## A Appendix 1: Training Details

**Training details** Both backward and forward models were three-layer LSTMs with $1,024$ hidden cells for each layer. There were $512$ hidden cells in the encoder because it was bi-directional. The embedding size was set to $768$, while the vocabulary size was set to $50,000$. The batch size was chosen from [$64$, $128$]. The learning rate was set to $0.0001$, and the Adam optimizer was used. All parameters were initialized by sampling from the normal distribution of mean $0$ and variance $1$. The gradients were clipped to avoid gradient explosion with a threshold of $5$. We used pre-trained word embeddings from BERT (Devlin et al., 2018). With respect to the word graph approach, we used this implementation[12]. The number of candidates for the word graph approach was chosen from [$50$, $200$]. and the minimal number of tokens for the compression was set to $10$.

## B Appendix 2: Human Evaluation Details

We followed previous works (Barzilay and McKeown, 2005; Filippova, 2010) and asked the raters to provide three ratings (points): excellent (2 points) if the generated compression was a completely grammatical sentence; good (1 point) if the generated compression was basically readable but required minor corrections, and ungrammatical (0 point) if it is none of the above. For informativeness: excellent (2 points) if the generated compression conveyed the gist of the main event or topic, good (1 point) if it was related to the main theme, but misses something important; and unrelated (0 point) if the generated compression was not related to the main theme.

[12] https://github.com/boudinfl/takahe