Leveraging Discourse Rewards for Document-Level Neural Machine Translation

Inigo Jauregi Unanue[1,3], Nazanin Esmaili[1], Gholamreza Haffari[2], Massimo Piccardi[1]
[1]University of Technology Sydney, NSW 2007, Australia
[2]Monash University, VIC 3800, Australia
[3]RoZetta Technology, NSW 2000, Australia
Inigo.Jauregi@rozettatechnology.com
{Nazanin.Esmaili,Massimo.Piccardi}@uts.edu.au
Gholamreza.Haffari@monash.edu

Abstract

Document-level machine translation focuses on the translation of entire documents from a source to a target language. It is widely regarded as a challenging task since the translation of the individual sentences in the document needs to retain aspects of the discourse at document level. However, document-level translation models are usually not trained to explicitly ensure discourse quality. Therefore, in this paper we propose a training approach that explicitly optimizes two established discourse metrics, *lexical cohesion* (LC) and *coherence* (COH), by using a reinforcement learning objective. Experiments over four different language pairs and three translation domains have shown that our training approach has been able to achieve more cohesive and coherent document translations than other competitive approaches, yet without compromising the faithfulness to the reference translation. In the case of the Zh-En language pair, our method has achieved an improvement of 2.46 percentage points (pp) in LC and 1.17 pp in COH over the runner-up, while at the same time improving 0.63 pp in BLEU score and 0.47 pp in $\mathrm{F_{BERT}}$.

1 Introduction

The recent advances in neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017) have provided the research community and the commercial landscape with effective translation models that can at times achieve near-human performance. However, this usually holds at phrase or sentence level. When using these models in larger units of text, such as paragraphs or documents, the quality of the translation may drop considerably in terms of discourse attributes such as lexical and stylistic consistency.

In fact, document-level translation is still a very open and challenging problem. The sentences that make up a document are not unrelated pieces of text that can be predicted independently; rather, a set of sequences linked together by complex underlying linguistics aspects, also known as the discourse (Maruf et al., 2019b; Jurafsky and Martin, 2019). The discourse of a document includes several properties such as grammatical cohesion (Halliday and Hasan, 2014), lexical cohesion (Halliday and Hasan, 2014), document coherence (Hobbs, 1979) and the use of discourse connectives (Kalajahi et al., 2012). Ensuring that the translation retain such linguistic properties is expected to significantly improve its overall readability and flow.

However, due to the limitations of current decoder technology, NMT models are still bound to translate at sentence level. In order to capture the discourse properties of the source document in the translation, researchers have attempted to incorporate more contextual information from surrounding sentences. Most document-level NMT approaches augment the model with multiple encoders, extra attention layers and memory caches to encode the surrounding sentences, and leave the model to implicitly learn the discourse attributes by simply minimizing a conventional NLL objective. The hope is that the model will spontaneously identify and retain the discourse patterns within the source document. Conversely, very little work has attempted to model the discourse attributes explicitly. Even the evaluation metrics typically used in translation such as BLEU (Papineni et al., 2002) are not designed to assess the discourse quality of the translated documents.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4467–4482
Barcelona, Spain (Online), December 8-13, 2020

For these reasons, in this paper we propose training an NMT model by directly targeting two specific discourse metrics: *lexical cohesion* (LC) and *coherence* (COH). LC is a measure of the frequency of semantically-similar words co-occurring in a document (or block of sentences) (Halliday and Hasan, 2014). For example, *car*, *vehicle*, *engine* or *wheels* are all semantically-related terms. There is significant empirical evidence that ensuring lexical cohesion in a text eases its understanding (Halliday and Hasan, 2014). At its turn, COH measures how well adjacent sentences in a text are linked to each other. In the following example from Hobbs (1979):

"John took a train from Paris to Istanbul. He likes spinach."

the two sentences make little 'sense' one after another. An incoherent text, even if grammatically and syntactically perfect, is anecdotally very difficult to understand and therefore coherence should be actively pursued. Relevant to translation, Vasconcellos (1989) has found that a high percentage of the human post-editing changes over machine-generated translations involves the improvement of cohesion and coherence.

Several LC and COH metrics that well correlate with the human judgement have been proposed in the literature. However, like BLEU and most other evaluation metrics, they are discrete, non-differentiable functions of the model's parameters. Hereafter, we propose to overcome this limitation by using the well-established *policy gradient* approach from reinforcement learning (Sutton et al., 1999; Sutton and Barto, 2018) which allows using any evaluation metric as a reward without having to differentiate it. By combining different types of rewards, the model can be trained to simultaneously achieve more lexically-cohesive and more coherent document translations, while at the same time retaining faithfulness to the reference translation.

2 Related Work

2.1 Document-level NMT

Many document-level NMT models have proposed taking the context into account by concatenating surrounding sentences or extra features to the current input sentence, with otherwise no modifications to the model. For example, Rios et al. (2017) have trained an NMT model that learns to disambiguate words given the context semantic landscape by simply extracting lexical chains from the source document, and using them as additional features. Other researchers have proposed concatenating previous source and target sentences to the current source sentence, so that the decoder can observe a proper amount of context (Agrawal et al., 2018; Tiedemann and Scherrer, 2017; Scherrer et al., 2019). Their work has shown that concatenating even just one or two previous sentences can result in a noticeable improvement. Macé and Servan (2019) have added an embedding of the entire document to the input, and shown promising results in English-French.

Conversely, other document-level NMT approaches have proposed modifications to the standard encoder-decoder architecture to more effectively account for the context from surrounding sentences. Jean et al. (2017) have introduced a dedicated attention mechanism for the previous source sentences. Multi-encoder approaches with hierarchical attention networks have been proposed to separately encode each of the context sentences before they are merged back into a single context vector in the decoder (Miculicich et al., 2018; Maruf et al., 2019a; Wang et al., 2017). These models have shown significant improvements over sentence-level NMT baselines on many different language pairs. Kuang et al. (2018) and Tu et al. (2018) have proposed using an external cache to store, respectively, a set of topical words or a set of previous hidden vectors. This information has proved to benefit the decoding step at limited additional computational cost. In turn, Maruf and Haffari (2018) have presented a model that incorporates two memory networks, one for the source and one for the target, to capture document-level interdependencies. For the inference stage, they have proposed an iterative decoding algorithm that incrementally refines the predicted translation.

However, all the aforementioned models assume that the model can implicitly learn the occurring discourse patterns. Moreover, the training objective is the standard negative log-likelihood (NLL) loss,

which simply maximizes the probability of the reference target words in the sentence. Only one work these authors are aware of (Xiong et al., 2019) has attempted to train the model by explicitly learning discourse attributes. Inspired by recent work in text generation (Bosselut et al., 2018), Xiong et al. (2019) have proposed automatically learning neural rewards that can encourage translation coherence at document level. However, it is not clear whether the learned rewards would be in good correspondence with human judgment. For this reason, in our work we prefer to rely on established discourse metrics as rewards.

2.2 Discourse evaluation metrics

As a matter of fact, several metrics have been proposed in the literature to measure discourse properties. For LC, Wong and Kit (2012) have proposed a metric that looks for repetitions of words and their related terms (e.g. hyponyms, hypernyms) by using WordNet (Miller, 1998). Gong et al. (2015) have proposed a similar metric that uses lexical chains. For COH, mainly two types of metrics have been proposed: *entity-based* and *topic-based*. The former follow the Centering Theory (Grosz et al., 1995) which states that documents with a high frequency of the same salient entities are more coherent. An entity-based coherence metric was proposed by Barzilay and Lapata (2008). At their turn, *topic-based* metrics assume that a document is coherent when adjacent sentences are similar in topic and vocabulary. Accordingly, Hearst (1997) has proposed the Texttiling algorithm which computes the cosine distance between the bag-of-word (BoW) vectors of adjacent sentences. Foltz et al. (1998) have proposed to replace the BoW vectors with topic vectors. Li et al. (2017) have learned topic embeddings with a self-supervised neural network. There is also a third group of COH metrics that are based solely in syntactic regularities (Smith et al., 2016) that have also shown to be effective at modelling textual coherence. Other metrics have been proposed to measure different discourse properties such as grammatical cohesion (Hardmeier and Federico, 2010; Miculicich and Popescu-Belis, 2017) and discourse connectives (Hajlaoui and Popescu-Belis, 2013).

2.3 Reinforcement learning in NMT

Researchers in NMT and other natural language generation tasks have used reinforcement learning (Sutton and Barto, 2018) techniques to train the models to maximize discrete sentence-level and document-level metrics as an alternative or a complement to the NLL. For example, Ranzato et al. (2016) have proposed training NMT systems targeting the BLEU score, showing consistent improvements with respect to strong baselines. In addition to training the model directly with the evaluation function, they claim that this approach mollifies the exposure bias problem (Bengio et al., 2015). Expected risk minimization has been proposed as an alternative reinforcement learning-style training to maximize the sentence-level (Edunov et al., 2018; Shen et al., 2016) and the document-level (Saunders et al., 2020) BLEU scores. Paulus et al. (2018) have proposed a similar approach for summarization using ROUGE as the training loss (Lin and Hovy, 2000). Tebbifakhr et al. (2019) have used a similar objective function to improve the sentiment classification of translated sentences. Finally, Edunov et al. (2018) have presented a comprehensive comparison of reinforcement learning and structured prediction losses for NMT model training.

3 Baseline Models

This section describes the baseline NMT models used in the experiments. In detail, subsection 3.1 recaps the standard sentence-level translation model while subsection 3.2 describes the recent, strong hierarchical baseline that we have augmented with discourse rewards.

3.1 Sentence-level NMT

Our first baseline is a standard sentence-level NMT model. Given the source document D with k sentences, the model translates each sentence $\mathbf{x}_i = \{\mathbf{x}_i^1, \ldots, \mathbf{x}_i^{n_i}\}, i = 1, \ldots, k$, in the document into a sentence in the target language, $\mathbf{y}_i^\star = \{\mathbf{y}_i^1, \ldots, \mathbf{y}_i^{m_i}\}$:

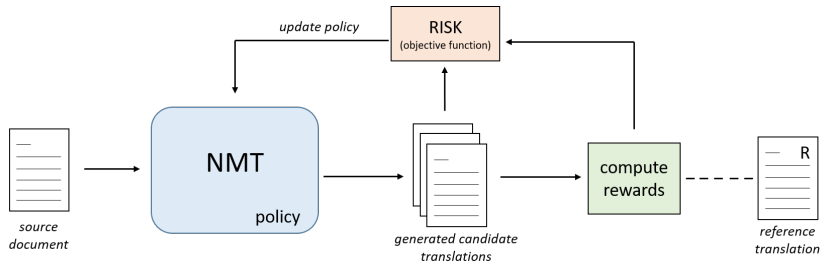

Figure 1: Risk training. Given the source document, the policy (NMT model) predicts l candidate translations. Then, a reward function is computed for each such translation. For supervised rewards, (e.g., BLEU) the reference translation is required, but not for LC and COH. Finally, the Risk loss is computed using the rewards and the probabilities of the candidate translations, differentiated, and backpropagated for parameter update.

$$\mathbf{y}_i^\star = \arg\max_{\mathbf{y}_i} p(\mathbf{y}_i|\mathbf{x}_i, \theta) \quad i = 1, \ldots, k \tag{1}$$

Thus, the model translates every sentence in the document independently. Our sentence model uses a standard transformer-based encoder-decoder architecture (Vaswani et al., 2017) where the model is trained to maximize the probability of the words in the training reference sentences using an NLL objective. We train this model for 20 epochs and select the best model over the validation set. For more details on training and the hyper-parameters please see Appendix A.

3.2 Hierarchical Attention Network

As a document-level translation baseline, we have used the Hierarchical Attention Network (HAN) of Miculicich et al. (2018). A HAN network is added to the sentence-level NMT model both in the encoder and in the decoder (referred to as HAN_{join} in the following), allowing the model to encode information from t previous source and target sentences. The prediction can be expressed as:

$$\mathbf{y}_i^\star = \arg\max_{\mathbf{y}_i} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{x}_{i-1}, \ldots, \mathbf{x}_{i-t}, \mathbf{y}_{i-1}, \ldots, \mathbf{y}_{i-t}, \theta) \quad i = 1, \ldots, k \tag{2}$$

where $(\mathbf{x}_{i-1}, \ldots, \mathbf{x}_{i-t})$ are the previous source sentences and $(\mathbf{y}_{i-1}, \ldots, \mathbf{y}_{i-t})$ the previous target sentences that make up the context. At inference time, the target sentences are the model's own predictions. Following the indications given by the authors, we have set $t = 3$. Additionally, we have used the weights of the sentence-level NMT baseline to initialize the common parameters of the HAN_{join} model, and we have initialized the extra parameters introduced by the HAN networks randomly. The model has been fine-tuned for 10 epochs and the best model over the validation set has been selected. For further information on the hyper-parameters see Appendix A.

4 Risk training with discourse rewards

In order to improve the baseline models, we propose to use the LC (Wong and Kit, 2012) and COH (Foltz et al., 1998) evaluation metrics as rewards during training, so that the model is explicitly rewarded for generating more cohesive and coherent translation at document level. For that, we use a reinforcement learning approach, which allows using discrete, non-differentiable functions as rewards in the objective. Following Edunov et al. (2018), we have used the structured loss that achieved the best results in their experiments, namely the *expected risk minimization (Risk)* objective:

$$\mathcal{L}_{Risk} = \sum_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} -r(\mathbf{u}, \mathbf{y}) p(\mathbf{u}|\mathbf{x}, \theta) \tag{3}$$

where $\mathbf{x}$ is the source sentence, $\mathbf{y}$ is the reference translation, $p(\mathbf{u}|\mathbf{x})$ is the conditional probability of a translation in our 'policy', or NMT model, $\mathcal{U}(x)$ is a set of candidate translations generated by the current policy, and $r(\cdot)$ is the reward function. In our work we have obtained the candidate translations using beam search, which achieved higher accuracy than sampling in Edunov et al. (2018). The conditional probability of a translation has been defined as:

$$
\begin{aligned}
p(\mathbf{u}|\mathbf{x},\theta) &= \frac{f(\mathbf{u},\mathbf{x},\theta)}{\sum_{\mathbf{u'}\in\mathcal{U}(\mathbf{x})} f(\mathbf{u'},\mathbf{x},\theta)} \\
f(\mathbf{u},\mathbf{x},\theta) &= \exp[\frac{1}{m}\sum_{j=1}^{m}\log p(u^j|u^1,\ldots,u^{j-1},\mathbf{x},\theta)]
\end{aligned}
\tag{4}
$$

where m is the number of words in the candidate translation. Note that in order to avoid underflow and put all the sentences on a similar scale, the (unnormalized) sentence score, $f(\mathbf{u},\mathbf{x},\theta)$, in Eq. 4 is computed as a sum of logarithms, divided by the number of tokens in the sequence and, finally, brought back to scale with the exponential function.

By minimizing this Risk objective, the NMT model is encouraged to give higher probability to candidate translations that obtain a higher reward. This function has been used at sentence level by Edunov et al. (2018). However, the same metrics could also be computed at document level by simply concatenating all the sentences from the same document together (both for the ground truth and the predictions). As a result, m now would be the number of words in a document, $\mathcal{U}(\mathbf{x})$ the candidate document translations, $\mathbf{x}$ the source document and $\mathbf{y}$ the reference document. Computing the Risk objective in this way permits having document-level reward functions as $r(\cdot)$.

4.1 Reward functions

We have explored the use and combination of different reward functions for training:

LC$_{\text{doc}}$: For LC, the metric proposed by Wong and Kit (2012) has been adopted. This metrics counts the number of lexical cohesive devices in the document and then divides that number by the total number of words in the document (Eq. 5). Cohesive devices include associations such as repetitions of words, synonyms, near-synonyms, hypernyms, meronyms, troponyms, antonyms, coordinating terms, and so on. WordNet (Fellbaum, 2012) has been used to classify the relationships between words. Note that this reward function is unsupervised since it does not require a ground-truth reference translation.

$$
LC = \frac{\#\ of\ \ cohesion\ \ devices\ \ in\ \ document}{\#\ of\ \ words\ \ in\ \ document} \tag{5}
$$

COH$_{\text{doc}}$: To calculate COH, we have used the approach proposed by Foltz et al. (1998). This approach first uses a trained LSA model to infer topic vectors ($\mathbf{t}_i$) for each sentence in the document, and then computes the average cosine distance between adjacent sentences (Eq. 6). For the topic vectors, we have used the pre-trained LSA model (Wiki-6) from Stefanescu et al. (2014), which was trained over Wikipedia. Note that COH also does not require a ground-truth reference translation.

$$
COH = \frac{1}{k-1}\sum_{i=2}^{k}\cos(\mathbf{t}_i,\mathbf{t}_{i-1}) \tag{6}
$$

BLEU$_{\text{doc}}$: In addition to the LC and COH rewards, we have decided to use a reference-based metric such as BLEU (Papineni et al., 2002). Due to the unsupervised nature of LC and COH, the model could trivially boost them by only repeating words and creating very similar sentences. However, this will come at the expense of producing translations that are increasingly unrelated to the reference translation (low adequacy) and grammatically incorrect (low fluency). As such, we encourage the model to also target a high BLEU score in its predictions.

Language pair	Domain	train	dev	test	Avg. # sent/doc
Zh-En	TEDtalks	0.2M	0.9K	3.9K	122
Cs-En	TEDtalks	0.1M	0.5K	5.2K	114
Es-En	TEDtalks	0.2M	0.8K	4.7K	114
Eu-En	subtitles	0.8M	0.8K	1.5K	1018
Es-En	subtitles	1.1M	1.9K	4.6K	774
Es-En	news	0.2M	2.1K	14K	37

Table 1: The datasets used for the experiments.

BLUE$_{\text{sen}}$: Finally, we have also used BLEU at sentence level as a reward. In this way, we can assess whether it is more beneficial to use this metric at document or sentence level.

These four rewards can be combined in several different ways. To limit the experiments, we have decided to use them in their natural range without reweighting. All the results with the different reward combinations are presented in Section 5.2.

4.2 Mixed objective

Similar to the MIXER training proposed by Ranzato et al. (2016), we have also explored mixing the Risk objective with the NLL. The rationale is similar to that of using BLEU$_{\text{doc}}$ and BLEU$_{\text{sen}}$ as rewards: the NLL loss can help the model to not deviate too much from the reference translation while improving discourse properties. To mix these losses, we have used an alternate batch approach: either loss is randomly selected in each training batch, with a certain probability (e.g. Risk(0.8) means that we have selected the Risk loss with 80% probability and the NLL with 20%).

5 Experiments

5.1 Datasets and experimental setup

We have performed a broad range of experiments over four different language pairs and three different translation domains (TED talks, movie subtitles and news) which have been used by other popular document-level NMT research (Miculicich et al., 2018; Tu et al., 2018). For translations of TED talks[1], we have used the datasets released in the IWSLT14 for Spanish-English (Es-En), in the IWSLT15 shared task for Chinese-English (Zh-En) and in IWSLT16 for Czech-English (Cs-En). For both language pairs, we have used their *dev2010* set as the validation set, and sets *tst2011-2013* (Zh-En), *tst2010-2013* (Cs-En) and *tst2010-2012* (Es-En), respectively, as test sets. For translations in the movie subtitles domain, we have used the OpenSubtitles-v2018 dataset (Lison et al., 2018) from OPUS[2], and the language pairs tested have been Basque-English (Eu-En) and Spanish-English (Es-En). For Eu-En we have used all the available data, but for Es-En we have only used a subset of the corpus to limit time and memory requirements. In both cases, we have divided the data into a training, validation and test sets[3]. The last translation domain is news, for which we have used the Es-En News-Commentary11 dataset[4]. As validation and test sets, we have used its *newstest2008* and *newstest2009-2013* sets, respectively, from WMT[5]. The document boundaries are given by the individual talks for the TED talks dataset, by movie scripts for the subtitles datasets and by single-author news commentaries for the news dataset. All the datasets have been tokenized using the *Moses tokenizer*[6], with the exception of Chinese for which we have used *Jieba*[7]. A *truecaser* model from Moses[7] has been learned over the training data of each dataset, and has been applied for consistent word casing as a final pre-processing step.

[1] https://wit3.fbk.eu/
[2] http://opus.nlpl.eu/
[3] All the datasets will be released publicly, and the reviewers can already see them as supplementary material.
[4] http://www.casmacat.eu/corpus/news-commentary.html
[5] http://www.statmt.org/wmt13/translation-task.html
[6] https://github.com/moses-smt/mosesdecoder
[7] https://github.com/fxsjy/jieba

As models, we have compared multiple models trained with the Risk objective with different combinations of reward functions. This has allowed us to select the best reward functions for the translation quality at document level. Then, the model trained with the best reward combination has been compared against the sentence-level NMT and HAN baselines. In our experiments, the Risk training objective has been used as fine-tuning of a pre-trained HAN_{join} baseline model, in order not to suffer from a "cold start" due to the large output label space. The main aim of our experiments is to show that the proposed training objectives can lead to performance improvements over HAN_{join}. Candidate translations have been obtained using beam search with a beam size of only 2, due to memory and computational time limitations. Furthermore, the training batch size has been set to 15 sentences. Since the objective is computed over the batch, this is equivalent to subdividing longer documents into sub-documents of 15 sentences each. Yet, our experimental results show that computing the rewards at such batch level is still effective for improving the translation quality.

Each model has been trained with three different seeds over its training set, and the validation set has been used at all times to select the best model. Then, the average results of the three runs over the test set have been reported. We have measured four different evaluation metrics: BLEU, LC, COH and F_{BERT}, an alternative metric to BLEU that compares the BERT sentence embeddings of the prediction and the reference and which has been shown to have better correlation with the human judgement than BLEU (Zhang et al., 2020). To select the best model over the validation set for the sentence-level NMT baseline, we have used the lowest perplexity. Instead, for the HAN_{join} baseline and our models, we have chosen the model with the best results in the majority of the four evaluation metrics (BLEU, LC, COH and F_{BERT}). This has not affected the relative ranking of the sentence-level NMT baseline since its performance has been generally lower than the other approaches. Complete details about the experimental set-up and other hyper-parameters are provided in Appendix A. The code is publicly available[8].

5.2 Results

Table 2 shows the main results from our experiments. Over all datasets, the HAN_{join} baseline has consistently outperformed the sentence-level NMT in terms of BLEU score and F_{BERT} which shows that including surrounding sentences can help to obtain better translation accuracy. However, HAN_{join} has not performed significantly better than the sentence-level model in terms of LC and COH (even worse in a few cases), showing that it has not been able to specifically learn these discourse properties in the document. The COH and LC values of both baselines have also been generally lower than those of the human reference translations for all datasets (with the exception of the LC in Zh-En (TED talks) and Es-En (movie subtitles)).

Table 2 also shows the results from our best models in comparison to these baselines. From preliminary experiments, we have seen that the Risk model that achieved the best results is the one that combines BLEU_{doc}, LC_{doc} and COH_{doc} as rewards. Yet, choosing the right proportion of Risk and NLL training has proven very important and dataset-dependent. In the TED talks domain (Table 2a), the Risk(1.0) model has outperformed the HAN_{join} baseline in all evaluated metrics over the Zh-En dataset, improving $+0.63$ percentage points (pp) in BLEU, 2.46 pp in LC, 1.17 pp in COH and 0.48 pp in F_{BERT}, while in the Cs-En dataset the same model has got an improvement of 2.68 pp in LC, 0.55 pp in COH and 0.22 pp in F_{BERT}, on a parity of BLEU score. Instead, over the Es-En dataset, even though the Risk(1.0) has achieved the highest LC and COH scores, this has come at a higher drop in translation accuracy (i.e. BLEU and F_{BERT}). Thus, we consider Risk(0.5) to be the best performing model over this dataset, as it still considerably improves LC and COH scores (1.28 pp and 0.23 pp respectively), while keeping similar translation accuracy in terms of BLEU ($+0.22$ pp) and F_{BERT} (-0.27 pp). In general, we had not anticipated the improvements in BLEU score and F_{BERT} since our main aim had only been to improve the translations in terms of discourse metrics. However, in some cases the improvements in discourse metrics have also translated into higher translation accuracy.

In turn, Table 2b shows the main results over the movie subtitles datasets which are characterized by documents with, on average, more, yet much shorter, sentences than the TED talks. On these datasets, the

[8] https://github.com/ijauregiCMCRC/DL_NMT_RL

Model	Zh-En (TED talks)				Cs-En (TED talks)				Es-En (TED talks)			
	BLEU	LC	COH	F_{BERT}	BLEU	LC	COH	F_{BERT}	BLEU	LC	COH	F_{BERT}
Sentence-level NMT	16.94	55.39	28.02	66.94	22.74	55.62	27.72	69.60	39.55	56.67	28.27	79.5
HAN_{join}	17.52	55.02	28.15	67.21	23.44	55.63	27.62	69.87	39.89	56.25	28.56	**79.88**
Human reference	–	55.13	29.33	–	–	55.91	29.7	–	–	57.84	30.79	–
Risk(1.0)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	**18.15**	**57.48***	**29.32***	67.69	23.40	**58.31***	**28.17**	**70.09**	37.4	59.41†	28.92	78.86
Risk(0.8)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	17.82	55.18	28.68	67.60	23.43	56.03*	27.62	70.01*	39.52	**57.53**	**28.79**	79.11
Risk(0.5)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	17.83	54.70	28.30	**67.73**	23.42	56.07	27.78	69.95*	**40.1**	57.4	28.78	79.61
Risk(0.2)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	17.80	55.10	28.35	67.62	**23.48**	55.85	27.62	69.95	40.07	56.83	28.61	79.62

(a) Results over the TED talks datasets.

Model	Eu-En (movie subtitles)				Es-En (movie subtitles)			
	BLEU	LC	COH	F_{BERT}	BLEU	LC	COH	F_{BERT}
Sentence-level NMT	9.12	37.08	19.34	59.18	29.34	58.31	22.70	67.57
HAN_{join}	9.74	37.19	19.63	59.72	**30.14**	58.11	22.58	**67.73**
Human reference	–	41.83	21.93	–	–	57.28	24	–
Risk(1.0)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	1.19	72.51†	27.67†	36.72	3.37	67.82†	19.53	48.07
Risk(0.8)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	9.67	**40.66***	19.60	**59.76**	29.51	58.34	22.82	67.51
Risk(0.5)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	9.77	38.85*	**19.80**	59.62	29.79	**58.44**	22.76	67.53
Risk(0.2)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	**9.99**	37.53	19.42	59.72	29.70	58.39	**22.96**	67.50

(b) Results over the movie subtitles datasets.

Model	Es-En (news)			
	BLEU	LC	COH	F_{BERT}
Sentence-level NMT	21.79	32.97	28.1	67.88
HAN_{join}	22.16	32.87	28.15	**68.28**
Human reference	–	38.66	30.97	–
Risk(1.0)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	20.67	32.81	28.14	67.84
Risk(0.8)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	22.26	**33.70***	**28.45***	68.14
Risk(0.5)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	22.34	33.51*	28.39	68.02
Risk(0.2)-$BLEU_{doc} + LC_{doc} + COH_{doc}$	**22.45***	33.32*	28.25	68.13

(c) Results over the news datasets.

Table 2: Main results. (*) means that the differences are statistically significant with respect to the HAN_{join} baseline with a p-value < 0.05 over a one-tailed Welch's t-test. ($\dagger$) indicates high LC and COH values that come at the expense of a considerable drop in translation accuracy (e.g. BLEU, F_{BERT}), and thus, likely undesirable.

Risk(1.0) model has been able to improve the LC and COH metrics to a large extent, but at a marked cost in BLEU score and F_{BERT}. Qualitatively, the translations generated by this model have often displayed many word and phrase repetitions that had little correspondence with the reference translation, showing that COH and LC can reach values that are undesirable. Conversely, training the model with the mixed objective has forced it to stay closer to the reference translations and helped it achieve higher BLEU and F_{BERT} scores. On Eu-En, the Risk(0.8) model has improved the LC by 3.47 pp at a substantial parity of all the other metrics. On Es-En, none of the proposed models has clearly outperformed the HAN_{join} baseline. For instance, the Risk(0.5) model has improved LC and COH by 0.33 pp and 0.18 pp, respectively, but at the cost of 0.35 pp in BLEU score and 0.20 pp in F_{BERT}.

Finally, Table 2c shows that the proposed models have delivered better results on the news domain dataset, where they have been able to simultaneously improve the BLEU score, LC and COH at a mild cost in F_{BERT}. In general, we can argue that the discourse rewards have proved more effective on documents such as talks and news commentaries – which come from single authors and are generally controlled in style – than on documents such as subtitles are more fragmented in nature.

5.2.1 Ablation study and translation example

To expand the analysis, Table 3 shows the results from an ablation study that explores the impact of the various reward functions over the Zh-En dataset. The best trade-off over the four evaluation metrics seems that returned by $BLEU_{doc}$ + LC_{doc} + COH_{doc} which has achieved the highest BLEU score, a high F_{BERT}, and high LC and COH. The results also show that using $BLEU_{sen}$ as a reward has contributed to improve the F_{BERT} score in all cases, but at the significant expense of the other evaluation

Model	BLEU	LC	COH	F_{BERT}
$BLEU_{doc} + LC_{doc} + COH_{doc}$	**18.15**	57.48	29.32	67.69
$BLEU_{sen} + LC_{doc} + COH_{doc}$	17.53	56.32	28.79	**67.96**
$BLEU_{doc} + LC_{doc}$	17.44	**59.21**	29.87	67.27
$BLEU_{doc} + COH_{doc}$	17.57	55.74	28.82	67.41
$BLEU_{sen} + LC_{doc}$	17.60	56.32	28.76	67.87
$BLEU_{sen} + COH_{doc}$	17.46	56.31	28.82	67.93
$LC_{doc} + COH_{doc}$	10.56	71.28†	31.25†	62.27
$BLEU_{sen}$	17.42	55.93	28.76	67.83
$BLEU_{doc}$	17.20	54.59	28.15	67.18
LC_{doc}	10.42	71.70†	31.61†	62.09
COH_{doc}	17.26	58.66	**29.98**	66.92

Table 3: Ablation study of the various reward functions over the Zh-En TED talks dataset with Risk(1.0). (†) indicates high LC and COH values that come at the expense of a considerable drop in translation accuracy (e.g. BLEU, F_{BERT}), and thus, likely undesirable.

Src:	. . . 女士们，先生们，见见你的近亲。 这就是野生倭黑猩猩的世界座落于刚果的丛林中 。 倭黑猩猩 和黑猩猩是我们大家生活里最密切相关的近亲。 这意味我们都享有一个共同的祖先，一个进化了的祖母，她生活在大约6 百万年前 。 . . .
Ref:	. . . ladies and gentlemen , meet your cousins . this is the world of wild bonobos in the jungles of Congo . bonobos are , together with chimpanzees , your living closest relative . that means we all share a common ancestor , an evolutionary grandmother , who lived around six million years ago . . .
HAN_{join}:	. . . ladies and gentlemen , meet your relatives . this is the world of the wildlife that is in the Congo . the chimps and chimpanzees are the most closely related to us . it means we all have a common ancestor , a grandmother who has evolved about six million years ago . . .
	BLEU: 32.21 **LC**: 9.52 **COH**: 19.36 **F_{BERT}**: 80.78
Risk(1.0):	. . . ladies and gentlemen , meet your close relatives . and that 's the world of the wild bonobos that are in the jungle in the Congo . the bonobos are the most closely related to the chimpanzees that we live in . it means that we all have a common ancestor , a grandmother who lived about six million years ago . . .
	BLEU: 23.99 **LC**: 20.00 **COH**: 20.77 **F_{BERT}**: 82.83

Table 4: Translation example. Excerpt of a document from the Zh-En TED talks test set.

metrics. However, when $BLEU_{doc}$ and $BLEU_{sen}$ have been compared head-to-head as the sole rewards, the sentence-level BLEU has been able to achieve higher scores in all metrics. In contrast, the $BLEU_{doc}$ reward has been most effective when used jointly with the cohesion and coherence rewards. At its turn, the LC_{doc} reward without a balance from a BLEU reward has led to LC and COH scores that are likely excessive and undesirable, with a corresponding drop in BLEU score and F_{BERT}. Conversely, the COH_{doc} reward has not displayed a comparable degradation. The main overall result from this ablation analysis is that the rewards need to be used in a calibrated combination to deliver the best trade-off across all the evaluation metrics, and that the selection of the best combination can be effectively carried out by validation.

Finally, Table 4 shows an example of the translation of a document excerpt from the Zh-En TED talks dataset made by our best model (Risk(1.0)-$BLEU_{doc} + LC_{doc} + COH_{doc}$), in comparison to that made by the HAN_{join} baseline, the reference translation (Ref) and the text in the source language (Src). In this example, we can clearly see the positive influence of the LC and COH rewards, as the model has been able to provide better lexical cohesion and coherence in the translation. The model has also been able to correctly translate words such as *bonobos* and *jungle* while the HAN_{join} model has uttered a more generic *chimps*. In addition, the translation generated by our model seems more faithful to the reference overall. Note also that these improvements have come at a significant drop in BLEU score. This may suggest that LC and COH can influence improvements that the BLEU score is not able to capture. Examples for the other language pairs are provided in Appendix B.

6 Conclusion

In this paper, we have presented a novel training method for document-level NMT models that uses discourse rewards to encourage the models to generate more lexically cohesive and coherent translations at document level. As training objective we have used a reinforcement learning-style function, named Risk, that permits using discrete, non-differentiable terms in the objective. Our results on four different language pairs and three translation domains have shown that our models have achieved a consistent improvement in discourse metrics such as LC and COH, while retaining comparable values of accuracy metrics such as BLEU and $\mathrm{F_{BERT}}$. In fact, on certain datasets, the models have even improved on those metrics. While the approach has proved effective in most cases, the best combination of discourse rewards, accuracy rewards and NLL has had to be selected by validation for each dataset. In the near future we plan to investigate how to automate this selection, and also explore the applicability of the proposed approach to other natural language generation tasks.

Acknowledgment

The authors would like to thank the RoZetta Institute (formerly CMCRC) for providing financial support to this research. Warmest thanks also go to Dr. Sameen Maruf for her feedback on an early version of this paper.

References

Ruchit Rajeshkumar Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *21st Annual Conference of the European Association for Machine Translation*, pages 11–20.

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)*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1171–1179.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation (WNMT)*, pages 18–27.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Christiane Fellbaum. 2012. Wordnet. *The encyclopedia of applied linguistics*.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics (AISTATS)*, pages 249–256.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2015. Document-level machine translation evaluation with gist consistency and text cohesion. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 33–40.

Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 236–247. Springer.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in english*. Routledge.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289.

Marti A Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive science*, 3(1):67–90.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

D Jurafsky and JH Martin. 2019. Speech and language processing (3rd (draft) ed.).

Seyed Ali Rezvani Kalajahi, Ain Nadzimah Abdullah, Jayakaran Mukundan, and Dan J Tannacito. 2012. Discourse connectors: An overview of the history, definition and classification of the term. *World Applied Sciences Journal*, 19(11):1659–1673.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2018. Modeling coherence for neural machine translation with dynamic and topic caches. In *Proceedings of the 27th Conference on Computational Linguistics (COLING)*.

Jiwei Li and Dan Jurafsky. 2017. Neural net models for open-domain discourse coherence. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics (COLING)*, pages 495–501.

Pierre Lison, Jörg Tiedemann, Milen Kouylekov, et al. 2018. Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Language resources evaluation conference (LREC)*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421.

Valentin Macé and Christophe Servan. 2019. Using whole document context in neural machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting on Association for Computational Linguistics (ACL)*.

Sameen Maruf, André FT Martins, and Gholamreza Haffari. 2019a. Selective attention for context-aware neural machine translation. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2019b. A survey on document-level machine translation: Methods and evaluation. *arXiv preprint arXiv:1912.08494*.

Lesly Miculicich and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (apt). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

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)*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. *International Conference on Learning Representations (ICLR)*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. *International Conference on Learning Representations (ICLR)*.

Annette Rios, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 11–19.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2020. Using context in neural machine translation training objectives. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level nmt in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Karin Sim Smith, Wilker Aziz, and Lucia Specia. 2016. The trouble with machine translation coherence. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 178–189.

Dan Stefanescu, Rajendra Banjade, and Vasile Rus. 2014. Latent semantic analysis models on wikipedia and tasa. In *Language resources evaluation conference (LREC)*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS)*, pages 3104–3112.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1057–1063.

Amirhossein Tebbifakhr, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Machine translation for machines: the sentiment classification use case. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation (DiscoMT 2017)*.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Muriel Vasconcellos. 1989. Cohesion and coherence in the presentation of machine translation products. *Georgetown University Round Table on Languages and Linguistics*, pages 89–105.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Billy Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1060–1068.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Appendix A: Training and hyper-parameters

For our experiments, we have used, and developed on top of, the code provided by Miculicich et al. (2018) based on the OpenNMT framework (Klein et al., 2017). Our code is publicly available[9].

Sentence-level NMT: For the sentence-level model, we have used the hyper-parameters proposed by Miculicich et al. (2018) in their code repository[10]. The model uses a 6-layer transformer network (Vaswani et al., 2017) as the encoder and decoder. The dimensions of the source word embeddings, the target word embeddings and the transformers' hidden vectors have all been set to 512. The default position encoding has been added to the input vectors, and a dropout of 0.1 to the hidden vectors. Additionally, a label smoothing of 0.1 has been applied to the output probabilities. During training, the batch size has been set to 4096 tokens with a gradient accumulation of 4. We have used the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 2, and $\beta_2 = 0.998$. The parameters of the network have been initialized with the *glorot* method (Glorot and Bengio, 2010), and the model has been warmed up for $8,000$ steps. Training has been performed for 20 epochs and the model with the best perplexity over the validation set has been selected.

HAN$_{\text{join}}$: The document-level baseline follows almost exactly the settings of the sentence-level one. The main difference is the added HAN networks in the encoder and decoder. During training, for memory reasons the batch size has been reduced to $1,024$ tokens, and the learning rate to 0.2. The parameters in common have been initialized with the pre-trained sentence-level baseline, while the extra HAN networks have been initialized with *glorot*. For computational reasons, this model has been trained for only 10 epochs. In the validation, the best model was selected as that with the highest values in the majority of the evaluation metrics (BLEU, LC, COH and $\mathrm{F_{BERT}}$).

Risk models: All our proposed models use the HAN$_{\text{join}}$ architecture. As such, they all have been initialized with the pre-trained weights of the HAN$_{\text{join}}$ baseline, and then fine tuned with the Risk objective. The candidate documents for the Risk objective have been obtained with beam search and limited to just 2, due to limitations in computational resources. For the same reason, the batch size has had to be limited to 15. This has led to splitting most of the documents into multiple batches, and the reward metrics have been computed at batch level. For the batches that contained a document boundary, we have computed the rewards separately for each document. The model has been fine-tuned until convergence of the perplexity on the validation set, and using simulated annealing (Denkowski and Neubig, 2017), which repeatedly halves the learning rate when perplexity convergence is reached. The number of annealing steps has been set to 5. After training, the model with the highest values in the majority of the evaluation metrics (BLEU, LC, COH and $\mathrm{F_{BERT}}$) has been selected.

Appendix B: Translation examples

In this appendix we show other translation examples that give evidence to the translation improvement achieved by our models. Table 5 shows a Cs-En translation example. The example shows that our model has successfully translated word *renewables* while the baseline predicted *electricity*. Additionally, it has properly constructed the phrase *to build a completely new system*. Table 6 shows an Eu-En example, where our model has properly predicted word *card* instead of the baseline's *ticket*. Our model has also predicted sentence *they 'll lock me in a mental hospital* which, by looking at the source excerpt, seems more adequate than the translation provided by the baseline, and even possibly the reference sentence itself. Table 7 shows an Es-En example in the news domain. Our model has predicted phrase *any late consequence can be avoided*, which, again, seems more appropriate than the baseline's prediction *any belated consequence is possible*. Finally, Table 8 shows how our model seems to have better captured the context of the excerpt, which revolves around money and payments, and has correctly translated the Spanish word *adelanto* for *advancement*. Conversely, the translation from the HAN$_{\text{join}}$ baseline has been *earlier*, which could be correct in a different context, but not in this one.

[9] https://github.com/ijauregiCMCRC/DL_NMT_RL

[10] https://github.com/idiap/HAN_NMT

Src:	. . . otázka zní : “ můžeme ho snížit na nulu ? ” pokud budeme spalovat uhlí , tak ne . ani při spalování zemního plynu ne . téměř každý současný způsob výroby elektřiny , s vyjímkou rozšiřujících se obnovitelných a jaderných zdrojů , produkuje CO2 . budeme muset v globálním měřítku vytvořit úplně nový systém . a potřebujeme energetické zázraky . . .
Ref:	. . . and so the question is : can you actually get that to zero ? if you burn coal , no . if you burn natural gas , no . almost every way we make electricity today , except for the emerging renewables and nuclear , puts out CO2 . and so , what we ’re going to have to do at a global scale , is create a new system . and so , we need energy miracles . . .
HAN$_{join}$:	. . . the question is , can we reduce it to zero ? if we keep burning coal , we don ’t . even burning , natural gas don ’t . almost every single way of production of electricity , except for the exception of electricity and nuclear resources , produces CO2 . we ’re going to have to have a completely new system on a global scale . and we need energy miracles . . .
Risk(1.0)-	. . . the question is , can we reduce it to zero ? if we keep burning coal , we don ’t . even burning , natural gas don ’t . almost every single way of producing electricity , except for example , with the exception of renewables and nuclear resources , produces CO2 . we ’re going to have to build a completely new system on a global scale . and we need energy miracles . . .

Table 5: Translation example. Excerpt of a document from the Cs-En TED talks test set.

Appendix C: Rewards during training

To show the behavior of the different rewards during training, Figure 2 shows the BLEU, LC and COH scores over the Cs-En validation set at different training iterations. This plot confirms the intuition that improving LC and COH comes at a cost of BLEU score. In the first 2000 training iterations, LC has improved by more than 2 pp and COH by more than 1 pp, while the BLEU score has dropped by approximately 0.3 pp. Moreover, the highest scores for LC and COH coincide with the lowest score for BLEU (iteration 4000). Overall, validation is needed to achieve a model with the best trade-off between BLEU, LC and COH (in this case, for instance, iteration 2000 or 6800).

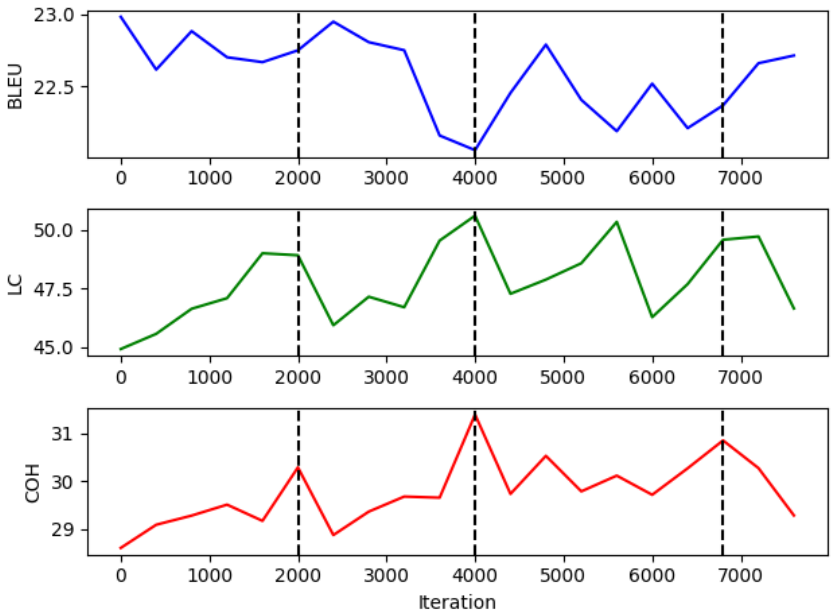

Figure 2: BLEU, LC and COH scores over the Cs-En validation set at different training iterations.

Src:	. . . ‘ zure izena ? baduzu txartelik ? ’ ‘ zure helbidea ? baduzu telefonorik etxean ? ’ ‘ eskerrik asko . joan zaitezke . harremanetan egongo gara ’ . ‘ nire familiak hau jakiten badu , eroetxe batean giltzapetuko naute ’. . .
Ref:	. . . ‘ your name ? do you have a card ? ’ ‘ your address ? do you have a telephone in your home ? ’ ‘ thank you , that ’s fine . you can leave . I will contact you later . ’ ‘ if my family learns about this , I will be forcefully detained . ’ . . .
HAN$_{join}$:	. . . your name ? do you have a ticket ? ’ your address ? do you have a phone at home ? ’ ‘ thank you . we ’ll be in touch . ’ ‘ if my family knows this , I ’ll be locked up in a mental institution . ’. . .
Risk(1.0)-	. . . ’ your name ? do you have a card ? ’ your address ? do you have a phone at home ? ’ thank you . we ’ll be in touch . ’ ‘ if my family knows this , they ’ll lock me in a mental hospital . ’ . . .

Table 6: Translation example. Excerpt of a document from the Eu-En subtitles test set.

Src:	. . . los preservativos pueden reducir el riesgo de contagio , pero no ofrecen protección al cien por cien . algunos agentes patógenos de enfermedades de transmisión sexual también pueden transmitirse a través de infecciones por suciedad y por contacto físico . por este motivo , los expertos recomiendan someterse regularmente a exámenes médicos , sobre todo si se cambia con frecuencia de pareja sexual . si se diagnostican de forma temprana , la mayoría de las ETS pueden curarse y es posible evitar cualquier consecuencia tardía . . .
Ref:	. . . condoms can reduce the risk of contraction , however , they do not offer 100 % protection . this is because occasionally , the pathogens of sexually transmitted diseases can also be passed on via smear infections and close bodily contact . therefore , first and foremost experts recommend that people with frequently changing sexual partners undergo regular examinations . if diagnosed early , the majority of STIs can be cured and long - term consequences avoided . . .
HAN$_{join}$:	. . . condoms can reduce the risk of contagion , but they do not provide protection to 100 per hundred . some <unk> agents of sexual transmission can also be cured by <unk> infections and physical contact . for this reason , experts recommend submitting regularly to medical tests , especially if sexual couple are often changed . if taken early , most of the ETS can collapse and any belated consequence is possible . . .
Risk(1.0)-	. . . condoms can reduce the risk of contagion , but they do not provide protection to a hundred per hundred . some immune agents of sexual transmission can also be channeled through infection by <unk> and physical contact . for this reason , experts report regularly to medical tests , especially if sexual couple are often changed . if they were early in , most of the ETS can collapse , and any late consequence can be avoided . . .

Table 7: Translation example. Excerpt of a document from the Es-En news test set.

Src:	. . . no voy a perdonar a ese bastardo ! Digaselo al Dr Chaddha , no me mienta . le dí el 30 % por adelantado . . . incluso después de haberme prometido , que él nos daria una esperma de calidad . digale que se joda !. . .
Ref:	. . . I wont spare that bas**** tell that Dr. Chaddha of yours , not to lie to me . he has taken 30 % advance from me . . . even after promising , he hasn ’t given us a quality sperm . you tell that f * * * er I ’ll hunt him down. . .
HAN$_{join}$:	. . . I ’m not going to forgive that bastard ! don ’t lie to me . I gave him 30 % earlier . even after I was promised , he ’d give us a quality sperm. . .
Risk(0.8):	. . . I ’m not going to forgive that bastard ! don ’t lie to me . I gave him 30 % advance . even after I was promised , he ’d give us a quality sperm . . .

Table 8: Translation example. Excerpt of a document from the Es-En subtitles test set.