VaLaR (Vastly Lacking Resources) NMT

Minhyung Kang, Kais Kudrolli
{dankang, kudrolli} @ stanford.edu

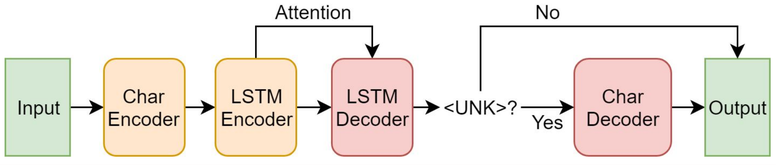
Background

- **Problem**: Translate Quenya, an Elvish language created by author J.R.R. Tolkien, to English
- **Low-resource setting**: As Elvish is a fictional language, we have very limited, mostly fan-sourced data
  - **Existing approaches**: **transfer learning**, universal representation, meta learning, **data augmentation**

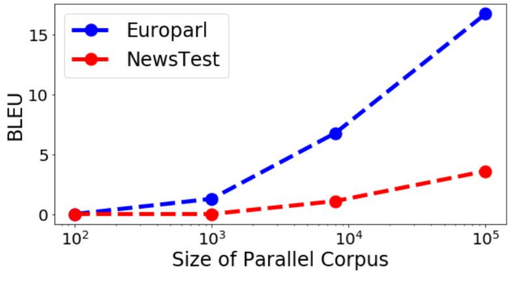
Data

Data set	Pair	Train	Val	Test	Usage
Bible	Elvish English	7,135	396	396	Main training
Misc		-	-	215	Test
Dictionary		5,107	-	-	Augmentation
Europarl	Finnish English	1M	1,000	3,000	Transfer, LM
Newstest		-	-	3,000	Scarcity

Model



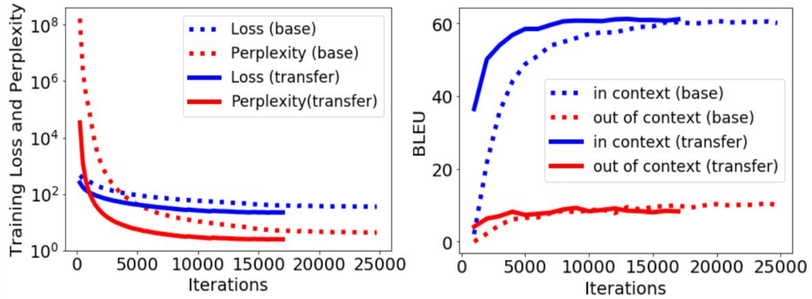
Data Scarcity and Performance



Amount of data is **highly correlated** with performance

Transfer Learning

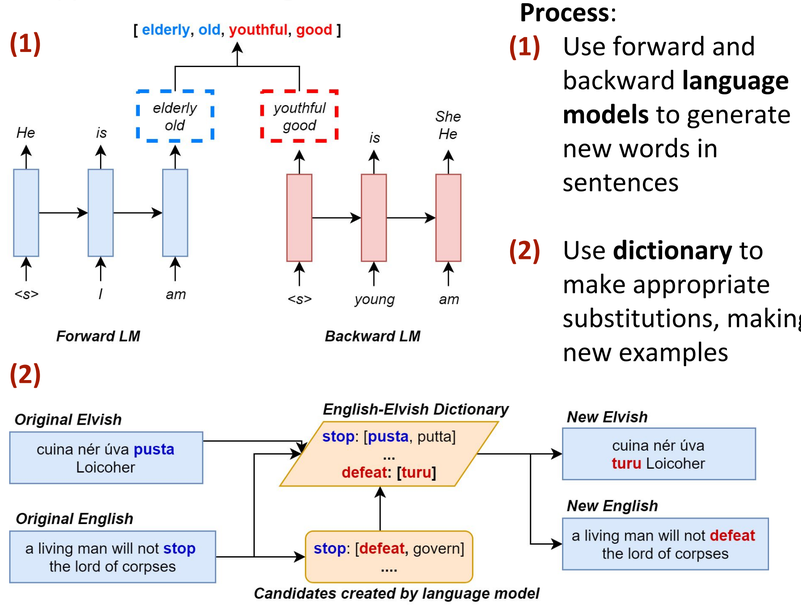
- **Idea**: create a parent NMT model between Finnish-English, train Elvish-English child model
- **Faster convergence**, not necessarily better performance



- **Ablation study**: freeze parts of network and observe performance
  - Best model: freeze **target char embedding** and **char-decoder**

Data Augmentation

- **Idea**: generate new examples that use vocabulary which do not appear in our training set



Process:

(1) Use forward and backward **language models** to generate new words in sentences

(2) Use **dictionary** to make appropriate substitutions, making new examples

Results

Model	Train Loss	Train PPL	Bible BLEU	Misc BLEU
Baseline	**35.71**	**4.33**	60.11	9.89
Transfer	44.18	6.14	**61.35**	10.41
Augmented	39.65	11.68	61.08	**11.38**

Translation Example

Elvish: Mernelye fire ar harya alcar
Gold Translation: You wanted to die and have glory
Our Translation: You wanted to die and possess glory

Augmentation Example

Original: For you do not **know** what day your Lord will come
Augment: For you do not **grasp** what day your Lord will come
Original: An ualde **ista** mi mana ré Herulda tuluva
Augment: An ualde **mapa** mi mana ré Herulda tuluva

Conclusion & Challenges

- Lack of data makes training and **evaluation** difficult
- It is difficult to **balance augmentation and repetition**
- Transfer learning does help by providing a good **initial model**
- **Quality over quantity**: better augmented examples more helpful than large number of examples

Future Work

- Different ways to combine forward and backward language models to generate new candidates
- **Named entity recognition** to prevent mistranslation
- Better parsing (normalize case, stemming)

References

[1] Zoph, B, Barret Zoph, Deniz Yuret, Jonathan May, Kevin Knight. *Transfer Learning for Low-Resource Neural Machine Translation.* In EMNLP 2016.
[2] Marzieh Fadaee, Arianna Bisazza, Christof Monz. *Data Augmentation for Low-Resource Neural Machine Translation*. In ACL 2017.