Multi-task Model Training

Task 1: Multilingual Skip-gram (similar to [2])

![Figure 1: Example context attachments for a bilingual skip-gram model (en-de).](image)

Task 2: Cross-lingual Sentence Similarity

![Figure 2: Architecture of the Sentence Encoder that we use for computing sentence representations Rs and Rt for input sentences S and T.](image)

Training Routines

- **JMT-Sent-LSTM**: Model is trained by alternating between mini-batches of the two tasks.
- **JMT-Sent-Avg**: Proposed joint multi-task model but does not include an LSTM layer in the sentence encoder.
- **Sent-LSTM** and **Sent-Avg** are the single-task variants of these models.

Data

- 500k parallel sentences for each language pair from Europarl Corpus.
- Additional 500k monolingual sentences for JMT models
- Vocabulary sizes for German (de) and English (en) are respectively 39K and 21K in the parallel corpus, 120K and 68K in the combined corpus
- Evaluated on the RCV1/RCV2 cross-lingual document classification task (same data splits as in literature)

Results

We construct document embeddings by averaging sentence representations produced by a trained sentence encoder.

![Table 1: Results for models trained on en-de language pair. *no-mono means no monolingual data was used in training.](image)

- **JMT-Sent-LSTM** model outperforms systems compared at 128 dimensions.
- When sentence embedding dimension is 512, our results are close to the best results from literature.
- Models with an LSTM layer perform better than those without one.
- Ablation experiments (*no-mono) suggest that gains are partly due to the addition of monolingual data.

References
