Multilingual Neural Machine Translation with Soft Decoupled Encoding

Carnegie Mellon University

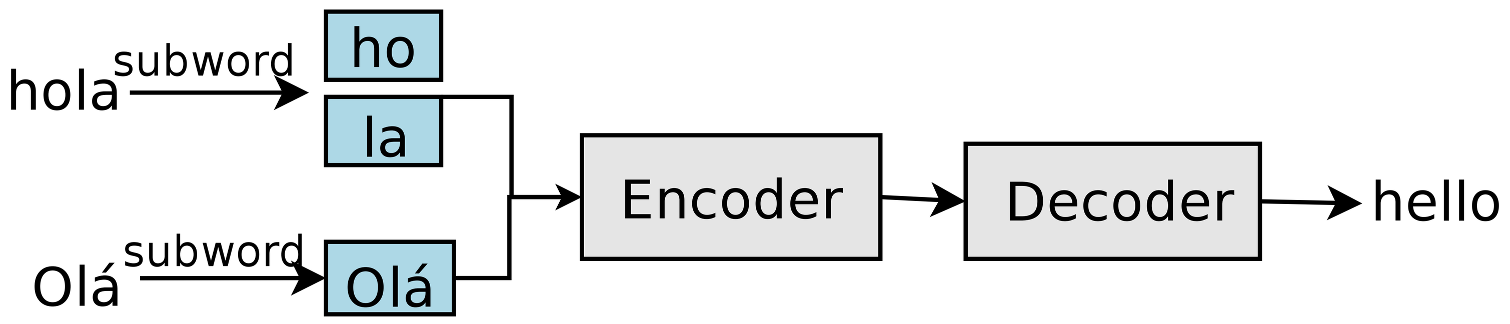
Xinyi Wang [1] Hieu Pham [1,2] Philip Arthur [3] Graham Neubig [1]

(1): Carnegie Mellon University (2): Google Brain (3): Monash University

MONASH University

Multilingual Neural Machine Translation

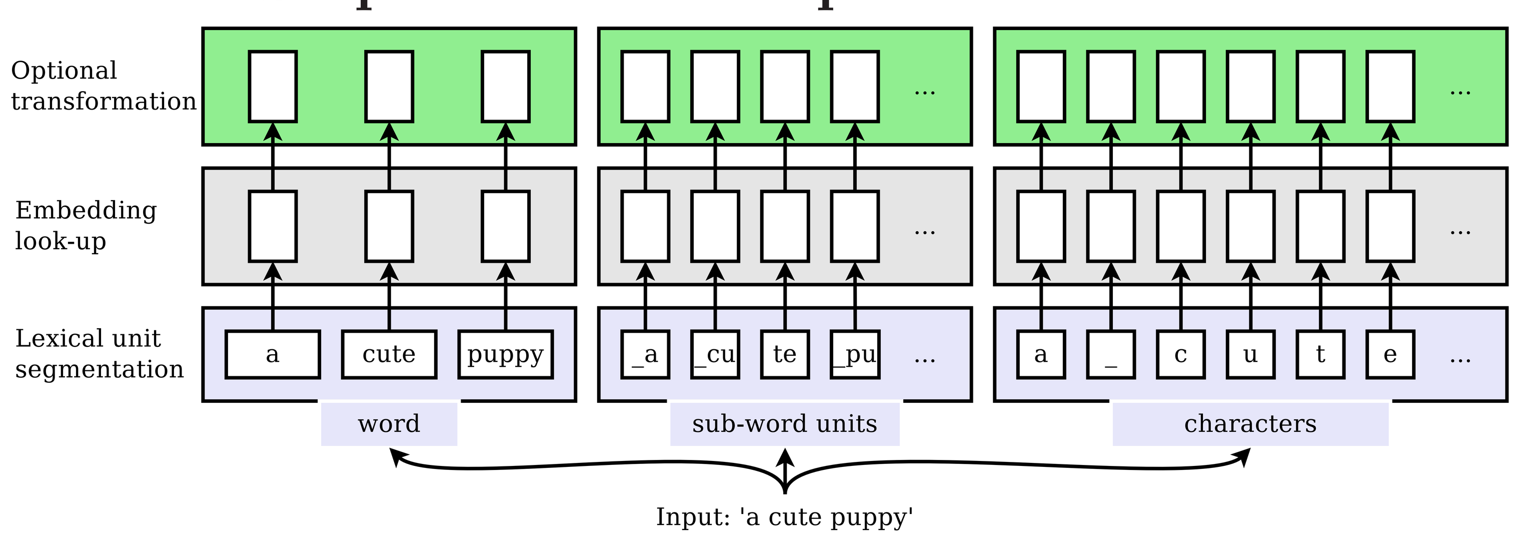
- **Parameter-efficient for low-resource languages**



- **Subword is suboptimal for multilingual NMT**
  - Joint: low-resource language gets bad segmentation
  - Separate: little lexical overlap between languages
- **Training subwords multilingually is difficult**

Lexical Representations

- **What is a good lexical representation?**
  - Accurate representation: words of similar meaning close to each other
  - Maximize parameter sharing between languages
- **Three Steps for lexical representation**



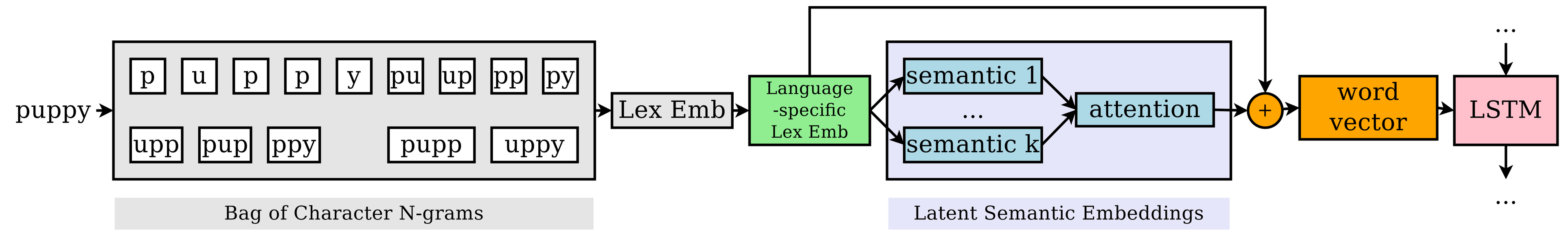
- **Existing Methods**

Method	Lex Unit	Embedding	Encoding
Johnson et. al. 2016	Subword	Lookup	Identity
Lee et. al. 2017	Character	Lookup	Identity
Gu et. al. 2018	Subword	Lookup	Lookup + Latent
Ataman et. al. 2018	Word	character n-gram	Identity

- **Their Properties**

Method	Acc	LexShare	Speed
Johnson et al. 2016	😐	☹	🙂
Lee et al. 2017	😐	🙂	☹
Gu et al. 2018	😐	🙂	🙂
Ataman et al. 2018	😐	🙂	☹
SDE (ours)	🙂	🙂	🙂

Soft Decoupled Encoding (SDE)



- **1: Lexical Embedding**

$$c(w) = \tanh(\mathrm{BoN}(w) \cdot \mathbf{W}_c).$$

  - Represents spelling
  - Fr: couleur; En: color

- **2: Language Specific Transform**

$$c_i(w) = \tanh(c(w) \cdot \mathbf{W}_{L_i}),$$

  - Captures consistent spelling shift
  - Cs: Kryštof; En: Chrsitopher

- **3: Latent Semantic Embedding**

$$e_{\text{latent}}(w) = \mathrm{Softmax}(c_i(w) \cdot \mathbf{W}_s^\top) \cdot \mathbf{W}_s.$$

  - Semantic meaning shared by all langs
  - Fr: bonjour; En: hello

- **4: Residual Connection**

$$e_{\text{SDE}}(w) = e_{\text{latent}}(w) + c_i(w).$$

  - Combines lexical and semantic meaning

Experiments

- **Datasets**
  - TED: 58 langs to Eng
  - 4 low-resource languages
  - paired with high-resource

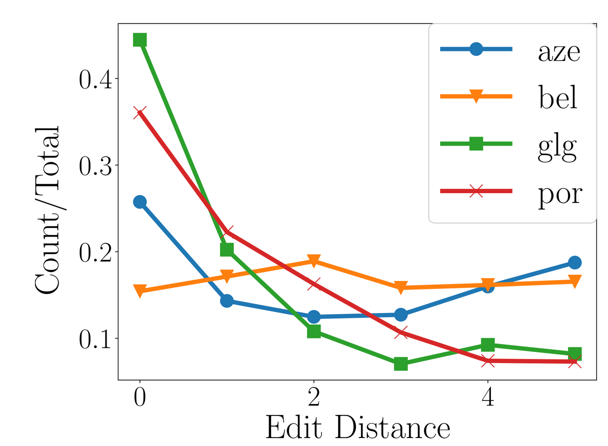
LRL	Train	Dev	Test	HRL	Train
aze	5.94k	671	903	tur	182k
bel	4.51k	248	664	rus	208k
glg	10.0k	682	1007	por	185k
slk	61.5k	2271	2445	ces	103k

- **Main Results**

Lex Unit	Model	aze	bel	glg	slk
Word	Lookup	7.66	13.03	28.65	25.24
Sub-joint	Lookup	9.40	11.72	22.67	24.97
Sub-sep	Lookup (re-imp Neubig et. al.)	10.90	16.17	28.10	28.50
Sub-sep	UniEnc (re-imp Gu et. al.)	4.80	8.13	14.58	12.09
Word	SDE	**11.82***	**18.71***	**30.30***	**28.77†**

- **Ablations**

Model	aze	bel	glg	slk
SDE	11.82	18.71	30.30	28.77
−Lang-Specific Trans.	12.89*	18.13†	30.07	29.16†
−Latent-Sem Emb.	7.77*	15.66*	29.25*	28.15*
−Lexical Emb.	4.57*	8.03*	13.77*	7.08*



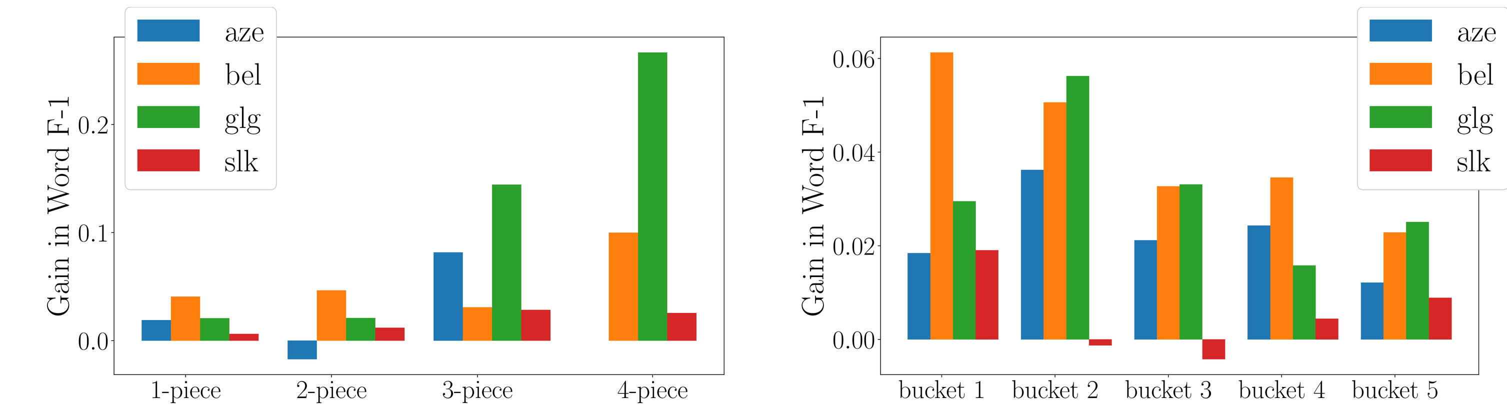
- Lexical embedding has the largest effect
- Lang-specific transform helps more for lang pairs with similar lexicons

Code

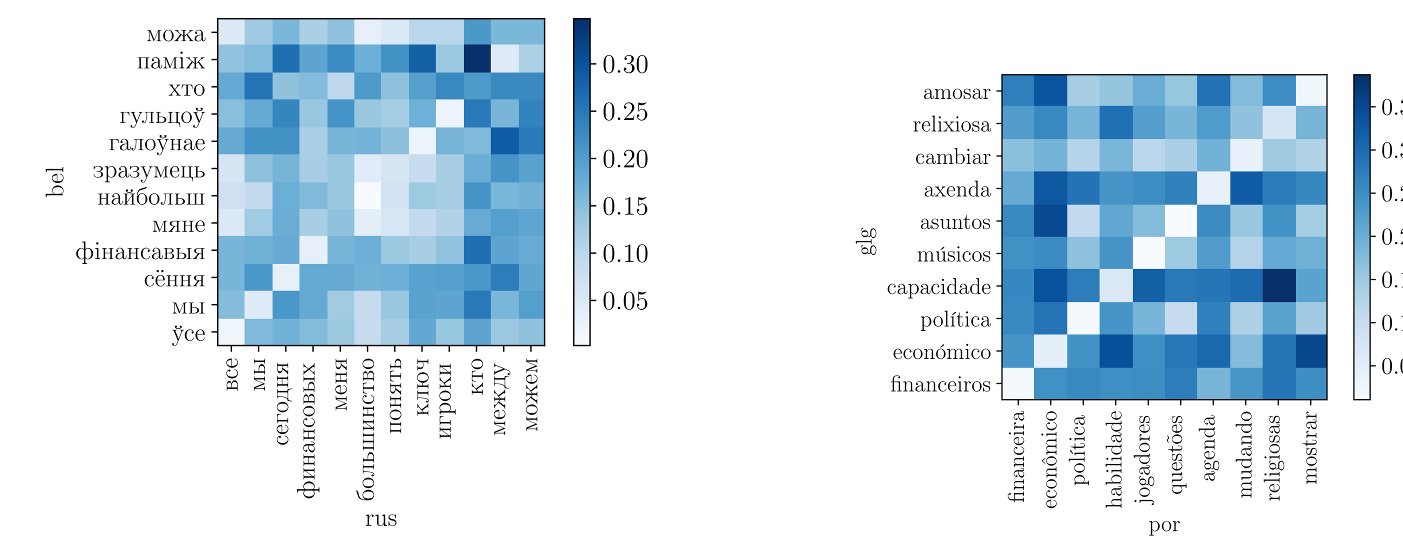
- **Available here:** `https://github.com/cindyxinyiwang/SDE`

Analysis

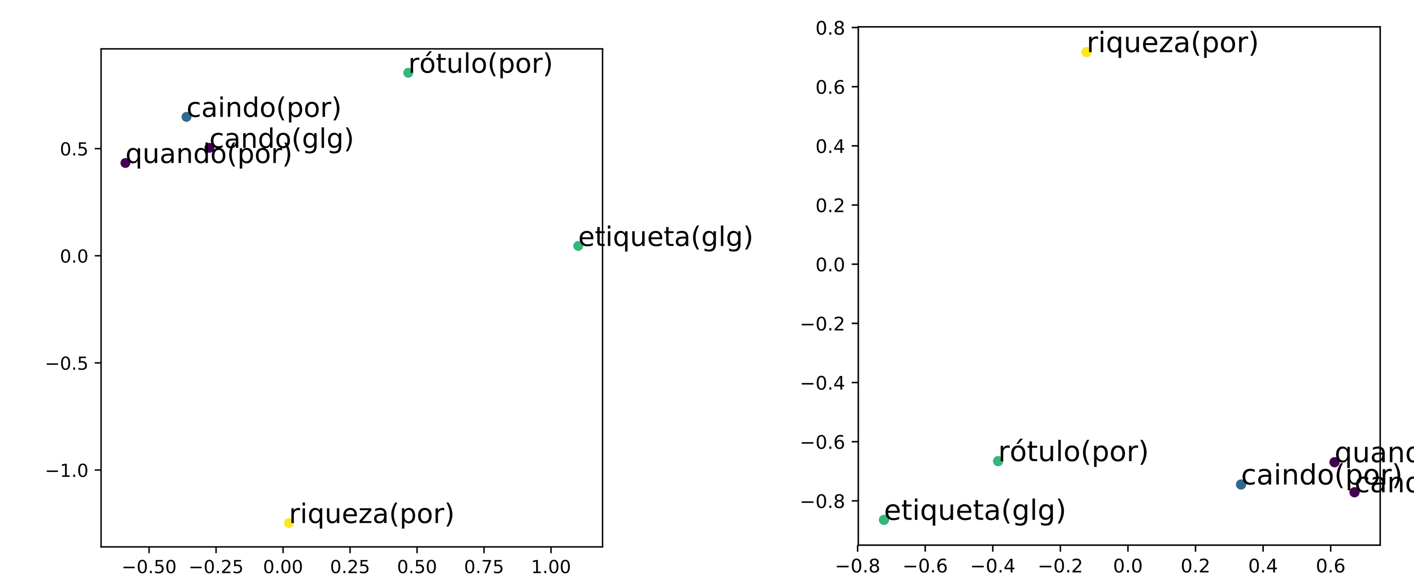
- **Why does SDE work?**



  - More gains for words split into more BPE pieces (left)
  - More gains for words with similar related language words (right)
- **What is the latent semantic embedding like?**



  - Words of similar meaning attend to similar latent space
- **Visualization of the word embedding from SDE**



  - Words of similar meaning are moved closer
- **Some translation examples!**

glg	eng	sub-sep	SDE
Pero non temos a tecnoloxía para resolver iso, temos?	But we don't have a technology to solve that, right?	But we don't have the technology to solve that , we have?	But we don't have the technology to solve that, do we?
Se queres saber sobre o clima, preguntas a un climatólogo.	If you want to know about climate, you ask a climatologist.	If you want to know about climate, you're asking a college friend.	If you want to know about climate, they ask for a weather.
Non é dicir que si tivesemos todo o diñeiro do mundo, non o quereríamos facer.	It's not to say that if we had all the money in the world, we wouldn't want to do it .	It's not to say that we had all the money in the world, we didn't want to do it .	It's not to say that if we had all the money in the world, we wouldn't want to do it.