Monolingual Embeddings for Low Resourced Neural Machine Translation

M. A. Di Gangi[1,2] and M. Federico[2]

{digangi,federico}@fbk.eu

[1]Università degli Studi di Trento
ICT International Doctoral School
Trento, Italy

[2]Fondazione Bruno Kessler
via Sommarive, 18, Trento

Problem

- NMT consists of millions of parameters and requires large parallel corpora to be trained.
- Unfortunately, most language pairs are low-resourced
- In other NLP tasks, embeddings trained on large monolingual corpora have been successfully used to compensate the lack of labeled data
- Can we use a similar approach also in NMT?

Solution

1. Train **external embeddings** on monolingual data
2. Extend input layer for **additional input**
3. Input at time t: word index and external embedding
4. **Merge the two embeddings**
5. Pass on result to RNN layer

Embedding Merging

Mix sum: $\hat{\mathbf{x}}_j = \mathbf{x}_j + \tilde{\mathbf{x}}_j$

Mix ctrl: $\hat{\mathbf{x}}_j = \mathbf{x}_j + w_{ext}\tilde{\mathbf{x}}_j$
$w_{ext} = \sigma(\tilde{\mathbf{x}}^\top \mathbf{w_{ctrl}} + b_{ctrl})$

Mix Gate: $\hat{\mathbf{x}}_j = \tanh(\mathbf{z}_j \odot ff_1(\mathbf{x}_j) + (1 - \mathbf{z}_j) \odot ff_2(\tilde{\mathbf{x}}_j))$
$\mathbf{z}_j = \sigma([\mathbf{x}_j; \tilde{\mathbf{x}}_j]^\top \mathbf{W_z} + \mathbf{b_z})$

$\mathbf{x}_j$ = NMT embedding of j-th word

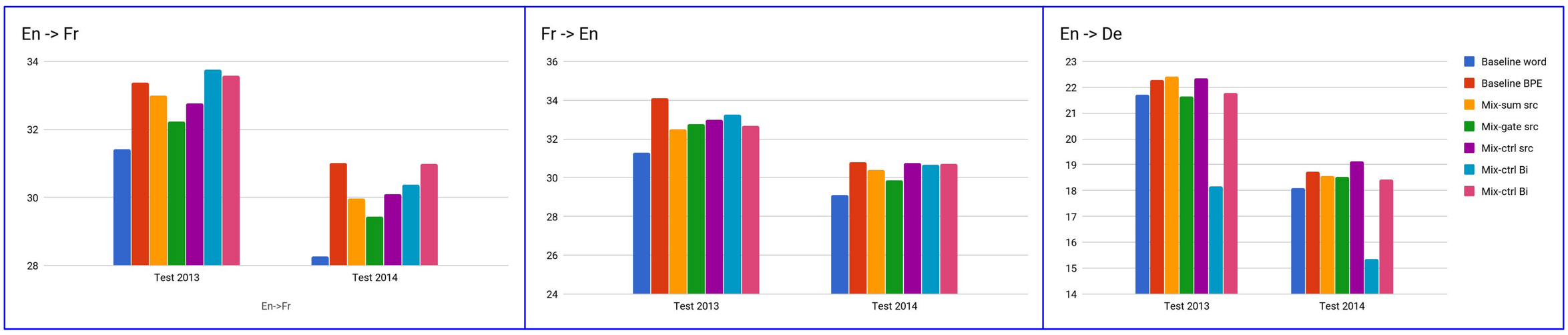
$\tilde{\mathbf{x}}_j$ = external embedding of j-th word

Experimental Setup

	EN <->FR		EN -> DE	
Parallel Words	~4.3M	~4.5M	~3.9M	~3.7M
Monolingual Words	~42B	~2.5B	~42B	~5B
Mono. Emb. Size	300	300	300	300
Mono. Emb. Vocab.	~1.9M	~900K	~1.9M	~4.7M

Embeddings size	500
RNN size	500
Dropout	0.2
Optimizer	Adam
Learning rate	0.0003

Results



Sample translations

Src	but this isn't what twentysomethings are [...]
Ref	mais ce n'est pas ce que les jeunes adultes [...]
Word	mais ce n'est pas ce que les jeunes de la vingtaine
Mix Ctrl	mais ce n'est pas ce que les jeunes de la vingtaine [...]
BPE	mais ce n'est pas ce que les gens de twitymer [...]
Src	Egyptologists have always known the site of Itjtawy was [...]
Ref	les égyptologues avaient toujours présumé qu'Itjtawy se [...]
Word	Nous avons toujours connu le site de Londres , situé [...]
Mix Ctrl	les UNK ont toujours connu le site de la UNK était [...]
BPE	les Egyptologistes ont toujours connu le site de Itjtawy a été

Conclusions

- Leverage external embeddings for NMT
  - Independently trained/pre-existing
- Effective merging with internal embeddings
  - With few additional parameters (ctrl)
- Improved word representations
  - for low-resourced settings.
- Competitive with NMT with BPE
  - less NE errors, more UNKs

QT21
ModernMT
Next Generation Machine Translation