Debiasing Word Embedding Improves Multimodal Machine Translation

Tosho Hirasawa, Mamoru Komachi
Tokyo Metropolitan University, hirasawa-tosho@ed.tmu.ac.jp

1 Task & Background

Multimodal Neural Machine Translation (MMT)

- MT using non-linguistic information
- WMT Multimodal Shared Task
  - Inputs: source sentence + image
  - Output: target sentence

Problem

- Only small amount of training data is available (≈30k).
  - Poor performance for translating less-frequent words.

Contributions

- GloVe word embeddings are useful for various multimodal NMT models irrespective of the extent to which visual features are used in them.
- All-but-the-Top debiasing technique for pretrained word embeddings to further improve multimodal NMT models.

2 Related works

Pretrained Word Embeddings for NMT [Qi et al., 2018]

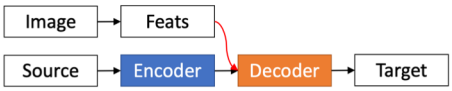
- Use pretrained word embedding to initialize NMT model.
- Better performance in low-resource scenario; while decreasing as the data size is growing up.
- Distance language pairs receive more profit.
- Use vanilla pretrained word embedding. Debiasing is not processed.

All-but-the-Top [Mu and Viswanath, 2018]

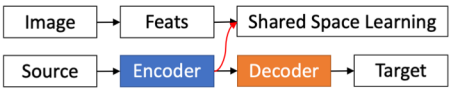
- Reduce the hubness of pretrained word embedding.
  - Hubness: word embeddings are used in the k-nearest neighbor (k-NN) problem, certain words appear frequently in the k-nearest neighbors for other words.
- Better performance for word similarity tasks and clustering tasks.
- Neural models are tested in a few scenario.

3 MMT model / Word embedding / Debiasing

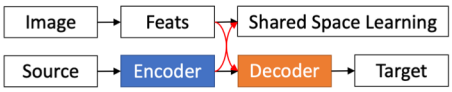
Doubly-Attentive: Visual-aware decoder



IMAGINATION: Multi-task learning

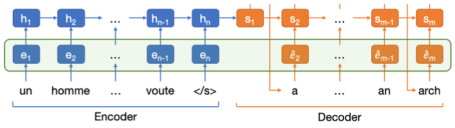


VAG-NET: Mixed model



Model initialization

- Initialize embedding layers in both encoder and decoder.



Pretrained word embeddings

- word2vec, GloVe, and FastText.
- Debiased after pretraining.
- OOV words are calculated as the average embedding over words that are a part of pretrained embeddings but are not included in the vocabularies.

Localized Centering [Hara et al., 2015]

- Debias using the local bias of each word
  1. Retrieve k-nearest neighbors (k-nn) of each word using cosine similarity.
  2. Subtract the mean vector of k-nn from original representation to get the debiased representation.

All-but-the-Top [Mu and Viswanath, 2018]

- Debias using the global bias of corpus
  1. Subtract the centroid of all words from each word.
  2. Compute the PCA components for the centered space.
  3. Subtract the top 3 PCA components from each centered word.

4 Experiment

Dataset: Multi30K (English -> German/French)

Model (RNN-based)

- Vocabulary: 10,000
- Embedding: 300D
- Encoder: bi-GRU, 256D
- Decoder: 256D
- Attention: 256D
- beam: 1

Pretrained word embedding

- Dataset: Wikidump
- word2vec: CBOW
- Glove: window size 10
- FastText: CBOW, 5-gram

Training

- Optimizer: Adam
- Learning rate: 0.0004
- Batch size: 64
- Dropout: 0.3

	Lines	Types	Tokens
English	96M	10M	2347M
German	35M	11M	829M
French	39M	4M	703M

Statistics of Wikdump (January 20, 2019)

5 Results

English -> German (METEOR, average of 3 runs)

Model (random)	Debiasing	None	LC	AbtT
NMT (54.68)	word2vec	52.71	53.10	52.09
	GloVe	54.75	**55.30**	55.21
Doubly-Attention (52.37)	word2vec	50.71	51.06	50.71
	GloVe	53.40	53.53	**54.39**
IMAGINATION (54.18)	word2vec	52.32	52.48	52.86
	GloVe	**55.08**	54.88	**55.08**
VAG-NET (55.07)	word2vec	53.11	52.53	52.43
	GloVe	55.27	54.51	**55.66**

English -> French (METEOR, average of 3 runs)

Model (random)	Debiasing	None	LC	AbtT
NMT (72.57)	word2vec	70.44	71.15	70.58
	GloVe	**73.84**	72.86	73.38
Doubly-Attention (71.16)	word2vec	67.84	71.15	68.29
	GloVe	71.84	71.20	**72.02**
IMAGINATION (72.44)	word2vec	70.80	70.97	71.03
	GloVe	72.71	73.30	**73.39**
VAG-NET (72.59)	word2vec	71.17	71.04	71.75
	GloVe	**73.44**	73.31	73.36

6 Discussion / Examples

Word Embedding (word2vec, GloVe, FastText)

- GloVe performs the best for MMT models.
- word2vec are reported to be cohesively clustered and not evenly distributed, making models to learn from some specific values.
- FastText learns both words and their subwords and requires more training data to get competitive performance.

Debiasing (Localized Centering, All-but-the-Top)

- All-but-the-Top improves most of models.

Languages (English -> German, English -> French)

- Better performance for distant language pairs (English -> German).

Translation examples

Source	a young boy wearing a blue jersey and yellow shorts is playing soccer .
Reference	ein junge in einem blauen trikot und gelben shorts spielt fußball .
VAG	ein junge in blauem trikot und gelben shorts spielt fußball .
VAG (GloVe)	ein junge in einem blauen trikot und gelben shorts spielt fußball .