Decoding as Continuous Optimization in Neural Machine Translation

Cong Duy Vu Hoang
University of Melbourne
Melbourne, VIC, Australia
vhoang2@student.unimelb.edu.au

Gholamreza Haffari
Monash University
Clayton, VIC, Australia
gholamreza.haffari@monash.edu

Trevor Cohn
University of Melbourne
Melbourne, VIC, Australia
t.cohn@unimelb.edu.au

Abstract

We propose a novel decoding approach for neural machine translation (NMT) based on continuous optimisation. The resulting optimisation problem is then tackled using constrained gradient optimisation. Our powerful decoding framework, enables decoding intractable models such as the intersection of left-to-right and right-to-left (bidirectional) as well as source-to-target and target-to-source (bilingual) NMT models. Our empirical results show that our decoding framework is effective, and leads to substantial improvements in translations generated from the intersected models where the typical greedy or beam search is infeasible.

1 Introduction

Sequence to sequence learning with neural networks (Graves, 2013; Sutskever et al., 2014; Lipton et al., 2015) is typically associated with two phases: training and decoding (*a.k.a.* inference). Model parameters are learned by optimising the training objective, so that the model generalises well when the unknown test data is decoded. The majority of literature have been focusing on developing better training paradigms or network architectures, but the decoding problem is arguably under-investigated. Conventional heuristic-based approaches for approximate inference include greedy, beam, and stochastic search. Greedy and beam search have been empirically proved to be adequate for many sequence to sequence tasks, and are the standard methods for decoding in NMT.

However, these approximate inference approaches have several drawbacks. Firstly, due to sequential decoding of symbols of the target sequence, the inter-dependencies among the target symbols are not fully exploited. For example, when decoding the words of the target sentence in a left-to-right manner, the right context is not exploited leading potentially to inferior performance (see Watanabe and Sumita (2002a) who apply this idea in traditional statistical MT). Secondly, it is not trivial to apply greedy or beam search to decode in NMT models involving global features or constraints, e.g., intersecting left-to-right and right-to-left models which do not follow the same generation order. These global constraints capture different aspects and can be highly useful in producing better and more diverse translations.

We introduce a novel decoding framework (§ 3) that effectively relaxes this *discrete* optimisation problem into a *continuous* optimisation problem. This is akin to linear programming relaxation approach for approximate inference in graphical models with discrete random variables where the exact inference is NP-hard (Sontag, 2010). Our continuous optimisation problems are challenging due to the non-linearirty and non-convexity of the relaxed decoding objective. We make use of stochastic gradient descent (SGD) and exponentiated gradient (EG) algorithms, which are mainly used for training in the literature, for decoding based on our relaxation approach. Our decoding framework is powerful and flexible, as it enables us to decode with global constraints involving intersection of multiple NMT models (§4). We present experimental results on Chinese-English and German-English translation tasks, confirming the effectiveness of our relaxed optimisation method for decoding (§5).[1]

[1]The source code will be released upon publication.

2 Neural Machine Translation

We briefly review the attentional neural translation model proposed by Bahdanau et al. (2015) as a sequence-to-sequence neural model onto which we will apply our decoding framework.

In neural machine translation (NMT), the probability of the target sentence $\boldsymbol{y}$ given a source sentence $\boldsymbol{x}$ is written as:

$$\begin{aligned} \mathrm{P}_{\boldsymbol{\Theta}}\left(\boldsymbol{y}|\boldsymbol{x}\right) &= \sum_{i=1}^{|\boldsymbol{y}|} \log \mathrm{P}_{\boldsymbol{\Theta}}\left(y_i|\boldsymbol{y}_{<i}, \boldsymbol{x}\right) \quad (1) \\ y_i|\boldsymbol{y}_{<i}, \boldsymbol{x} &\sim \mathrm{softmax}\left(\boldsymbol{f}(\boldsymbol{\Theta}, \boldsymbol{y}_{<i}, \boldsymbol{x})\right) \end{aligned}$$

where f is a non-linear function of the previously generated sequence of words $\boldsymbol{y}_{<i}$, the source sentence $\boldsymbol{x}$, and the model parameters $\boldsymbol{\Theta}$. In this paper, we realise $\boldsymbol{f}$ as follows:

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{\Theta}, \boldsymbol{y}_{<i}, \boldsymbol{x}) &= \boldsymbol{W}_o \cdot \mathrm{MLP}\left(\boldsymbol{c}_i, \boldsymbol{E}_T^{y_{i-1}}, \boldsymbol{g}_i\right) + \boldsymbol{b}_o \\ \boldsymbol{g}_i &= \mathrm{RNN}_{dec}^{\phi}\left(\boldsymbol{c}_i, \boldsymbol{E}_T^{y_{i-1}}, \boldsymbol{g}_{i-1}\right) \end{aligned}$$

where MLP is a single hidden layer neural network with $\tanh$ activation function, and $\boldsymbol{E}_T^{y_{i-1}}$ is the embedding of the target word y_{i-1} in the embedding matrix $\boldsymbol{E}_T \in \mathbb{R}^{n_e \times |V_T|}$ of the target language vocabulary V_T and n_e is the embedding dimension. The state $\boldsymbol{g}_i$ of the decoder RNN is a function of y_{i-1}, its previous state $\boldsymbol{g}_{i-1}$, and the *context* $\boldsymbol{c}_i = \sum_{j=1}^{|\boldsymbol{x}|} \alpha_{ij} \boldsymbol{h}_j$ summarises parts of the source sentence which are *attended* to, where

$$\begin{aligned} \boldsymbol{\alpha}_i &= \mathrm{softmax}(\boldsymbol{e}_i) \quad ; \quad e_{ij} = \mathrm{MLP}\left(\boldsymbol{g}_{i-1}, \boldsymbol{h}_j\right) \\ \boldsymbol{h}_j &= \mathrm{biRNN}_{enc}^{\theta}\left(\boldsymbol{E}_S^{x_j}, \overrightarrow{\boldsymbol{h}}_{j-1}, \overleftarrow{\boldsymbol{h}}_{j+1}\right) \end{aligned}$$

In above, $\overrightarrow{\boldsymbol{h}}_i$ and $\overleftarrow{\boldsymbol{h}}_i$ are the states of the left-to-right and right-to-left RNNs encoding the source sentence, and $\boldsymbol{E}_S^{x_j}$ is the embedding of the source word x_j in the embedding matrix $\boldsymbol{E}_S \in \mathbb{R}^{n'_e \times |V_S|}$ of the source language vocabulary V_S and n'_e is the embedding dimension.

Given a bilingual corpus $\mathcal{D}$, the model parameters are learned by maximizing the (regularised) conditional log-likelihood:

$$\boldsymbol{\Theta}^* := \mathrm{argmax}_{\boldsymbol{\Theta}} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}} \log \mathrm{P}_{\boldsymbol{\Theta}}\left(\boldsymbol{y} \mid \boldsymbol{x}\right). \quad (2)$$

The model parameters $\boldsymbol{\Theta}$ include the weight matrix $\boldsymbol{W}_o \in \mathbb{R}^{|V_T| \times n_h}$ and the bias $\boldsymbol{b}_o \in \mathbb{R}^{|V_T|}$ – with n_H denoting the hidden dimension size – as well as the RNN encoder $\mathrm{biRNN}_{enc}^{\theta}$ / decoder $\mathrm{RNN}_{dec}^{\phi}$ parameters, word embedding matrices, and those of the attention mechanism. The model is trained end-to-end by optimising the training objective using stochastic gradient descent (SGD) or its variants. In this paper, we are interested in the decoding problem though which is outlined in the next section.

3 Decoding as Continuous Optimisation

In decoding, we are interested in finding the highest probability translation for a given source sentence:

$$\text{minimise}_{\boldsymbol{y}} \quad -P_{\boldsymbol{\Theta}}\left(\boldsymbol{y} \mid \boldsymbol{x}\right) \quad \text{s.t.} \quad \boldsymbol{y} \in \mathcal{Y}_{\boldsymbol{x}} \quad (3)$$

where $\mathcal{Y}_{\boldsymbol{x}}$ is the space of possible translations for the source sentence $\boldsymbol{x}$. In general, searching $\mathcal{Y}_{\boldsymbol{x}}$ to find the highest probability translation is intractable due to long-range dependency terms in eqn (1) which prevents dynamic programming for efficient search algorithms in this exponentially-large space of possible translations with respect to the input length $|\boldsymbol{x}|$.

We now formulate this discrete optimisation problem as a continuous one, and then use standard algorithms for continuous optimisation for decoding. Let us assume that the maximum length of a possible translation for a source sentence is known and denote it by ℓ. The best translation for a given source sentence solves the following optimisation problem:

$$\begin{aligned} \boldsymbol{y}^* = \underset{y_1,\ldots,y_\ell}{\arg\min} \sum_{i=1}^{\ell} -\log \mathrm{P}_{\boldsymbol{\Theta}}\left(y_i \mid \boldsymbol{y}_{<i}, \boldsymbol{x}\right) \quad (4) \\ \text{s.t.} \quad \forall i \in \{1 \ldots \ell\} : y_i \in V_T. \end{aligned}$$

Equivalently, we can re-write the above discrete optimisation problem as follows:

$$\begin{aligned} &\underset{\tilde{\boldsymbol{y}}_1,\ldots,\tilde{\boldsymbol{y}}_\ell}{\arg\min} -\sum_{i=1}^{\ell} \tilde{\boldsymbol{y}}_i \cdot \log \mathrm{softmax}\left(\boldsymbol{f}\left(\boldsymbol{\Theta}, \tilde{\boldsymbol{y}}_{<i}, \boldsymbol{x}\right)\right) \\ &\text{s.t.} \quad \forall i \in \{1 \ldots \ell\} : \tilde{\boldsymbol{y}}_i \in \mathbb{I}^{|V_T|} \quad (5) \end{aligned}$$

where $\tilde{\boldsymbol{y}}_i$ are vectors using the one-hot representation of the target words $\mathbb{I}^{|V_T|}$.

We now convert the optimisation problem (5) to a continuous one by dropping the integrality constraints $\tilde{\boldsymbol{y}}_i \in \mathbb{I}^{|V|}$ and require the variables to take values from the probability simplex:

$$\begin{aligned} &\underset{\hat{\boldsymbol{y}}_1,\ldots,\hat{\boldsymbol{y}}_\ell}{\arg\min} -\sum_{i=1}^{\ell} \hat{\boldsymbol{y}}_i \cdot \log \mathrm{softmax}\left(\boldsymbol{f}\left(\boldsymbol{\Theta}, \hat{\boldsymbol{y}}_{<i}, \boldsymbol{x}\right)\right) \\ &\text{s.t.} \quad \forall i \in \{1 \ldots \ell\} : \hat{\boldsymbol{y}}_i \in \Delta_{|V_T|} \end{aligned}$$

Algorithm 1 The EG Algorithm for Decoding by Optimisation

1: For all i initialise $\hat{\boldsymbol{y}}_i^0 \in \Delta_{|V_T|}$
2: **for** $t = 1, \ldots,$ MaxIter **do** ▷ $Q(.)$ is defined as eqn (6)
3: For all i, w : calculate $\nabla_{i,w}^{t-1} = \frac{\partial Q(\hat{\boldsymbol{y}}_1^{t-1}, \ldots, \hat{\boldsymbol{y}}_\ell^{t-1})}{\partial \hat{\boldsymbol{y}}_i(w)}$ ▷ using backpropagation
4: For all i, w : update $\hat{\boldsymbol{y}}_i^t(w) \propto \hat{\boldsymbol{y}}_i^{t-1}(w) \cdot \exp\left(-\eta \nabla_{i,w}^{t-1}\right)$ ▷ η is the step size
5: **return** $\arg\min_t Q(\hat{\boldsymbol{y}}_1^t, \ldots, \hat{\boldsymbol{y}}_\ell^t)$

where $\Delta_{|V_T|}$ is the $|V_T|$-dimensional probability simplex, i.e., $\{\hat{\boldsymbol{y}}_i \in [0,1]^{|V_T|} : \|\hat{\boldsymbol{y}}\|_1 = 1\}$. Intuitively, this amounts to replacing $\boldsymbol{E}_T^{y_i}$ with the *expected* embedding of target language words $\mathbb{E}_{\hat{\boldsymbol{y}}_i(w)}[\boldsymbol{E}_T^w]$ under the distribution $\hat{\boldsymbol{y}}_i$ in the NMT model.

After solving the above constrained continuous optimisation problem, there is no guarantee that the resulting solution $\{\hat{\boldsymbol{y}}_i^*\}_{i=1}^\ell$ to include one-hot vectors corresponding to target language words. It instead will have *distributions* over target language vocabulary for each random variable of interest in prediction, so we need a technique to round up this *fractional* solution. Our method is to put all of the probability mass on the word with the highest probability[2] for each $\hat{\boldsymbol{y}}_i^*$. We leave exploration of more elaborate projection techniques to the future work.

In the context of graphical models, the above relaxation technique gives rise to linear programming for approximate inference (Sontag, 2010). However, our decoding problem is much harder due to the non-linearity and non-convexity of the objective function for deep models. We now turn our attention to optimisation algorithms to effectively solve the decoding optimisation problem.

3.1 Exponentiated Gradient (EG)

Exponentiated gradient (Kivinen and Warmuth, 1997) is an elegant algorithm for solving optimisation problems involving simplex constraints. Recall our constrained optimisation problem:

$$\arg\min_{\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_\ell} Q(\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_\ell)$$
$$\text{s.t.} \quad \forall i \in \{1 \ldots \ell\} : \hat{\boldsymbol{y}}_i \in \Delta_{|V_T|}$$

where $Q(\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_\ell)$ is defined as

$$-\sum_{i=1}^{\ell} \hat{\boldsymbol{y}}_i \cdot \log \operatorname{softmax}\left(\boldsymbol{f}\left(\boldsymbol{\Theta}, \hat{\boldsymbol{y}}_{<i}, \boldsymbol{x}\right)\right). \quad (6)$$

[2] If there are multiple words with the same highest probability mass, we choose one of them arbitrarily.

EG is an iterative algorithm, which updates each distribution $\hat{\boldsymbol{y}}_i^t$ in the current time-step t based on the distributions of the previous tilmestep as follows:

$$\forall w \in V_T : \quad \hat{\boldsymbol{y}}_i^t(w) = \frac{1}{Z_i^t} \hat{\boldsymbol{y}}_i^{t-1}(w) \exp\left(-\eta \nabla_{i,w}^{t-1}\right)$$

where η is the step size, $\nabla_{i,w}^{t-1} = \frac{\partial Q(\hat{\boldsymbol{y}}_1^{t-1}, \ldots, \hat{\boldsymbol{y}}_\ell^{t-1})}{\partial \hat{\boldsymbol{y}}_i(w)}$ and Z_i^t is the normalisation constant

$$Z_i^t = \sum_{w \in D_T} \hat{\boldsymbol{y}}_i^{t-1}(w) \exp\left(-\eta \nabla_{i,w}^{t-1}\right).$$

The partial derivatives $\nabla_{i,w}$ are calculated using the back propagation algorithm treating $\hat{\boldsymbol{y}}_i$'s as *parameters* and the original parameters of the model $\boldsymbol{\Theta}$ as constants. Adapting EG to our decoding problem leads to Algorithm 1. It can be shown that the EG algorithm is a gradient descent algorithm for minimising the following objective function subject to the simplex constraints:

$$Q(\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_\ell) - \gamma \sum_{i=1}^{\ell} \sum_{w \in V_T} \hat{\boldsymbol{y}}_i(w) \log \frac{1}{\hat{\boldsymbol{y}}_i(w)}$$
$$= Q(\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_\ell) - \gamma \sum_{i=1}^{\ell} \text{Entropy}(\hat{\boldsymbol{y}}_i) \quad (7)$$

In other words, the algorithm looks for the maximum entropy solution which also maximizes the log likelihood under the model.

3.2 Stochastic Gradient Descent (SGD)

To be able to apply SGD to our optimisation problem, we need to make sure that the simplex constraints are kept intact. One way to achieve this is by changing the optimisation variables from $\hat{\boldsymbol{y}}_i$ to $\hat{\boldsymbol{r}}_i$ through the softmax transformation, i.e. $\hat{\boldsymbol{y}}_i = \operatorname{softmax}(\hat{\boldsymbol{r}}_i)$. The resulting *unconstrained* optimisation problem then becomes

$$\arg\min_{\hat{\boldsymbol{r}}_1, \ldots, \hat{\boldsymbol{r}}_\ell} -\sum_{i=1}^{\ell} \operatorname{softmax}(\hat{\boldsymbol{r}}_i) \cdot \log \operatorname{softmax}\left(\text{f}\left(\boldsymbol{\Theta}, \hat{\boldsymbol{y}}_{<i}, \boldsymbol{x}\right)\right)$$

Algorithm 2 The SGD Algorithm for Decoding by Optimisation

For all i initialise $\hat{\boldsymbol{r}}_i^0$
for $t = 1, \ldots,$ MaxIter **do** ▷ $Q(.)$ is defined in eqn (6) and $\hat{\boldsymbol{y}}_i = \text{softmax}(\hat{\boldsymbol{r}}_i)$
  For all i, w : calculate $\nabla_{i,w}^{t-1} = \sum_{w' \in V_T} \frac{\partial Q(\hat{\boldsymbol{y}}_1^{t-1}, \ldots, \hat{\boldsymbol{y}}_\ell^{t-1})}{\partial \hat{\boldsymbol{y}}_i(w')} \frac{\partial \hat{\boldsymbol{y}}_i(w')}{\partial \hat{\boldsymbol{r}}_i(w)}$ ▷ using backpropagation
  For all i, w : update $\hat{\boldsymbol{r}}_i^t(w) = \hat{\boldsymbol{r}}_i^{t-1}(w) - \eta \nabla_{i,w}^{t-1}$ ▷ η is the step size
return $\arg\min_t Q(\text{softmax}(\hat{\boldsymbol{r}}_1^t), \ldots, \text{softmax}(\hat{\boldsymbol{r}}_\ell^t))$

where $\boldsymbol{E}_T^{y_i}$ is replaced with the expected embedding of the target words under the distribution resulted from the $\mathbb{E}_{\text{softmax}(\hat{\boldsymbol{r}}_i)}[\boldsymbol{E}_T^w]$ in the model.

To apply SGD updates, we need the gradient of the objective function with respect to the new variables $\hat{\boldsymbol{r}}_i$ which can be derived with the back-propagation algorithm based on the chain rule:

$$\frac{\partial Q}{\partial \hat{\boldsymbol{r}}_i(w)} = \sum_{w' \in V_T} \frac{\partial Q(.)}{\partial \hat{\boldsymbol{y}}_i(w')} \frac{\partial \hat{\boldsymbol{y}}_i(w')}{\partial \hat{\boldsymbol{r}}_i(w)}$$

The resulting SGD algorithm is summarized in Algorithm 2.

4 Decoding in Extended NMT

Our decoding framework allows us to effectively and flexibly add additional global factors over the output symbols during inference. This in enabling by allowing decoding for richer global models, for which there is no effective means of greedy decoding or beam search. We outline several such models, and their corresponding relaxed objective functions for optimisation-based decoding.

Bidirectional Ensemble Standard NMT generates the translation in a left-to-right manner, conditioning each target word on its left context. However, the joint probability of the translation can be decomposed in a myriad of different orders; one compelling alternative would be to condition each target word on its right context, i.e., generating the target sentence from right-to-left. We would not expect a right-to-left model to outperform a left-to-right, however, as the left-to-right ordering reflects the natural temporal order of spoken language. However, the right-to-left model is likely to provide a complementary signal in translation as it will be bringing different biases and making largely independent prediction errors to those of the left-to-right model. For this reason, we propose to use both models, and seek to find translations that have high probability according both models (this mirrors work on bidirectional decoding in classical statistical machine translation by Watanabe and Sumita (2002b).) Decoding under the ensemble of these models leads to an intractable search problem, not well suited to traditional greedy or beam search algorithms, which require a fixed generation order of the target words. This ensemble decoding problem can be formulated simply in our linear relaxation approach, using the following objective function:

$$\begin{aligned}\mathcal{C}_{+\text{bidir}} := &- \alpha \log \mathrm{P}_{\boldsymbol{\Theta}_{\leftarrow}}(\boldsymbol{y} \mid \boldsymbol{x}) \\ &- (1-\alpha) \log \mathrm{P}_{\boldsymbol{\Theta}_{\rightarrow}}(\boldsymbol{y} \mid \boldsymbol{x})\,; \end{aligned} \quad (8)$$

where α is an interpolation hyper-parameter, which we set to 0.5; $\boldsymbol{\Theta}_{\rightarrow}$ and $\boldsymbol{\Theta}_{\leftarrow}$ are the pre-trained left-to-right and right-to-left models, respectively. This bidirectional agreement may also lead to improvement in translation diversity, as shown in (Li and Jurafsky, 2016) in a re-ranking evaluation.

Bilingual Ensemble Another source of complementary information is in terms of the translation direction, that is forward translation from the source to the target language, and reverse translation in the target to source direction. The desire now is to find a translation which is good under both the forward and reverse translation models. This is inspired by the direct and reverse feature functions commonly used in classical discriminative SMT (Och and Ney, 2002) which have been shown to offer some complementary benefits (although see (Lopez and Resnik, 2006)). More specifically, we decode for the best translation in the intersection of the source-to-target and target-to-source models by minimizing the following objective function:

$$\begin{aligned}\mathcal{C}_{+\text{biling}} := &- \alpha \log \mathrm{P}_{\boldsymbol{\Theta}_{\text{s}\rightarrow\text{t}}}(\boldsymbol{y} \mid \boldsymbol{x}) \\ &- (1-\alpha) \log \mathrm{P}_{\boldsymbol{\Theta}_{\text{s}\leftarrow\text{t}}}(\boldsymbol{x} \mid \boldsymbol{y}) \\ &- \gamma \,\mathrm{tr}\left(\boldsymbol{\alpha}_{\text{s}\rightarrow\text{t}} \boldsymbol{\alpha}_{\text{s}\leftarrow\text{t}}^T\right); \end{aligned} \quad (9)$$

where α is an interpolation hyper-parameter to be fine-tuned; and $\boldsymbol{\Theta}_{\text{s}\rightarrow\text{t}}$ and $\boldsymbol{\Theta}_{\text{s}\leftarrow\text{t}}$ are the pre-trained

	# tokens	# types	# sents
BTEC zh↔en			
train	422k / 454k	3k / 3k	44,016
dev	10k / 10k	1k / 1k	1,006
test	5k / 5k	1k / 1k	506
TED Talks de↔en			
train	4067k / 4329k	26k / 19k	194,181
dev	33k / 35k	4k / 3k	1,565
test2013	22k / 23k	3k / 3k	993
test2014	26k / 27k	4k / 3k	1,305

Table 1: Statistics of the training and evaluation sets; token and types are presented for both source/target languages.

source-to-target and target-to-source models, respectively. Inspired by Cohn et al. (2016), the last term in (9) is the so-called trace bonus which encourages agreement between the two models' alignments. Decoding for the best translation under the above objective function leads to an intractable search problem, as the reverse model is global over the target language, meaning there is no obvious means of search with greedy algorithm or alike.

Discussion Our framework is highly general, and several other model architectures can be easily supported, such as the noisy-channel model (Koehn, 2010), which combines source to target translation with a target language model; or decoding with additional constraints, e.g., source coverage (Xu et al., 2015; Mi et al., 2016) or word fertility (Cohn et al., 2016).[3] We leave these and other extensions for future work.

There are two important considerations on how best to initialise the relaxed optimisation in the above settings, and how best to choose the step size. As the relaxed optimisation problem is, in general, non-convex, finding a plausible initialisation is likely to be important for avoiding local optima. Furthermore, a proper step size is a key in the success of the EG-based and SGD-based optimisation algorithms, and there is no obvious method how to best choose its value. We may also adaptively change the step size using (scheduled) annealing or via the line search. We return to this considerations in the experimental evaluation.

[3]These constraints are only enforced only during training but not decoding, presumably due to intractable search. Our framework could allow their use in decoding.

5 Experiments

5.1 Setup

Datasets. We conducted our experiments on two datasets, translating between Chinese↔English using the BTEC corpus, and German↔English using the IWSLT 2015 TED Talks corpus (Cettolo et al., 2014). The statistics of the datasets can be found in Table 1.

NMT Models. We implemented our continuous-optimisation based decoding method on top of the mantis toolkit[4] (Cohn et al., 2016), and using the *dynet* deep learning library[5] (Neubig et al., 2017). All neural network models were configured with 512 input embedding and hidden layer dimensions, and 256 alignment dimension, with 1 and 2 hidden layers in the source and target, respectively. We used a LSTM recurrent structure (Hochreiter and Schmidhuber, 1997) for both source and target RNN sequences. For vocabulary sizes, we have chosen the word frequency cut-off 5 for creating the vocabularies for all datasets. For training our neural models, the best perplexity scores on the development set is used for early stopping, which usually occurs after 5-8 epochs.

Evaluation Metrics. We evaluated in terms of search error, measured using the model score of the inferred solution (either continuous or discrete), as well as measuring the end translation quality with BLEU (Papineni et al., 2002). The continuous cost measures $-\frac{1}{|\hat{\boldsymbol{y}}|}\log P_{\boldsymbol{\Theta}}(\hat{\boldsymbol{y}} \mid \boldsymbol{x})$ under the model $\boldsymbol{\Theta}$; the discrete model score has the same formulation, albeit using the discrete rounded solution $\boldsymbol{y}$ (see §3). Note the cost can be used as a tool for selecting the best inference solution, as well as assessing convergence, as we illustrate below.

5.2 Results and Analysis

Initialisation and Step Size. As our relaxed optimisation problems are not convex, local optima are likely to be a problem. We test this empirically, focusing on the effect that initialisation and step size, η, have on the inference quality.

For plausible initialisation states, we evaluate different strategies: *uniform* in which the relaxed variables $\hat{\boldsymbol{y}}$ are initialised to $\frac{1}{|V_T|}$; and *greedy* or *beam* whereby $\hat{\boldsymbol{y}}$ are initialised based on an al-

[4]https://github.com/trevorcohn/mantis
[5]https://github.com/clab/dynet

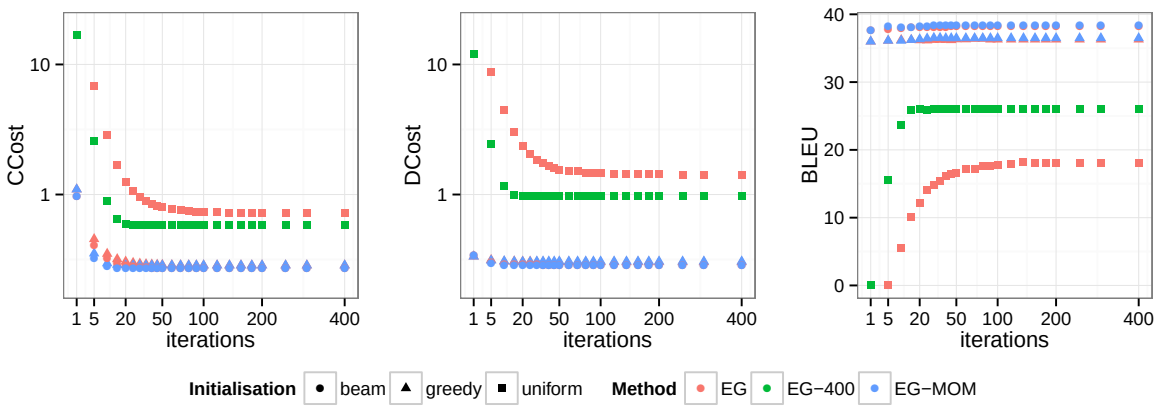

Figure 1: Analysis on effects of initialisation states (uniform vs. greedy vs. beam), step size annealing, momentum mechanism from BTEC zh→en translation. **EG-400**: EG algorithm with step size $\eta = 400$ (otherwise $\eta = 40$); **EG-MOM**: EG algorithm with momentum.

ready good solution produced by a baseline decoder. Instead of using the Viterbi outputs as a one-hot representation, we initialise to the probability prediction[6] vectors, which serves to limit attraction of the initialisation condition, which is likely to be a local (but not global) optima.

Figure 1 illustrates the effect of initialisation on the EG algorithm, in terms of search error (left and middle) and translation quality (right), as we vary the number of iterations of inference. There is clear evidence of non-convexity: all initialisation methods can be seen to converge using all three measures, however they arrive at highly different solutions. Uniform initialisation is clearly not a viable approach, while greedy and beam initialisation both yield much better results. The best initialisation, beam, outperforms both greedy and beam decoding in terms of BLEU.

Note that the EG algorithm has fairly slow convergence, requiring at least 100 iterations, irrespective of the initialisation. To overcome this, we use momentum (Qian, 1999) to accelerate the convergence by modifying the term $\nabla^t_{i,w}$ in Algorithm 1 with a weighted moving average of past gradients:

$$\nabla^{t-1}_{i,w} = \gamma \nabla^{t-2}_{i,w} + \eta \frac{\partial Q(\hat{\boldsymbol{y}}^{t-1}_1, \ldots, \hat{\boldsymbol{y}}^{t-1}_\ell)}{\partial \hat{\boldsymbol{y}}_i(w)}$$

where we set the momentum term $\gamma = 0.9$. The EG with momentum (**EG-MOM**) converges after fewer iterations (about 35), and results in marginally better BLEU scores. The momentum technique is usually used for SGD involving additive updates; it is interesting to see it also works in EG with multiplicative updates.

The step size, η, is another important hyperparameter for gradient based search. We tune the step size using line search over $[10, 400]$ over the development set. Figure 1 illustrates the effect of changing step size from 50 to 400 (compare **EG** and **EG-400** with **uniform**), which results in a marked difference of about 10 BLEU points, underlining the importance of tuning this value. We found that EG with momentum had less of a reliance on step size, with optimal values in $[10, 50]$; we use this setting hereafter.

Continuous vs Discrete Costs. Another important question is whether the assumption behind continuous relaxation is valid, i.e., if we optimise a continuous cost to solve a discrete problem, do we improve the discrete output? Although the continuous cost diminishes with inference iterations (Figure 1 centre), and appears to converge to an optima, it is not clear whether this corresponds to a better discrete output (note that the BLEU scores do show improvements Figure 1.) Figure 2 illustrates the relation between the two cost measures, showing that in almost all cases the discrete and continuous costs are identical. Linear relaxation effectively fails only for a handful of cases, where the nearest discrete solution is significantly worse than it would appear using the continuous cost.

EG vs SGD. Both the EG and SGD algorithms are iterative methods for solving the relaxed optimisation problem with simplex constraints. We measure empirically their different in terms of

[6]Here, the EG algorithm uses softmax normalization whereas the SGD algorithm uses pre-softmax one.

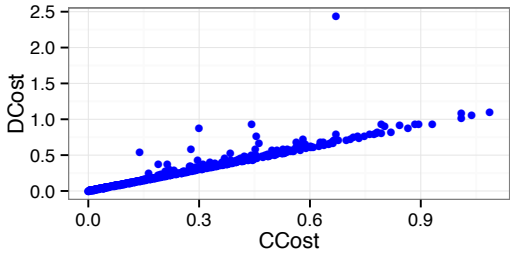

Figure 2: Comparing discrete vs continuous costs from BTEC zh→en translation, using the EG algorithm with momentum, $\eta = 50$. Each point corresponds to a sentence.

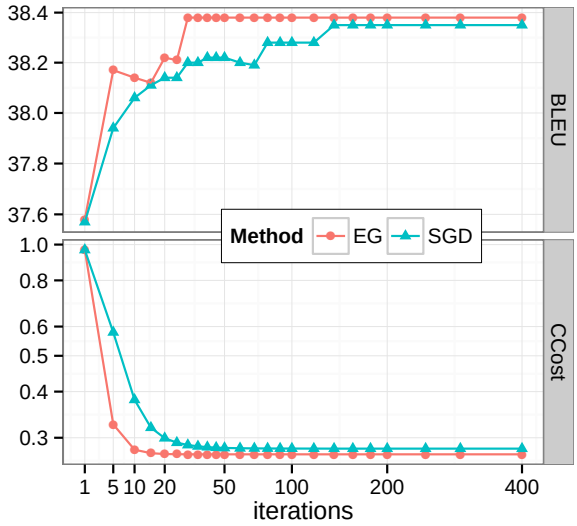

Figure 3: Analysis on convergence and performance comparing SOFTMAX and EG algorithms from BTEC zh→en translation. Both algorithms use momentum and step size 50.

quality of inference and speed of convergence, as illustrated in Figure 3. Observe that SGD requires 150 iterations for convergence, whereas EG requires many fewer (50). This concurs with previous work on learning structured prediction models with EG (Globerson et al., 2007). Further, the EG algorithm consistently produces better results in terms of both model cost and BLEU.

Main Results. Table 2 shows our experimental results across all datasets,[7] evaluating the EG algorithm and its variants.[8] For the EG algorithm with greedy initialisation (top), we see small but consistent improvements in terms of BLEU. Beam initialisation led to overall higher BLEU scores, and again demonstrating a similar pattern of improvements, albeit of a lower magnitude, over the initialisation values.

Next we evaluate the capability of our inference method with extended NMT models, where approximate algorithms such as greedy or beam search are infeasible. With the **bidirectional** ensemble, we obtained the statistically significant BLEU score improvements compared to the unidirectional models, for either greedy or beam initialisation. This is interesting in the sense that the unidirectional right-to-left model always performs worse than the left-to-right model (sometimes badly, e.g., en→de'13). However, our method with bidirectional ensemble is capable of combining their strengths in a unified setting. For the **bilingual** ensemble, we see similar effects, with significant BLEU score improvements in most cases. Extending inference to use a trace bonus is not always beneficial, leading to a small degradation in performance in several cases. This is likely to be due to a disparity with the training condition for the models, which were learned independently of one another, without the trace bonus term.

Overall, decoding in extended NMT models leads to performance improvements compared to the baselines. This is one of the main findings in this work, and augurs well for its extension to other global model variants.

6 Related Work

Decoding (inference) for neural models is an important task; however, there is limited research in this space perhaps due to the challenging nature of this task. The most widely-used inference methods include sampling (Cho, 2016), greedy and beam search (Sutskever et al., 2014; Bahdanau et al., 2015, *inter alia*), and reranking (Birch, 2016; Li and Jurafsky, 2016).

Cho (2016) proposed to perturb the neural model by injecting noise(s) in the hidden transition function of the conditional recurrent neural language model during greedy or beam search, and execute multiple parallel decoding runs. This strategy can improves over greedy and beam search; however, however it is not clear how, when and where noise should be injected to be beneficial. Recently, Wiseman and Rush (2016) proposed beam search optimisation while *training* neural models, where the model parameters are updated in case the gold standard falls outside of

[7] Comparative translation examples are included in the *Supplementary Material*.

[8] Given the aforementioned analysis and space constraints, here we reported the results for the EG algorithm only.

	zh→en	en→zh	de → en'13	en → de'13	de → en'14	en → de'14
greedy $\text{dec}_{\text{left-to-right}}$	35.98	20.23	27.58	23.16	23.16	19.33
greedy $\text{dec}_{\text{right-to-left}}$	35.86	20.01	26.83	20.78	21.95	17.18
relopt $\text{dec}_{\text{ginit}}$	36.34	20.42	27.81	23.29	23.28	19.37
+bidirectional	**36.67**	**21.50**†	**28.65**†	**23.76**	**23.91**	**19.92**
+bilingual	**36.88**†	**21.16**	**28.41**	**24.07**†	**24.01**†	**20.22**†
+bilingual+trace	**36.59**	**21.50**†	**28.23**	**23.82**	**23.65**	**19.85**
beam $\text{dec}_{\text{left-to-right}}$	38.02	21.20	28.74	24.62	23.95	20.64
beam $\text{dec}_{\text{right-to-left}}$	37.38	17.60	28.30	22.17	23.13	19.45
relopt $\text{dec}_{\text{binit}}$	38.38	21.11	28.75	24.58	24.02	20.54
+bidirectional	**39.13**†	**21.87**	**29.92**†	**25.18**	**24.72**†	21.06†
+bilingual	38.25	**21.86**	**29.20**	**25.19**†	**24.60**	21.01
+bilingual+trace	**38.64**	**22.10**†	28.87	**25.09**	**24.52**	20.57

Table 2: The BLEU evaluation results across datasets for all **relopt** variants with EG algorithm against the baselines; **bold**: statistically significantly better than the best greedy or beam baseline, †: best performance on dataset.

the beam. This exposes the model to its past incorrect predicted labels, hence making the training more robust.[9] This is orthogonal to our approach where we focus on the decoding problem with a pre-trained model.

Reranking has also been proposed as a means of global model combination: Birch (2016) and Li and Jurafsky (2016) re-rank the left-to-right decoded translations based on the scores of a right-to-left model, learning to more diverse translations. Related, Li et al. (2016) learn to adjust the beam diversity with reinforcement learning.

Perhaps most relevant is Snelleman (2016), performed concurrently to this work, who also proposed an inference method for NMT using linear relaxation. Snelleman's method was similar to our SGD approach, however he did not manage to outperform beam search baselines with an encoder-decoder. In contrast we go much further, proposing the EG algorithm, which we show works much more effectively than SGD, and demonstrate how this can be applied to inference in an attentional encoder-decoder. Moreover, we demonstrate the utility of related optimisation for inference over global ensembles of models, resulting in consistent improvements in search error and end translation quality.

[9] See also imitation learning, which supports training-time conditioning on model errors to allow more robust test inferences with greedy search (Chang et al., 2015; Ballesteros et al., 2016, *inter alia*).

7 Conclusions

We have presented a novel framework for decoding in neural translation models, where decoding is formulated as a continuous optimisation problem. The core idea is to drop the integrality (i.e. one-hot vector) constraints from the prediction variables, and treat them by soft assignments belonging to the probability simplex with the goal of minimising the loss produced by the neural model. We have provided two optimisation algorithms, i.e. exponentiated gradient (EG) and stochastic gradient descent (SGD), for optimising the resulting contained optimisation problem, where our findings show the effectiveness of EG compared to SGD. Thanks to our framework, we have been able to decode and intersect left-to-right and right-to-left NMT models as well as source-to-target and target-to-source NMT models. Our results show that that our decoding framework is effective, and lead to substantial improvements in translations generated from the intersected models where the typical greedy or beam search are not applicable.

This work raises several compelling possibilities, which we intend to address in future work, including the integration of additional constraints, such as word coverage, fertility etc into *decoding*, which previously have only been included during training (Cohn et al., 2016; Mi et al., 2016), as well as applying the method to other intractable structured prediction tasks beyond translation.

Acknowledgments

Cong Duy Vu Hoang was supported by Australian Government Research Training Program Scholarships at the University of Melbourne, Australia. Dr Trevor Cohn was supported by the ARC Future Fellowship.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of 3rd International Conference on Learning Representations (ICLR2015)*.

Miguel Ballesteros, Yoav Goldberg, Chris Dyer, and Noah A. Smith. 2016. Training with exploration improves a greedy stack lstm parser. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2005–2010.

Rico Sennrich; Barry Haddow; Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers. Berlin, Germany*.

M. Cettolo, J. Niehues, S. Stuker, L. Bentivogli, and M. Federico. 2014. Report on the 11th IWSLT Evaluation Campaign. In *Proc. of The International Workshop on Spoken Language Translation (IWSLT)*.

Kai-wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daume, and John Langford. 2015. Learning to search better than your teacher. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. pages 2058–2066.

K. Cho. 2016. Noisy Parallel Approximate Decoding for Conditional Recurrent Language Model. *ArXiv e-prints* .

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating Structural Alignment Biases into an Attentional Neural Translation Model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 876–885.

Amir Globerson, Terry Y. Koo, Xavier Carreras, and Michael Collins. 2007. Exponentiated Gradient Algorithms for Log-linear Structured Prediction. In *Proceedings of the 24th International Conference on Machine Learning*. ACM, New York, NY, USA, ICML '07, pages 305–312.

A. Graves. 2013. Generating Sequences With Recurrent Neural Networks. *ArXiv e-prints* .

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9(8):1735–1780.

Jyrki Kivinen and Manfred K. Warmuth. 1997. Exponentiated Gradient Versus Gradient Descent for Linear Predictors. *Inf. Comput.* 132(1):1–63.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

J. Li and D. Jurafsky. 2016. Mutual Information and Diverse Decoding Improve Neural Machine Translation. *ArXiv e-prints* .

J. Li, W. Monroe, and D. Jurafsky. 2016. A Simple, Fast Diverse Decoding Algorithm for Neural Generation. *ArXiv e-prints* .

Z. C. Lipton, J. Berkowitz, and C. Elkan. 2015. A Critical Review of Recurrent Neural Networks for Sequence Learning. *ArXiv e-prints* .

Adam Lopez and Philip Resnik. 2006. Word-based alignment, phrase-based translation: Whats the link. In *Proceedings of 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage Embedding Models for Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 955–960.

G. Neubig, C. Dyer, Y. Goldberg, A. Matthews, W. Ammar, A. Anastasopoulos, M. Ballesteros, D. Chiang, D. Clothiaux, T. Cohn, K. Duh, M. Faruqui, C. Gan, D. Garrette, Y. Ji, L. Kong, A. Kuncoro, G. Kumar, C. Malaviya, P. Michel, Y. Oda, M. Richardson, N. Saphra, S. Swayamdipta, and P. Yin. 2017. DyNet: The Dynamic Neural Network Toolkit. *ArXiv e-prints* .

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. pages 295–302.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318.

Ning Qian. 1999. On the momentum term in gradient descent learning algorithms . *Neural Networks* 12(1):145 – 151.

Emanuel Snelleman. 2016. *Decoding neural machine translation using gradient descent*. Master's thesis, Chalmers University of Technology, Gothenburg, Sweden.

David Sontag. 2010. *Approximate Inference in Graphical Models using LP Relaxations*. Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, NIPS'14, pages 3104–3112.

Taro Watanabe and Eiichiro Sumita. 2002a. Bidirectional Decoding for Statistical Machine Translation. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '02, pages 1–7.

Taro Watanabe and Eiichiro Sumita. 2002b. Bidirectional decoding for statistical machine translation. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. pages 1–7.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-Sequence Learning as Beam-Search Optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1296–1306.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. pages 2048–2057.

Supplementary material with translation examples for *Decoding as Continuous Optimization in Neural Machine Translation*

	BTEC zh→en
Source	我 确定 我 昨天 给 旅馆 打 过 电话 并且 做 了 预定 。
Reference	i am sure that i called the hotel yesterday and made a reservation .
beam dec	i 'm sure i called the hotel reservation and i made a reservation .
relopt dec	i 'm sure i called the hotel yesterday and i made a reservation .
Source	当 我 到 路口 时 我 这 边 的 灯 是 绿色 的 。
Reference	my light was green when i got to the intersection .
beam dec (l2r)	this was the green UNK i came in my room .
beam dec (r2l)	this is green when this is on the intersection .
relopt dec	this was the green UNK i came in my room .
+bidirectional	this UNK the green when i was on the intersection .
Source	请 大 点 声讲 。
Reference	please speak a little louder .
beam dec (l2r)	please speak some more UNK .
beam dec (r2l)	a little UNK , please .
relopt dec	please speak some more UNK .
+bidirectional	please speak a little louder .
Source	请来 一 个 中 瓶 的 可乐 。
Reference	a medium coke , please .
beam dec (l2r)	a medium cola , please .
beam dec (r2l)	a large coke , please .
relopt dec	a medium bottle , please
+bidirectional	a medium coke , please .
Source	我 很 快 会 变 好 吗 ?
Reference	will i get better soon ?
beam dec (l2r)	will i be ready soon ?
beam dec (r2l)	will i get well soon ?
relopt dec	will i be ready soon ?
+bidirectional	will i be better soon ?
Source	驾驶员 , 能 在 这 停 吗 ?
Reference	driver , can you stop there ?
beam dec (l2r)	UNK , can i park here ?
beam dec (r2l)	could you stop here , please ?
relopt dec	UNK , can i park here ?
+bidirectional	UNK , can i stop here ?
Source	那些 是 免税 物品 。
Reference	those are duty- free items .
beam dec (l2r)	those are tax- free items .
beam dec (r2l)	these are duty- free items .
relopt dec	those are tax- free items .
+bidirectional	those are duty- free items .
Source	有 对 孩子们 的 折扣 吗 ?
Reference	is there a discount for children ?
beam dec (l2r)	do you have a discount for that ?
beam dec (r2l)	do you have any reduction for children ?
relopt dec	do you have a discount for the ?
+bidirectional	do you have a discount for children ?

Supplementary material with translation examples for *Decoding as Continuous Optimization in Neural Machine Translation*

	TED Talks de→en
Source	durchhaltevermögen ist leidenschaft und ausdauer für sehr langfristige ziele .
Reference	grit is passion and perseverance for very long-term goals .
beam dec	persistence is passion and resilience for very long-term targets .
relopt dec	persistence is passion and resilience for very long-term targets .
+bilingual	resilience is passion and perseverance for very long-term targets .
Source	man muss aber nicht erst eifersüchtig werden , um einzuräumen , dass es harte arbeit ist . stimmt ' s ?
Reference	but you do n't have to be that jealous to concede that it 's hard work . right ?
beam dec	but you do n't need to get to get in order to say that it 's hard work . right ?
relopt dec	but you do n't need to get to get in order to say that it 's hard work . right ?
+bilingual	but you do n't need to become jealous , in order to say that it 's hard work . right ?
Source	wir sind doch alle gute bürger der sozialen medien , bei denen die währung neid ist . stimmt ' s ?
Reference	i mean , we 're all good citizens of social media , are n't we , where the currency is envy ?
beam dec	we 're all great UNK of social media , where the currency is envy . right ?
relopt dec	we 're all great UNK of social media , where the currency is envy . right ?
+bilingual	we 're all good citizens of social media , where the currency is envy . right ?
Source	wir alle wollen alternative energiequellen , die preislich mithalten können . aber es existieren keine .
Reference	we want alternative energy sources that can compete on price . none exist .
beam dec	we all want to have the source of energy sources , which are the things that we want .
relopt dec	we all want to have the source of energy sources , which are the things that we want .
+bilingual	we all want to have alternative sources of energy out , which are the things that we can .
Source	wie lehre ich kindern eine solide arbeitsmoral ?
Reference	what do i do to teach kids a solid work ethic ?
beam dec	how do i teach kids a commercial work ?
relopt dec	how do i teach kids a commercial work ?
+bilingual	how do i teach kids a robust ethic ?
Source	wir anderen liegen dazwischen . übrigens liegt der durchschnittliche bürger fast genau in der mitte .
Reference	the rest of us are in between , and by the way , the average person in the street is almost exactly midway .
beam dec	we are in between them , and by the way , the average citizens are almost right in the middle .
relopt dec	we are in between them , and by the way , the average citizens are almost right in the middle .
+bilingual	we lie in between them . and by the way , the average citizen is almost exactly in the middle .
Source	" ach , ich kann nicht glauben , dass pixar einen prinzessinnen-film gemacht hat . "
Reference	" aw , i ca n't believe pixar made a princess movie . "
beam dec	" oh , i ca n't believe that UNK has a UNK . "
relopt dec	" oh , i ca n't believe that UNK has a UNK . "
+bilingual	" oh , i ca n't believe that pixar has made UNK . "
Source	dieses hier wurde nur wochen nach dem 11. september aufgenommen und ich musste erklären , was an dem tag passiert war , so dass eine fünfjährige es verstehen kann .
Reference	this one was taken just weeks after 9 / 11 , and i found myself trying to explain what had happened that day in ways a five-year-old could understand .
beam dec (l2r)	this one was taken just after a 9 / 11 , and i had to explain what was happening on the day , so that five UNK could understand it .
beam dec (r2l)	this was taken just weeks after 9 / 11 , and i had to explain what happened on that day , so that a five-year-old can understand it .
relopt dec	this one was taken just after a 9 / 11 , and i had to explain what was happening on the day , so that five UNK could understand it .
+bidirectional	this one was taken weeks after the 9 / 11 , and i had to explain what was happening on the day , so that five-year-old could understand it .
Source	mit sieben jahren sah ich zum ersten mal eine öffentliche hinrichtung , aber ich dachte , mein leben in nordkorea sei normal .
Reference	when i was seven years old , i saw my first public execution , but i thought my life in north korea was normal .
beam dec (l2r)	at the age of seven , i first saw a public art , but i thought my life in north korea was normal .
beam dec (r2l)	in seven years , at the first time i was a UNK survivor , but i thought , my life in north korea was normal .
relopt dec	at the age of seven , i first saw a public art , but i thought my life in north korea was normal .
+bidirectional	at the age of seven , i first saw a public execution , but i thought my life in north korea was normal .