Beyond Shared Vocabulary: Increasing Representational Word Similarities across Languages for Multilingual Machine Translation

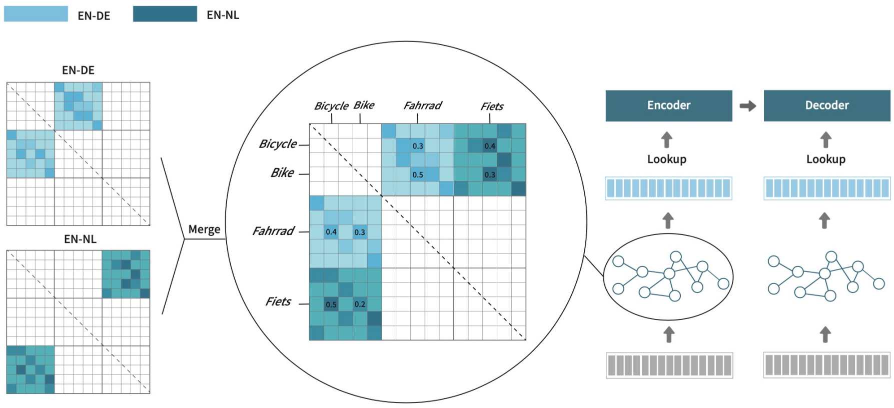
Di Wu, Christof Monz

Language Technology Lab, Univeristy of Amsterdam

Introduction

1. Shared vocabulary is common practice:
   a. Multilingual Translation, mBert, LLama, GPT, ...
2. Shared vocabulary is good:
   a. Simple design, easy to scale
   b. **Word overlap** encourge knowledge transfer, when they refer to similar meanings across languages
3. But has limitations:
   a. **When languages use different writing system, there is little word overlap and knowledge transfer suffers.**
   b. Even if language use similar writing systems, shared tokens may have completely different meanings.
4. **What we do:**
   a. **Mine priors of word equivalence based on word alignments, then model them into a graph.**
   b. **Inject such priors into embedding table via graph networks, thereby enhancing knowledge transfer.**

Reparameterization Framework



$$g_{i,j}^{s,t} = \frac{c_{i,j}^{s,t}}{\sum_{k=1}^{|V|} c_{i,k}^{s,t}} \Rightarrow G = \sum_{D_{s,t} \subseteq D} G^{s,t} \Rightarrow E^{h+1} = \rho(E^h W_1^h + GE^h W_2^h + B^h)$$

Experiments

1. IWSLT14: 8 English-centric language pairs, the size of each is range from 89K to 169K
2. EC30: 30 language pairs, 5 different writing systems, High (5M), Medium (1M), Low (100K)
3. **Results in short:**
   a. **High-level consistent improvement: 1) for all of the language direction, and 2) as graph networks goes deeper.**
   b. **Zero-shot: Also, get improved.**
   c. **Ablation: Tying new embeddings with the decoder's projection matters**

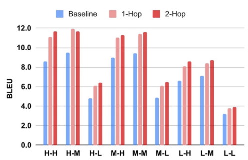
Results on IWSLT14 & Ablation

Model	DE	ES	FA	AR	HE	NL	PL	IT	EN→X	X→EN	AVG
Baseline (Lin et al., 2021)	28.1	35.2	16.9	20.9	29.0	30.9	16.4	29.2	-	-	25.8
LASS (Lin et al., 2021)	29.8	37.3	17.9	22.9	30.9	33.0	17.9	30.9	-	-	27.6
Our Baseline	28.5	36.0	17.4	20.2	27.9	31.5	17.6	29.7	24.4	27.8	26.1
Weighted Sum	29.2	36.7	18.1	20.9	28.5	32.2	18.2	30.5	24.8	28.7	26.8
GraphMerge-1hop	30.2	37.5	19.0	21.7	30.0	33.4	18.8	31.3	25.4	30.0	27.7
GraphMerge-2hop	30.4	37.9	19.0	21.9	30.0	33.7	19.2	31.6	**25.5**	30.5	28.0
GraphMerge-3hop	**30.7**	**38.2**	**19.9**	**22.3**	**30.1**	**34.0**	**19.4**	**32.2**	25.4	**31.3**	**28.4**
3-hop Gain	+2.2	+2.2	+2.5	+2.1	+2.2	+2.5	+1.8	+2.5	+1.0	+3.5	+2.3

Settings	EN→X	X→EN	AVG
Baseline	24.4	27.8	26.1
1-hop	25.4	30.0	27.7
1-hop w/o Tie	25.4	28.7	27.0
2-hop	25.5	30.5	28.0
2-hop w/o Tie	25.1	29.6	27.4
3-hop	25.4	31.3	28.4
3-hop w/o Tie	25.3	29.4	27.4
2-hop	25.5	30.5	28.0
eflomal → FastAlign	25.4	30.1	27.8
intersect → gdfa	25.2	29.9	27.6

Results on EC30: English-centric & Zero-shot

Model	High		Medium		Low		ALL		
	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN	EN→X	X→EN	AVG
Baseline (Trans.-Big)	28.7	31.3	31.0	31.4	20.0	25.6	26.5	29.4	28.0
GraphMerge-1hop	29.5	32.0	31.7	31.8	20.6	27.0	27.3	30.3	28.8
GraphMerge-2hop	**29.7**	**32.2**	**32.0**	**32.0**	20.9	**27.4**	**27.6**	**30.5**	**29.1**
GraphMerge-3hop	29.4	31.8	32.0	31.9	**21.0**	27.4	27.5	30.4	29.0
2-hop Gain	+1.0	+0.9	+1.0	+0.6	+0.9	+1.8	+1.1	+1.1	+1.1



Analysis: Performance & Word Similarity

1. Settings:
   a. Using MUSE (bilingual dictionary) as ground truth:
   b. Estimation similarities between equivalent words in MUSE
2. High-level Consistence:
   a. Deeper Graph -> Better Crosslinguality
   b. Better Crosslinguality -> Higher BLEU
   c. Works for all of the language pairs
3. Beyond English-Centric Word Similarity:
   a. Consistently works as well
   b. Transfer beyond English-centric language pairs, even though only English-centric data are leveraged.

English-Centric Cross-lingual Word Similarity

Model	EN↔ DE		EN↔NL		EN↔AR		EN↔HE	
	Similarity	BLEU	Similarity	BLEU	Similarity	BLEU	Similarity	BLEU
Baseline	0.24	28.5	0.25	31.5	0.23	20.2	0.23	27.9
GraphMerge-1hop	0.35	30.2	0.37	33.4	0.32	21.7	0.32	30.0
GraphMerge-2hop	0.42	30.4	0.44	33.7	0.38	21.9	0.38	30.0
GraphMerge-3hop	**0.46**	**30.7**	**0.48**	**34.0**	**0.41**	**22.4**	**0.41**	**30.1**
Model	EN↔ES		EN↔FA		EN↔PL		EN↔IT	
	Similarity	BLEU	Similarity	BLEU	Similarity	BLEU	Similarity	BLEU
Baseline	0.25	36.0	0.22	17.4	0.24	17.6	0.27	29.7
GraphMerge-1hop	0.38	37.5	0.31	19.0	0.35	18.8	0.40	31.3
GraphMerge-2hop	0.45	37.9	0.37	19.0	0.43	19.2	0.48	31.6
GraphMerge-3hop	**0.49**	**38.2**	**0.40**	**19.9**	**0.47**	**19.4**	**0.52**	**32.2**

Beyond English-Centric Cross-lingual Word Similarity

Model	DE↔NL	DE↔AR	DE↔HE	NL↔AR	NL↔HE	AR↔HE
Baseline	0.29	0.23	0.25	0.24	0.26	0.29
GraphMerge-1hop	0.36	0.28	0.30	0.30	0.31	0.33
GraphMerge-2hop	0.42	0.32	0.34	0.35	0.35	0.37
GraphMerge-3hop	**0.47**	**0.36**	**0.38**	**0.39**	**0.39**	**0.41**

Analysis: Speed & Memory

1. Extra Latency:
   a. Limited
   b. Consistant when vocabulary is fixed
2. Extra Params and Memory:
   a. Sparse, so they are "nothing"
3. Easy to scale:
   a. Works for BIG model
   b. Works for BIG vocalbuary

Model	WPS	Times
Transformer (30K)	201,378	1.00
GraphMerge-1hop	192,367	1.04
GraphMerge-2hop	188,851	1.06
Transformer-Big (128K)	69,702	1.00
GraphMerge-1hop	43,416	1.61
GraphMerge-2hop	33,912	2.05
Model	**Params**	**Times**
Transformer (30K)	62.3M	1.00
GraphMerge-1hop	63.3M	1.01
GraphMerge-2hop	64.4M	1.02
Transformer-Big (128K)	438.6M	1.00
GraphMerge-1hop	442.8M	1.01
GraphMerge-2hop	447.0M	1.02

Conclusion

1. Broadly Speaking:
   a. Mine meaning-equivalent priors and inject into embeddling table.
2. Goodness:
   a. A framework to reparameteriz embedding table for better multilinguality, resulting in significant transfer improvements.
   b. Leading to consistent improvements in MNMT.
   c. Remain Practical as well:
      i. Through multi-hop mechanism, the pivot language bridge the way of knowledge transfer amony non-English-centric pairs.
      ii. A negligible number of additional parameters and computational cost.
      iii. Identical inference cost via storing the re-parameterized embeddings for online systems.