A Multi-task Approach to Learning Multilingual Representations

Karan Singla[1] (singlak@usc.edu), Dogan Can[1], Shrikanth Narayanan[1]

[1]Signal Analysis and Interpretation Lab (SAIL), University of Southern California

USC University of Southern California

Summary

Our system learns word and sentence embeddings jointly by training a multilingual skip-gram model together with a cross-lingual sentence similarity model

Highlights

- Uses both monolingual and bilingual parallel corpora to learn multilingual embeddings
- BiLSTM layer to contextualize word embeddings
- Trained end-to-end
- Shows competitive performance in a standard cross-lingual document classification task using limited resources
- Can capture the similarity between words in different languages even if they are not present in the bilingual corpora (see Figure 3)

Multi-task Model Training

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

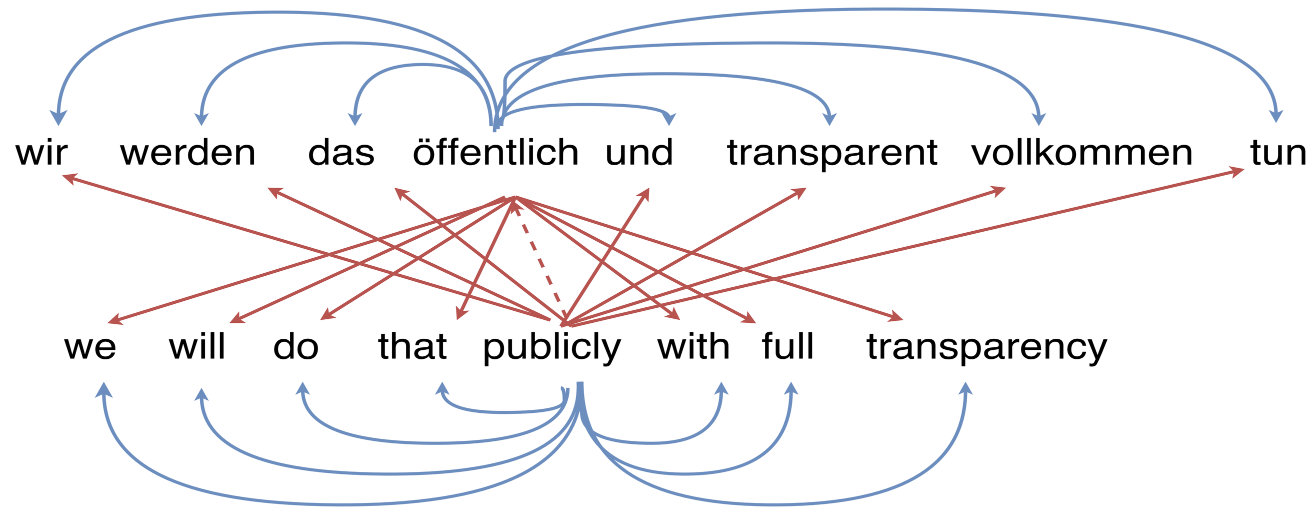

Figure 1: Example context attachments for a **bilingual skip-gram** model (en-de).

Task 2: Cross-lingual Sentence Similarity

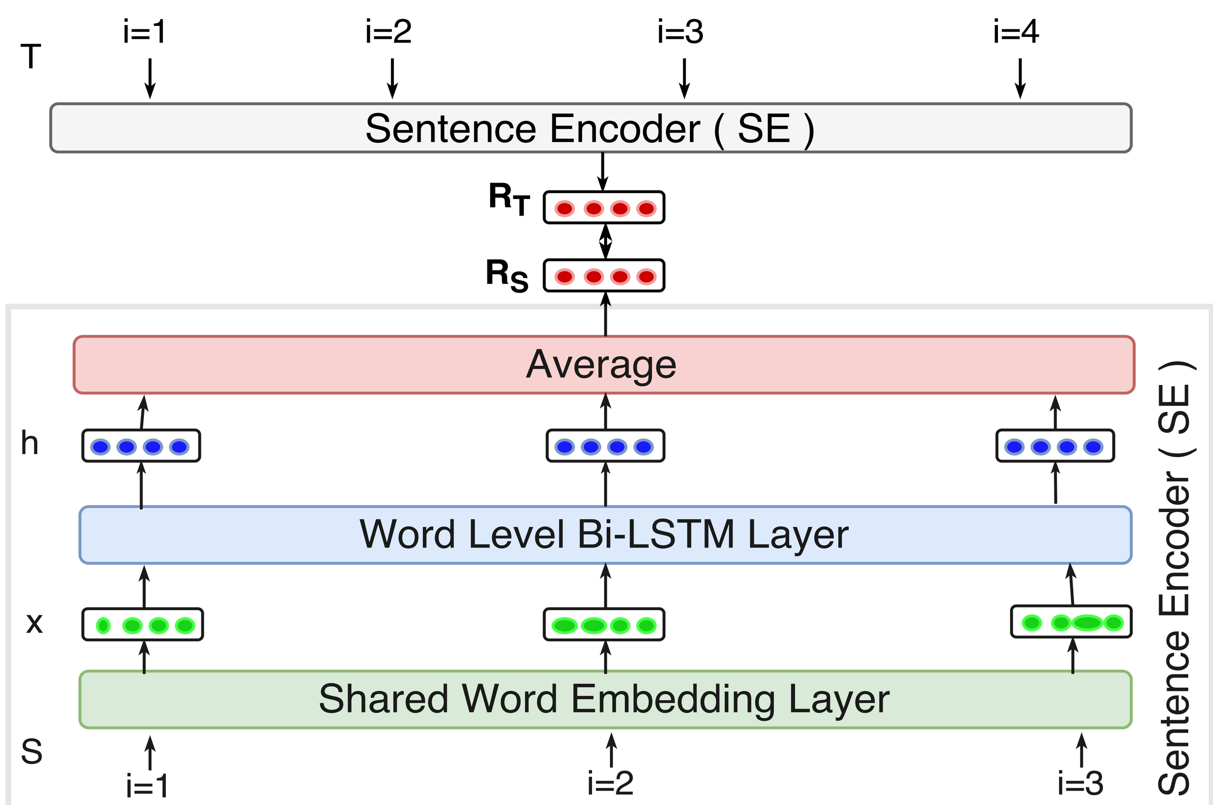

Figure 2: Architecture of the **Sentence Encoder** that we use for computing sentence representations R_S and R_T for input sentences S and T.

$$\textbf{Loss} : l(S,T) = \sum_{i=1}^{k} \max(0, m + d(S,T) - d(S,N_i))$$

Without the LSTM layer, this loss is similar to the BiCVM loss [1]

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.

Model	en → de	de → en
500k parallel sentences, dim=128		
BiCVM-add+ [1]	86.4	74.7
BiCVM-bi+ [1]	86.1	79.0
BiSkip-UnsupAlign [2]	88.9	77.4
Our Models		
Sent-Avg	88.2	80.0
JMT-Sent-Avg	88.5	80.5
Sent-LSTM	89.5	80.4
JMT-Sent-LSTM	**90.4**	**82.2**
JMT-Sent-Avg*no-mono	88.8	80.3
JMT-Sent-LSTM*no-mono	89.5	81.5
100k parallel sentences, dim=128		
Sent-Avg	81.6	75.2
JMT-Sent-Avg	85.3	79.1
Sent-LSTM	82.1	76.0
JMT-Sent-LSTM	87.4	80.7
JMT-Sent-LSTM*no-mono	83.4	76.5

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

- 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.

Example Word Embeddings

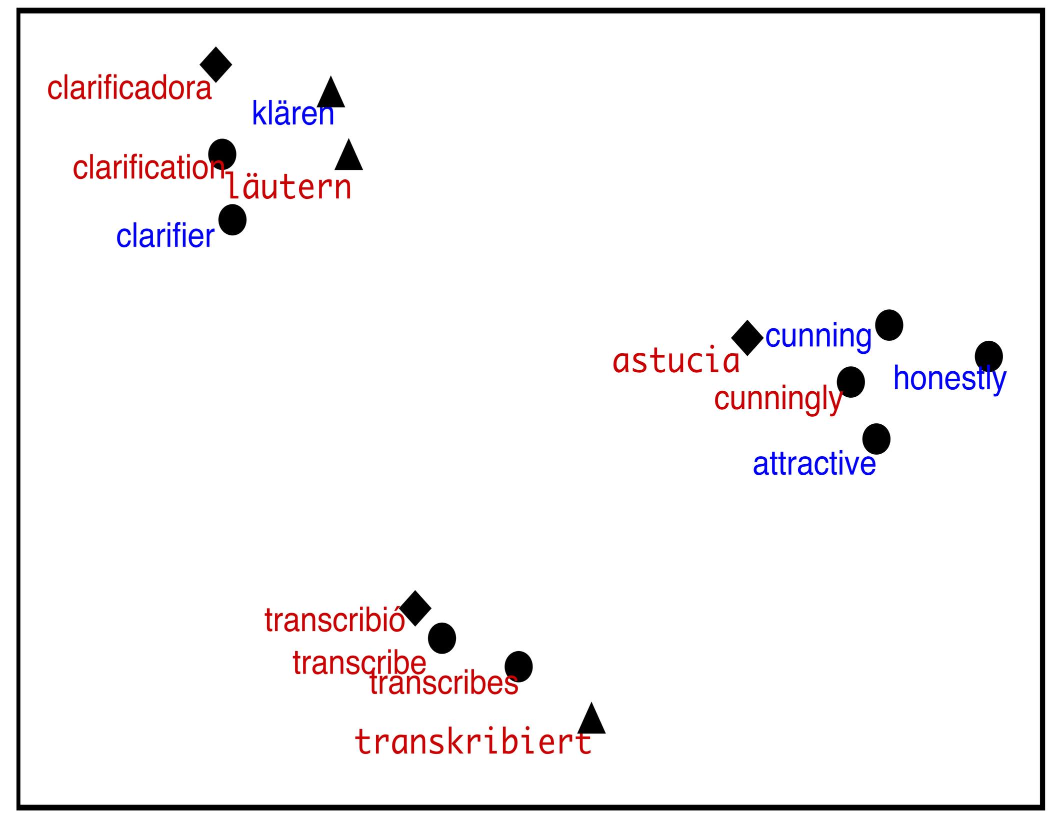

Figure 3: t-SNE projections for 3 English words (clarification, transcribe, cunningly) and their nearest neighbors. **Red** words are in the monolingual corpora only. **Blue** words are in both the monolingual and the parallel corpora. Points are styled based on language.

Monolingual vs Parallel Data (en-de, dim=128)

Mono \ Parallel	20K	50K	100K	500K
no-mono	60.3	68.3	82.1	89.5
20K	57.4	68.7	80.2	89.5
50K	**62.7**	69.0	83.5	89.5
100K	61.5	71.9	85.1	89.6
200K	58.1	**72.1**	85.5	90.0
500K	52.6	64.8	**87.4**	**90.4**

JMT-Sent-LSTM produces better embeddings as long as the amount of additional monolingual data is not too large or small.

Multilingual vs Bilingual* Models (dim=128)

Model	en-es	en-de	de-en	es-en	es-de
Sent-Avg	49.8	86.8	78.4	63.5	69.4
Sent-LSTM	53.1	89.9	77.0	67.8	65.3
JMT-Sent-Avg	51.5	87.2	75.7	60.3	**72.6**
JMT-Sent-LSTM	**57.4**	**91.0**	75.1	63.3	68.1
JMT-Sent-LSTM*	54.1	90.4	**82.2**	**68.4**	-

Multilingual models perform better than bilingual ones when English is the source language

References

[1] K. M. Hermann and P. Blunsom. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*, 2014.

[2] T. Luong, H. Pham, and C. D. Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, 2015.

http://sail.usc.edu

Supported by the U.S. DARPA LORELEI Program