On Target Representation in Continuous-output Neural Machine Translation

Overview

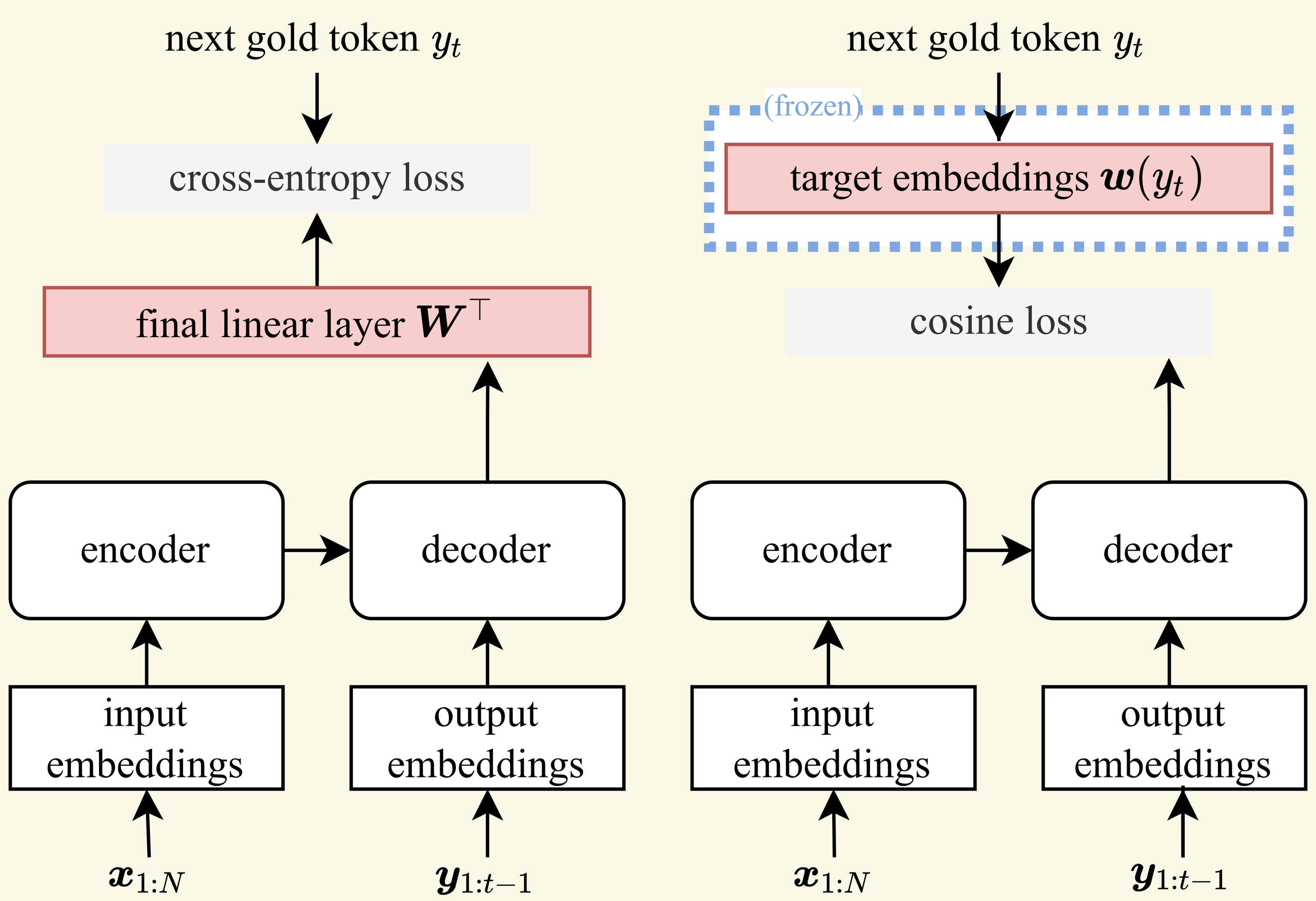
- NMT models are typically discrete
- Can they be continuous?
  - Yes, by learning to predict word embeddings directly
  - No moving target: must choose good embeddings
- This work:

  How to choose target embeddings?

Background

- **Output layer**: treat hidden states as embeddings
- **Objective function**: cosine similarity between target and output embeddings
- **Decoding**: Nearest Neighbors search with $K = 1$

Parallels between
the discrete (left) and continuous (right) Transformers:



Target Embeddings

Types of embeddings used in our analysis:

Data	Euclidean	Non-Euclidean
Monolingual	fastText	JoSe($\mathbb{S}$)
Bilingual	MT-transfer	MT-transfer($\mathbb{S}$)
External	fastText mBART	

fastText is pretrained with subword information
mBART is fine-tuned on NMT many-to-many data

Results

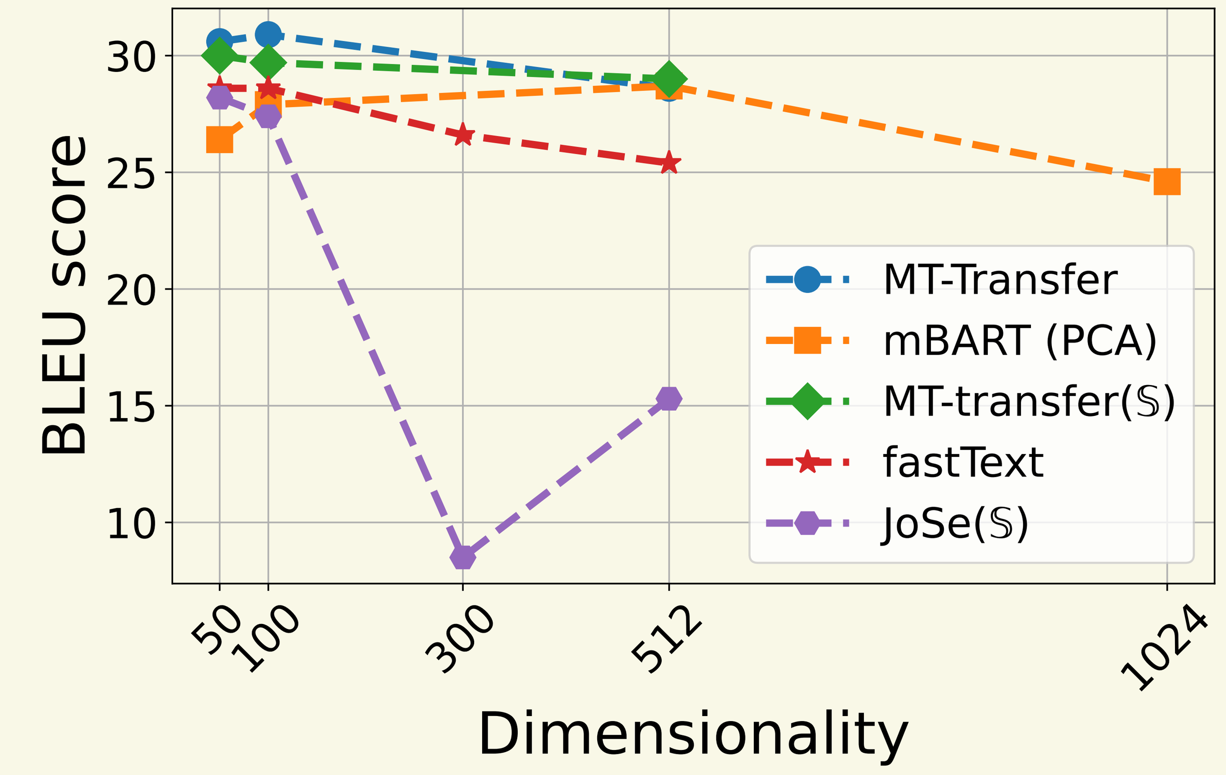
BLEU scores on newstest data.

embeddings	Ro→En *dim*	Ro→En test16	En→Tr *dim*	En→Tr test16	En→Tr test17
discrete		31.6		12.2	12.2
+beam=5		**32.3**		**12.8**	**13.0**
Trained on target monolingual data					
fastText	*(100)*	28.6	*(100)*	9.6	9.5
JoSe ($\mathbb{S}$)	*(50)*	28.2	*(50)*	9.4	9.9
Trained on bilingual data					
MT-transfer	*(100)*	**30.9**	*(50)*	8.6	8.9
MT-transfer ($\mathbb{S}$)	*(50)*	30.0	*(100)*	**11.2**	**11.6**
Pretrained on external data					
fastText◆	*(300)*	27.0	*(300)*	9.1	9.3
$\text{fastText}_{\text{PCA}}$	*(100)*	28.6	*(100)*	9.3	9.5
mBART-MT◆	*(1024)*	24.6	*(1024)*	0	0
$\text{mBART-MT}_{\text{PCA}}$	*(512)*	28.7	*(100)*	9.2	9.8

Embedding dimension is chosen
on the dev set, except for the fixed pretrained models (◆)

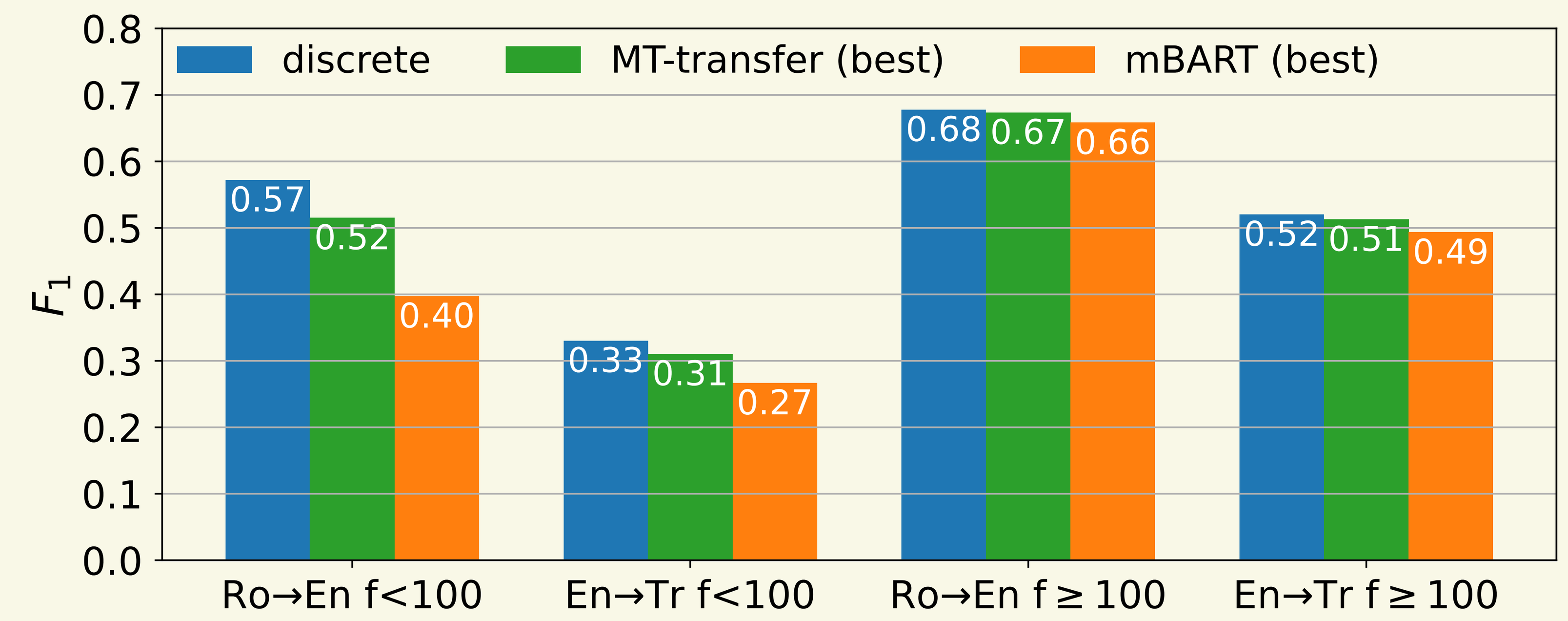
Embeddings Dimensionality

Lower dimensions is often better (Ro→En, test16):



Rare Words

Word-level F_1 score by word training frequency (f):



Conclusion

- Choice of target embeddings matters ()
- Dimensionality and geometry plays important role ()
- Large-scale pretraining ( )
  is not superior to MT-data only
- MT-Transfer embeddings
  outperforms all other embedding choices ()

Evgeniia Tokarchuk, Vlad Niculae

e.tokarchuk@uva.nl, v.niculae@uva.nl

University of Amsterdam · Language Technology Lab

VN partially funded by the NWO Zwaartekracht Hybrid Intelligence Centre

UvA