Data Augmentation for Low-Resource Neural Machine Translation

Marzieh Fadaee **Arianna Bisazza** **Christof Monz**
Informatics Institute, University of Amsterdam

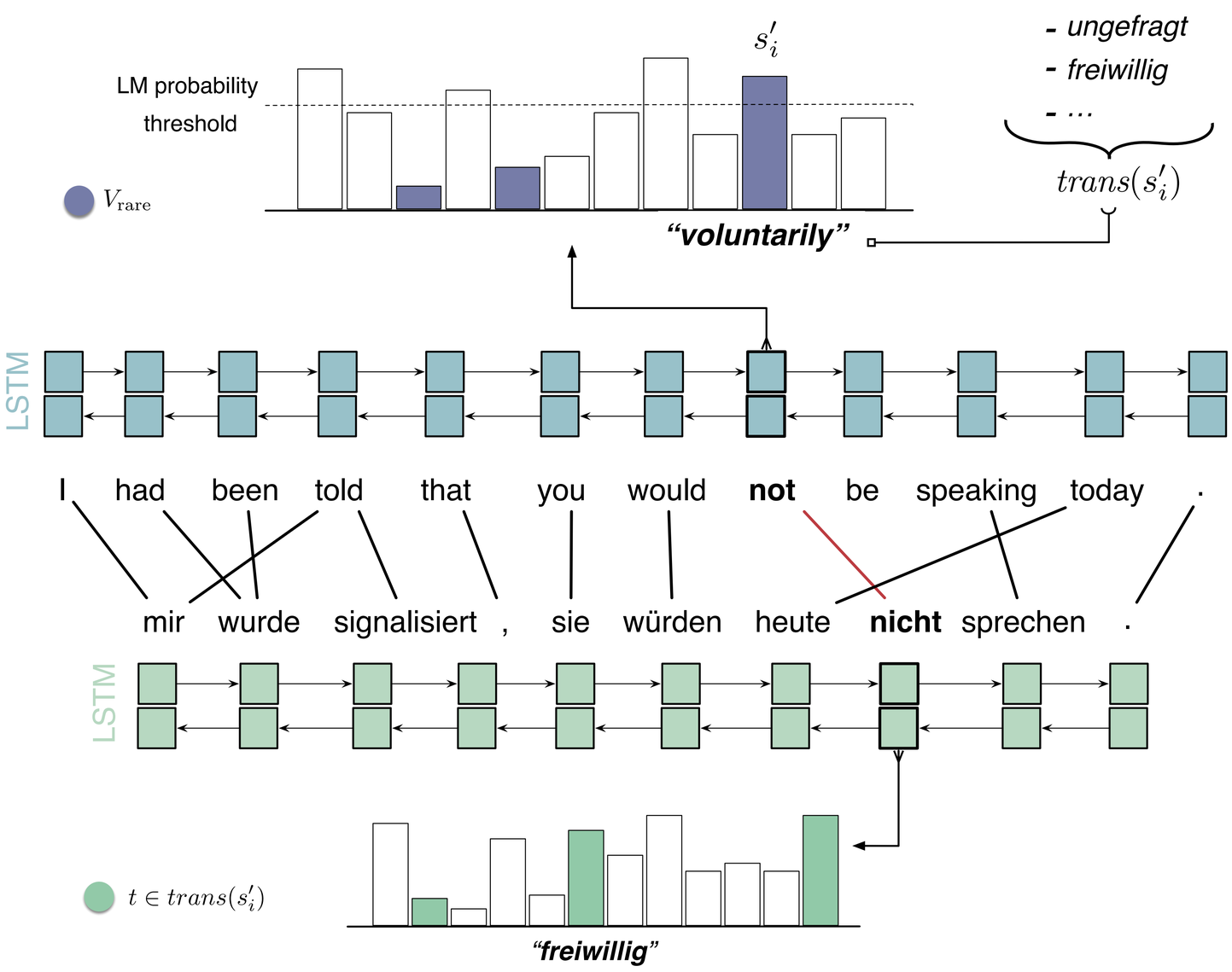
Summary

- Neural Machine Translation models perform best when an abundance of parallel data is available
- Acquiring human translations for low-resource language pairs is costly
- Hence translation of low-frequency words is difficult and often inaccurate
- Our approach
  - alters existing parallel sentences **targeting low-frequency words**
  - augments the data by generating new diverse context for low-frequency words and the corresponding translations
- As a result training with the augmented bitext achieves significant BLEU improvements in a simulated low-resource English↔German translation setting

Data Augmentation

- Image Processing
  - Flipping, cropping, tilting, altering the RGB channels
- Has not been done in Natural Language Processing
- One possible approach is paraphrasing which is meaning-preserving
- Our approach focuses on non meaning-preserving augmentation
- Closest work is back-translation of monolingual data (Sennrich et al. ACL 2016)

Approach: Translation Data Augmentation (TDA)



Choosing the best translation of s'_i ...

... that is compatible with the context

$$t'_j = \underset{t \in trans(s'_i)}{\arg\max}\ P_{lex}(t|s'_i) P_{lex_{inv}}(s'_i|t) P_{LM_T}(t|t_1^{j-1})$$

New sentence pair:

*I had been told that you would **voluntarily** be speaking today.*

*mir wurde signalisiert, sie würden heute **freiwillig** sprechen.*

NMT Results (BLEU)

DE→EN

Model	Data	testset2014	testset2015	testset2016
Baseline	371K	10.6	11.3	13.1
Back-trans$_{1:1}$	731K	11.4 (+0.8)$^{\blacktriangle}$	12.2 (+0.9)$^{\blacktriangle}$	14.6 (+1.5)$^{\blacktriangle}$
TDA$_{r=1}$	4.5M	11.9 (+1.3)$^{\blacktriangle,-}$	13.4 (+2.1)$^{\blacktriangle,\blacktriangle}$	15.2 (+2.1)$^{\blacktriangle,\blacktriangle}$
TDA$_{r\geq 1}$	6M	**12.6** (+2.0)$^{\blacktriangle,\blacktriangle}$	**13.7** (+2.4)$^{\blacktriangle,\blacktriangle}$	**15.4** (+2.3)$^{\blacktriangle,\blacktriangle}$

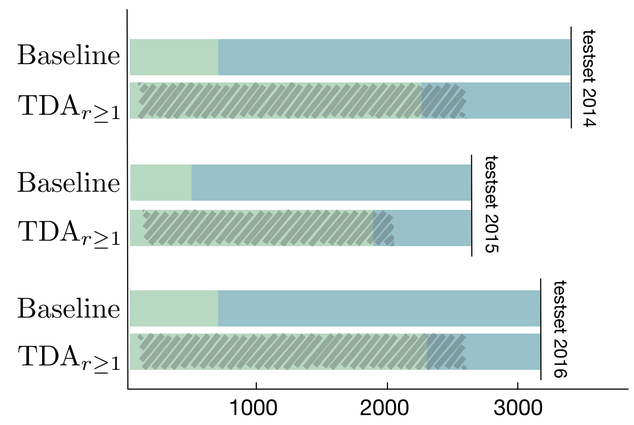
EN→DE

Model	Data	testset2014	testset2015	testset2016
Baseline	371K	8.2	9.2	11.0
Back-trans$_{1:1}$	731K	9.0 (+0.8)$^{\blacktriangle}$	10.4 (+1.2)$^{\blacktriangle}$	12.0 (+1.0)$^{\blacktriangle}$
TDA$_{r=1}$	4.5M	10.4 (+2.2)$^{\blacktriangle,\blacktriangle}$	11.2 (+2.0)$^{\blacktriangle,\blacktriangle}$	13.5 (+2.5)$^{\blacktriangle,\blacktriangle}$
TDA$_{r\geq 1}$	6M	**10.7** (+2.5)$^{\blacktriangle,\blacktriangle}$	**11.5** (+2.3)$^{\blacktriangle,\blacktriangle}$	**13.9** (+2.9)$^{\blacktriangle,\blacktriangle}$

→ Altering only one word per sentence
→ Altering one or multiple words per sentence

- Simulated low-resource MT setting
- TDA significantly improves translation quality
- Substituting several rare words is preferable even though the augmented sentences are likely to be noisier

Rare Translation Generation (DE→EN)



- Words in $V_{\text{rare}} \cap V_{\text{reference}}$ generated during translation
- Words in $V_{\text{rare}} \cap V_{\text{reference}}$ not generated during translation
- Words in $V_{\text{rare}} \cap V_{\text{reference}}$ affected by augmentation

Conclusions

- We present a data augmentation technique to enrich the training data targeting rare words
  - Increasing the size of the training data by diversifying the context of rare words yields better translations
  - Generation of *correct* rare words during translation increases
  - The attention scores of rare words are on average 8.8% higher than the baseline model
  - The generated translation length to reference length ratio is on average 7% higher

UNIVERSITY OF AMSTERDAM

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers 639.022.213 and 639.021.646

NWO

m.fadaee@uva.nl