hin zu Esa-Pekka Salonen.
	target1	We then began to talk about music, starting from bach on Beethoven, Brahms, Bruckner and all the other bs, from Bartók to esa-pekka salons.
	target2	We started talking about music from bach, Beethoven, Brahms, Bruckner and all the other bs, from Bartok to esa-pekka salons.
	rating	$\sigma = 0.99, \varnothing = -0.14$
#6	source	Heinrich muss auf all dies warten, nicht weil er tatsächlich ein anderes biologische Alter hat, nur aufgrund des Zeitpunktes seiner Geburt.
	target1	Heinrich has to wait for all of this, not because he's actually having another biological age, just because of the time of his birth.
	target2	Heinrich must wait for all this, not because he actually has another biological age, only due to the time of his birth.
	rating	$\sigma = 0.99, \varnothing = -0.14$

Table 2: Items with lowest (top) and highest (bottom) deviation in pairwise ratings. Preferences of target1 are treated as "-1"-ratings, preferences of target2 as "1", no preference as "0", so that a mean ratings of e.g. -0.14 expresses a slight preference of target1. Problematic parts of source and targets are underlined, namely hallucinated or inadequate target words (#1, #2, #3, #4), incorrect target logic (#2), omissions (#3), ungrammatical source (#4), capitalization (#5), over-literal translations (#5, #6).

C NMT

C.1 NMT Hyperparameters

The NMT has a bidirectional encoder and a single-layer decoder with 1,024 GRUs each, and subword embeddings of size 500 for a shared vocabulary of subwords obtained from 30k byte-pair merges (Sennrich et al., 2016). Maximum input and output sequence length are set to 60. For the MLE training of the out-of-domain model, we optimize the parameters with Adam ($\alpha = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) (Kingma and Ba, 2014). For further in-domain tuning (supervised, OPL and RL), α is reduced to 10^{-5}. To prevent the models from overfitting, dropout with probability 0.2 (Srivastava et al., 2014) and l2-regularization with weight 10^{-8} are applied during training. The gradient is clipped to its norm when its norm exceeds 1.0 (Pascanu et al., 2013). Early stopping points are determined on the respective development sets. For model selection we use greedy decoding, for test set evaluation beam search with a beam of width 10. For MLE and OPL models, mini-batches of size 60 are used. For the RL models, we reduce the batch size to 20 to fit $k = 5$ samples for each source into memory. The temperature is furthermore set to $\tau = 0.5$. We found that learning rate and temperature were the most critical hyperparameters and tuned both on the development set.

References

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. Atlanta, GA, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*. Atlanta, GA, USA.

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 (ACL)*. Berlin, Germany.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15:1929–1958.