A Experimental setup

A.1 Data preprocessing

We use the publicly available OpenSubtitles2018 corpus (Lison and Tiedemann, 2016) for English and Russian.[1] We pick sentence pairs with an overlap of at least 0.9 to reduce noise in the data. For context, we take the previous sentence if its timestamp differs from the current one by no more than 7 seconds.

We use the tokenization provided by the corpus.

Sentences were encoded using byte-pair encoding (Sennrich et al., 2016), with source and target vocabularies of about 32000 tokens. Translation pairs were batched together by approximate sequence length. Each training batch contained a set of translation pairs containing approximately 5000 source tokens.

A.2 Model parameters

We follow the setup of Transformer base model (Vaswani et al., 2017). More precisely, the number of layers in the encoder and decoder is $N = 6$. We employ $h = 8$ parallel attention layers, or heads. The dimensionality of input and output is $d_{model} = 512$, and the inner-layer of a feed-forward networks has dimensionality $d_{ff} = 2048$.

We use regularization as described in (Vaswani et al., 2017).

A.3 Optimizer

The optimizer we use is the same as in (Vaswani et al., 2017). We use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\varepsilon = 10^9$. We vary the learning rate over the course of training, according to the formula:

$$l_{rate} = d_{model}^{0.5} \cdot \min(step\_num^{0.5}, step\_num \cdot warmup\_steps^{1.5})$$

We use $warmup\_steps = 4000$.

[1] http://opus.nlpl.eu/OpenSubtitles2018.php