A Detailed model description

Word embeddings We embed source and target words in a low dimensional space with embedding matrices $E_S \in \mathbb{R}^{|V_S| \times d_{\text{emb}}}$, $E_T \in \mathbb{R}^{|V_T| \times d_{\text{emb}}}$. Each word vector is initialized at random from $\mathcal{N}(0, \frac{1}{\sqrt{d_{\text{emb}}}})$. We use $d_{\text{emb}} = 512$.

Encoder Our encoder is a one layer bidirectional LSTM with dimension $d_h = 512$. For a source sentence $e = e_1, \ldots, e_{|e|}$ the concatenated output of the encoder is thus of shape $|e| \times 2d_h$.

Attention We use a multilayer perceptron attention mechanism: given a query h_t at step t of decoding and encodings $x_1, \ldots, x_{|e|}$, the context vector c_t is computed according to:

$$\begin{aligned} \alpha_{it} &= V_a^T \tanh(W_a e_i + W_{ah} h_t + b_a) \\ c_t &= \sum_i \alpha_{it} x_i, \end{aligned} \tag{5}$$

where V_a, W_a, W_{ah}, b_a are learned parameters. We choose $d_a = 256$ as the dimension of the intermediate layer.

Decoder The decoder is a single layer LSTM of dimension $d_h = 512$. At each timestep t, it takes as input the previous word embedding w_{t-1} and the previous context c_{t-1}. Its output h_t is used to compute the next context vector c_t and the distribution over the next possible target words w_t:

$$\begin{aligned} o_t &= W_{oh} h_t + W_{oc} c_t + W_{ow} E_{T w_{t-1}} + b_o \\ p_t &= \text{softmax}(E_T o_t + b_T), \end{aligned} \tag{6}$$

where W_{o*}, b_o, b_T are learned parameters, E_T is the target word embedding matrix and $E_{T w_{t-1}}$ is the embedding of the previous target word.

Learning paradigm We employ several techniques to improve training. First, we are using the same parameters for the target word embeddings and the weights of the softmax matrix (Press and Wolf, 2017). This reduces the number of total parameters and in practice this gave slightly better BLEU scores.

We apply dropout (Srivastava et al., 2014) between the output layer and the softmax layer, as well as within the LSTM (using the variant presented in Gal and Ghahramani (2016)). We also drop words in the target sentence with probability 0.1 according to Iyyer et al. (2015). Intuitively, this forces the decoder to use the conditional information.

In addition to this, we implement label smoothing as proposed in Szegedy et al. (2016) with a smoothing coefficient 0.1. We noticed improvements of up to 1 BLEU point with this additional regularization term.

B Training process

We first train each model using the Adam optimizer (Kingma and Ba, 2014) with learning rate 0.001 (we clip the gradient norm to 1). The data is split into batches of size 32 where every source sentence has the same length. We evaluate the validation perplexity after each epoch. Whenever the perplexity doesn't improve, we restart the optimizer with a smaller learning rate from the previous best model (Denkowski and Neubig, 2017). Training is stopped when the perplexity doesn't go down for 3 epochs. We then perform a tuning step: we restart training with the same hyper-parameters except for using simple stochastic gradient descent and gradient clipping at a norm of 0.1, which improved the validation BLEU by 0.3-0.9 points.

C User classifier

In our analysis, we use a classifier to estimate which user wrote each output, which we describe more in this section.

The model uses a continuous bag of n-grams where $v_{\text{n-gram}}$ is a parameter vector for a paticular n-gram and the probability of speaker s for sentence f is given by:

$$\begin{aligned} p(s \mid f) &\propto w_s^T h_f + b_s \\ h_f &= \frac{1}{\#\{\text{n-gram} \in f\}} \Big(\sum_{\text{n-gram} \in f} v_{\text{n-gram}} \Big) \end{aligned} \tag{7}$$

The size of hidden vectors is 128. We limit n-grams to unigrams and bigrams. We estimate the parameters with Adam and a batch size of 32 for 50 epochs.

D Qualitative examples

Table 5 shows examples where our `full_bias/fact_bias` model helped translation by favoring certain words as opposed to the baseline in `en-fr`.

Talk	Andrew McAfee : What will future jobs look like?
Source	but the middle class is clearly under huge threat right now .
Reference	mais la classe moyenne fait aujourd' hui face à une grande menace .
base	mais la classe moyenne est clairement une menace énorme en ce moment .
full_bias	mais la classe moyenne est clairement maintenant dans une grande menace .
fact_bias	mais la classe moyenne est clairement en grande menace en ce moment .
Talk	Olafur Eliasson : Playing with space and light
Source	the show was , in a sense , about that .
Reference	le spectacle était , dans un sens , à propos de cela .
base	le spectacle était , en un sens , à propos de ça .
full_bias	le spectacle était , dans un sens , à propos de cela .
fact_bias	le spectacle était , dans un sens , à ce sujet .
Talk	Lona Szabo de Carvalho : 4 lessons I learned from taking a stand against drugs and gun violence
Source	we need to make illegal drugs legal .
Reference	nous avons besoin de rendre les drogues illégales , légales .
base	nous devons faire des médicaments illégaux .
full_bias	nous devons faire des drogues illégales .
fact_bias	nous devons produire des drogues illégales .
Talk	Wade Davis: On the worldwide web of belief and ritual
Source	a people for whom blood on ice is not a sign of death , but an affirmation of life .
Reference	un peuple pour qui du sang sur la glace n' est pas un signe de mort mais une affirmation de la vie .
base	une personne pour qui le sang sur la glace n' est pas un signe de mort , mais une affirmation de la vie .
full_bias	un peuple pour qui le sang sur la glace n' est pas un signe de mort , mais une affirmation de la vie .
fact_bias	un peuple pour qui sang sur la glace n' est pas un signe de mort , mais une affirmation de vie .

Table 5: Examples where our proposed method helped improve translation.