Engaging Content
Engaging People

Neural Machine Translation with and without parallel data.

Dimitar Shterionov, Alberto Poncelas, Andy Way

This work has been supported by the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund, by the European Commission as part of the FALCON project (contract number 610879)

DCU

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union

European Union
European Regional Development Fund

SFI

Problem

- Problem:
  - Neural Machine Translation (NMT) relies heavily on the amount and quality of the available parallel data.
  - In many scenarios the data is not enough to train a good NMT system.

- Solutions:
  - Synthetic data, e.g., backtranslation.
  - Pivot and Zero-shot Translation (ZST) systems.
  - …

- Research questions:
  - If no parallel data is available, can we still do MT?
  - How can we employ ‘foreign’ data for MT?
  - Is more data = higher quality?

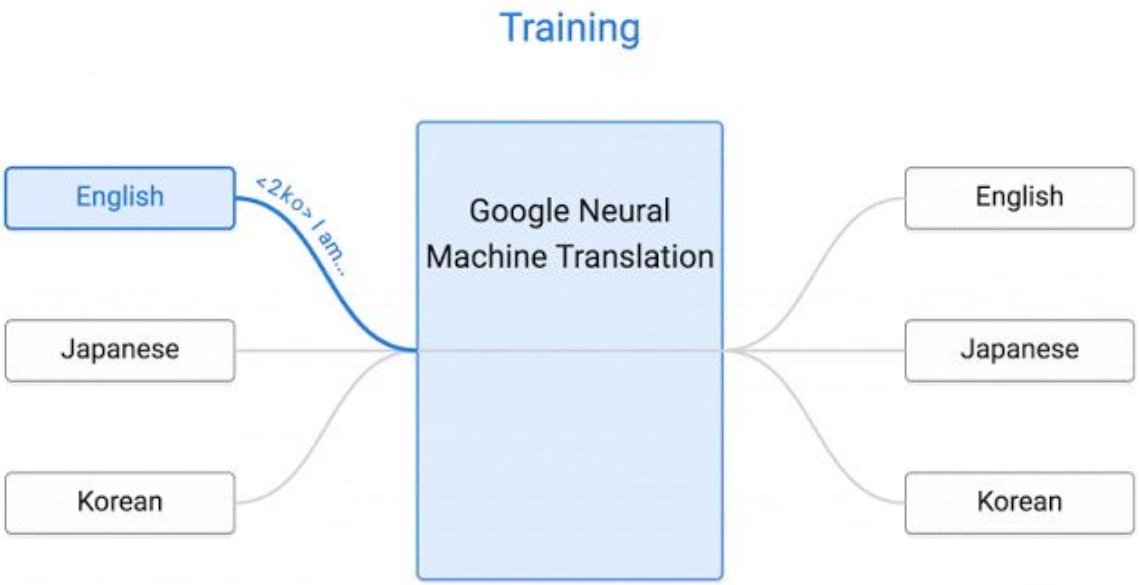
Presentation Outline

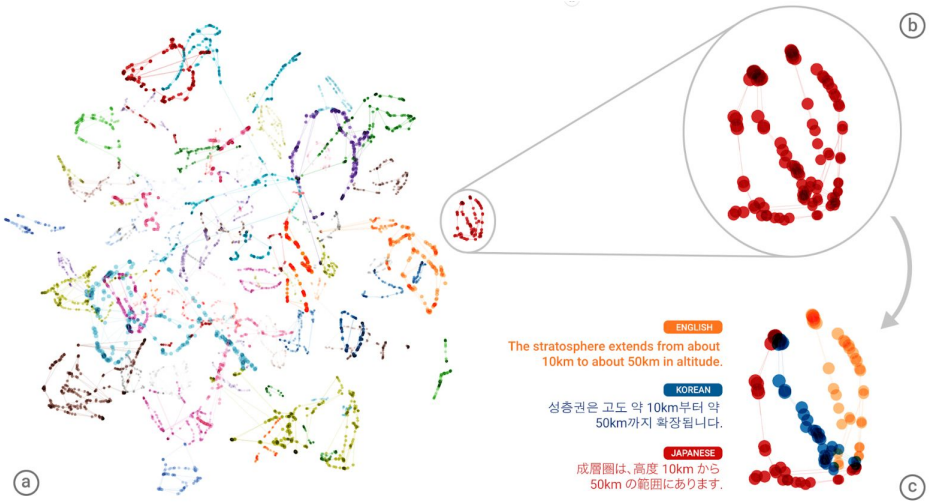
- Related Work
- Our work
- Data
- Experiments
- Discussion

Presentation Outline

- **Related Work**
- **Our work**
- **Data**
- **Experiments**
- **Discussion**

Related work





[https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html]

Related work

- Johnson et al. 2016:
  - a single shared attention mechanism and a single ‘universal’ encoder-decoder across all languages is used.
  - one prefixed token to indicate the translation direction.
- Ha et al. 2016:
  - a single shared attention mechanism and a single ‘universal’ encoder-decoder across all languages is used.
  - a language code to differentiate words from different languages.
  - a prefix and postfix on the source side of the training and validation data.
- Firat et al. (2016):
  - a shared attention mechanism multiple encoders/decoders for each source and target language.
  - investigates multiple strategies for multi-way, multilingual MT engines.
  - also a more basic multilingual NMT engine - trained on two parallel corpora (with or without a fine-tuning corpus).
- Mattoni et al. 2017:
  - first commercial custom zero-shot system.
  - two tokens - one to specify the source language (for tokenisation and segmentation); another to specify the target language.

Related work

- *Ha, T., Niehues, J., and Waibel, A. H. (2016). Toward multilingual neural machine translation with universal encoder and decoder. In Proceedings of the Thirteenth International Workshop on Spoken Language Translation (IWSLT '16), Seattle, WA, USA.*
- *Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Yonghui Chen, Z., and Thorat, N. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation.*
- *Firat, O., Sankaran, B., Al-Onaizan, Y., Yarman-Vural, F. T., and Cho, K. (2016). Zero resource translation with multi-lingual neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pages 268–277.*
- *Mattoni, Giulia, Pat Nagle, Carlos Collantes, and Dimitar Shterionov. (2017) "Zero-Shot Translation for Indian Languages with Sparse Data.", user track, MT Summit XVI.*
- *Lakew, Surafel Melaku & Lotito, Quintino & Turchi, Marco & Negri, Matteo & Federico, Marcello. (2017). FBK's Multilingual Neural Machine Translation System for IWSLT 2017.*

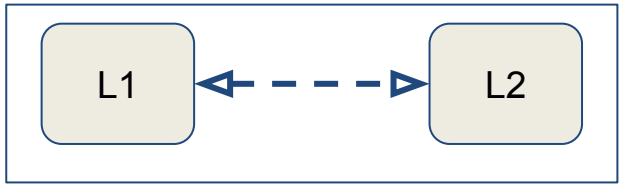
Presentation Outline

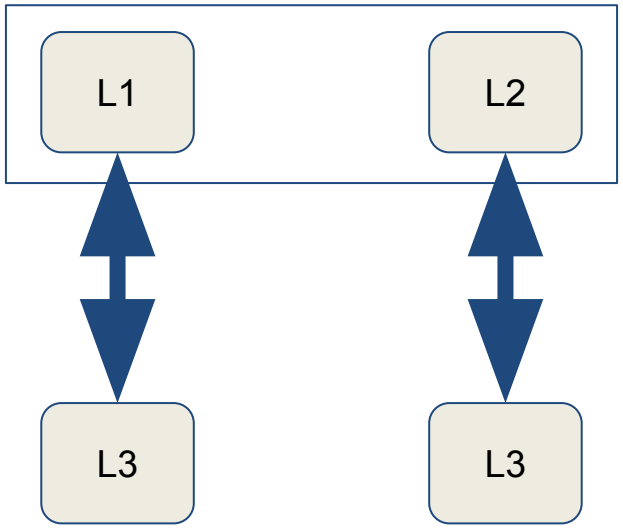
- Related Work
- Our work
- Data
- Experiments
- Discussion

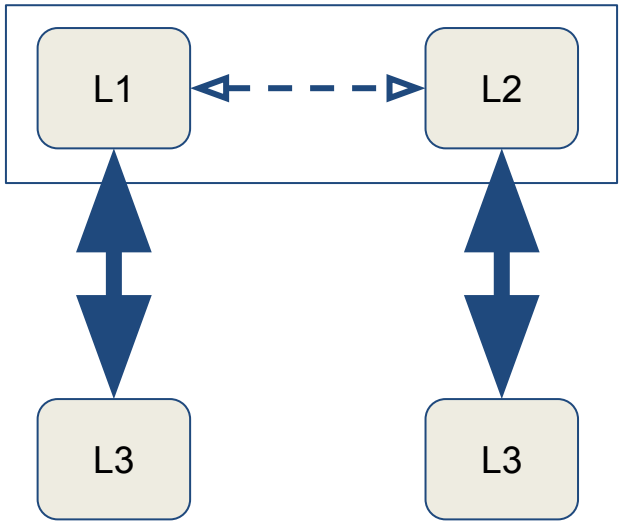
Presentation Outline

- Related Work
- Our work
- Data
- Experiments
- Discussion

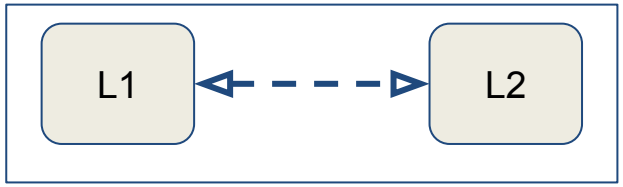
Our work

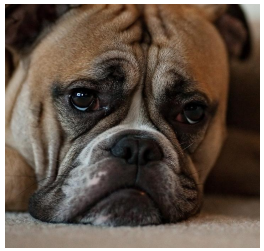


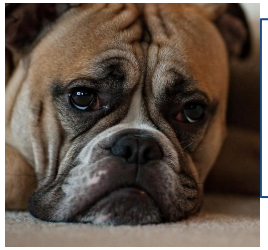


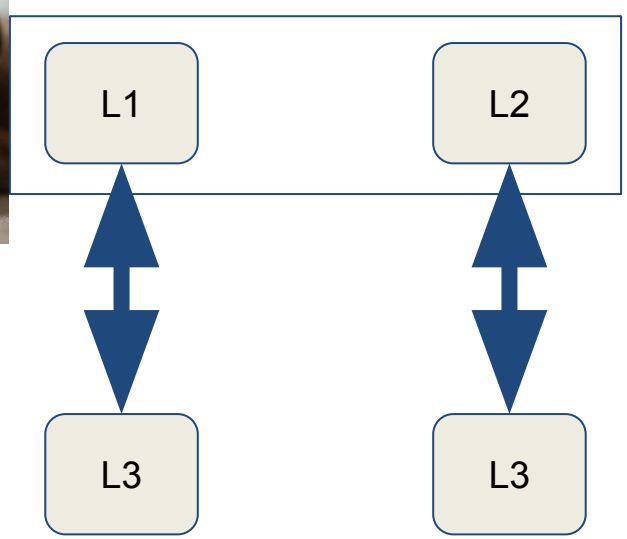


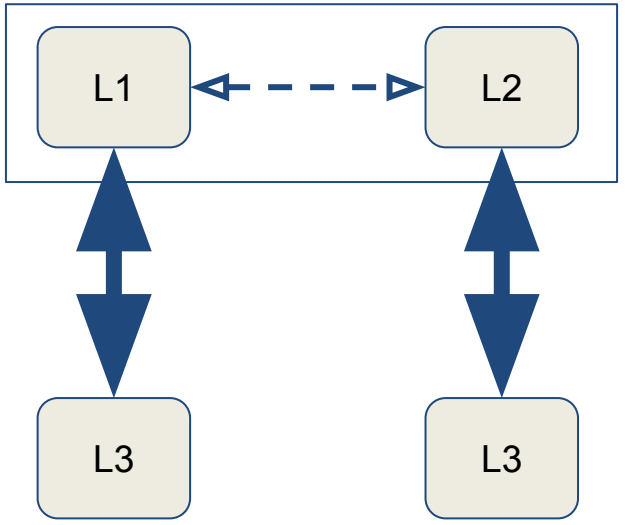
Our work



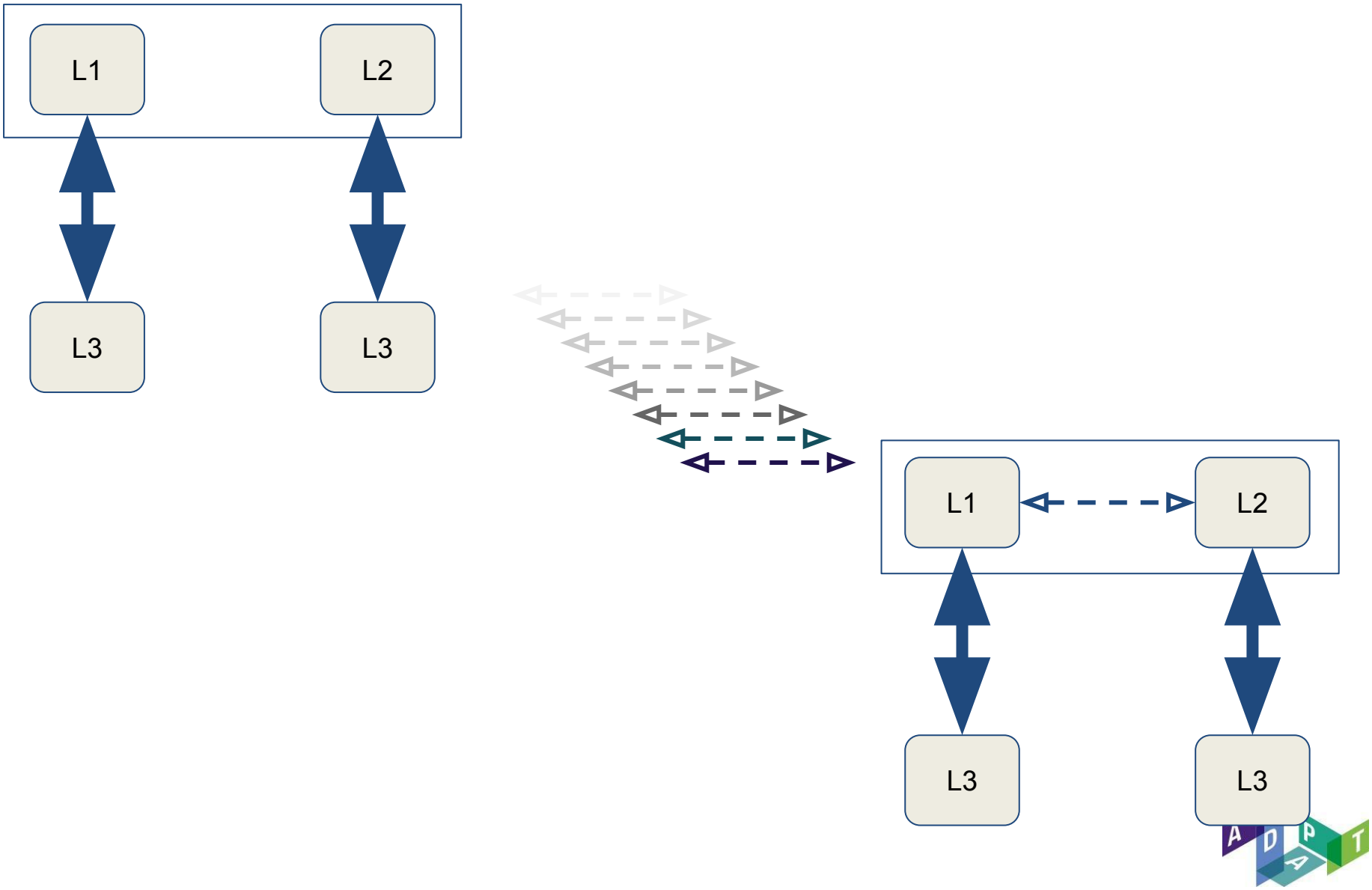




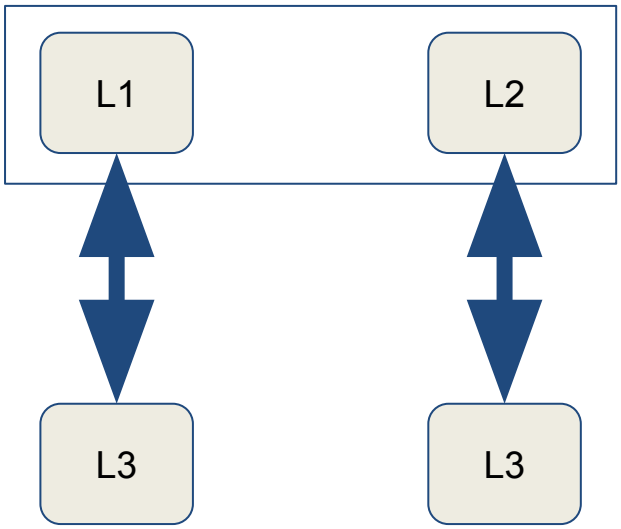




Our work

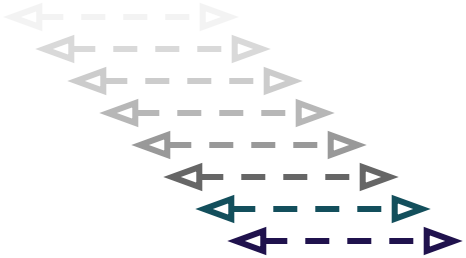


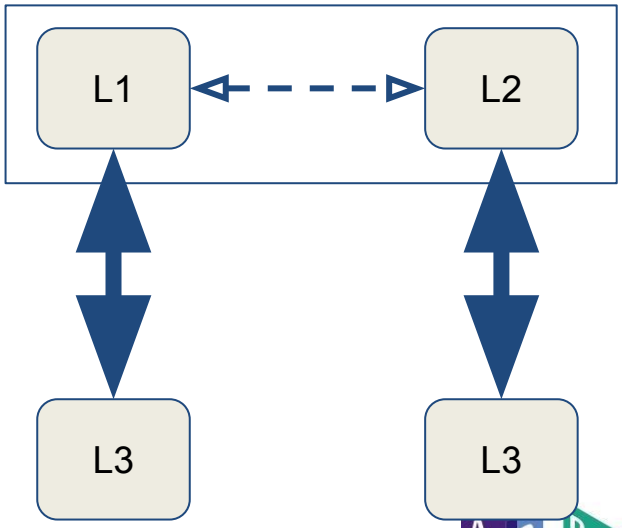
Our work



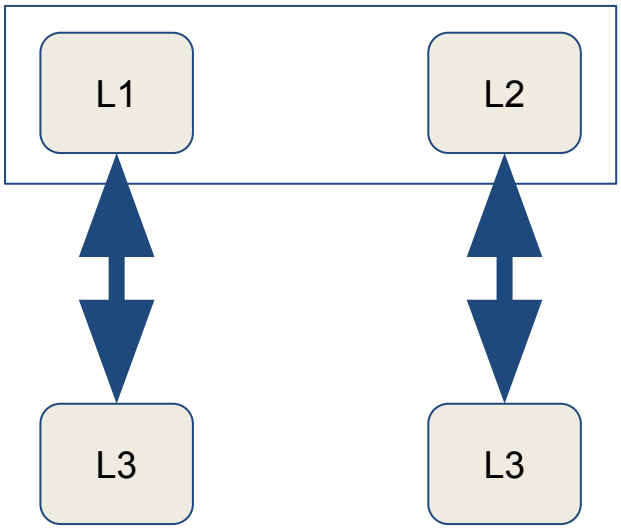
Three language pairs/triplets:

- CA - ES - EU
- HU - EN - RU
- AM - EN - TI



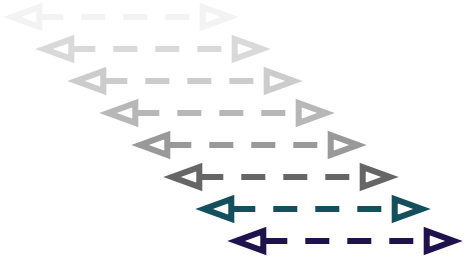


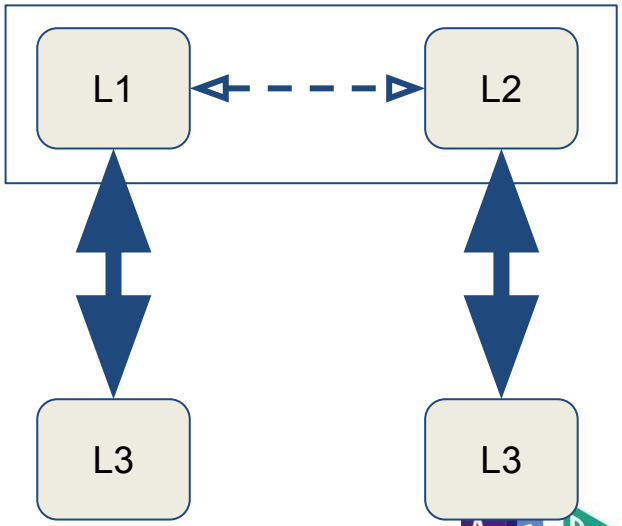
Our work



Three language pairs/triplets:

- CA - ES - EU
- HU - EN - RU
- AM - EN - TI





Two NMT architectures:

- seq2seq (LSTM)
- Transformer

Presentation Outline

- Related Work
- Our work
- **Data**
- **Experiments**
- **Discussion**

Presentation Outline

- Related Work
- Our work
- Data
- Experiments
- Discussion

Data

- Catalan - Basque (Ca-Eu) [KDE, Gnome, Ubuntu, http://opus.nlpl.eu/]
- Catalan - Spanish (Ca-Es) [DOGC dataset (journal of the Catalan Goverment: http://opus.nlpl.eu/DOGC.php)]
- Basque - Spanish (Eu-Es)
  [Basque-Spanish parallel and monolingual data from the Open Data Euskadi http://hltshare.fbk.eu/IWSLT2018/OpendataBasqueSpanish.tgz]

  CA-EU: 100K, CA-ES: +/- 900K, EU-ES: +/- 1M

- Hungarian - Russian (Hu-Ru) [Books (http://opus.nlpl.eu/Books-v1.php)]
- Hungarian - English (Hu-En) [DGT, http://opus.nlpl.eu/DGT-v4.php]
- Russian - English (Ru-En) [MultiUN, http://opus.nlpl.eu/MultiUN-v1.php]

  HU-RU: 23K, HU-EN: +/- 950K, RU-EN: +/- 950K

- Amharic - Tigrigna (Am-Tg) [Bible, thanks to Yalemisew Abgaz, DCU]
- Amharic - English (Am-En) [Various domains, https://github.com/adtsegaye/Amharic-English-Machine-Translation-Corpus]
- Tigrigna - English (Tg-En) [Bible, thanks to Yalemisew Abgaz, DCU]

  AM-TI: 11K, AM-EN: +/- 66K, TI-EN: 11K

Data

- Sentences prefix with target language ID token
- Byte Pair Encoding has been applied using 89500 merge operation
- Tokenized and truecased.
- Sentences have been shuffled and split into train, dev and test
- Bidirectional

Examples:

<2es> persones i col · lectius afectats o obligats a subministrar les dades : públic comprador d ' abon@@ aments .

<2eu> text a cer@@ car@@ Fin@@ d and go to the n@@ ext search mat@@ ch

Presentation Outline

- Related Work
- Our work
- Data
- **Experiments**
- **Discussion**

Presentation Outline

- Related Work
- Our work
- Data
- **Experiments**
- **Discussion**

Setup

- OpenNMT-py
- Vocabulary size of max 50000 for each language

- LSTM:
  - # units 500
  - trained for 13 epochs
  - batch_size: 64
  - SGD as optimization method
  - learning rate 1
  - learning decay rate 0.5
  - decaying starts after epoch 8

- Transformer:
  - # layers: 6
  - rnn_size: 512
  - word_vec_size: 512
  - transformer_ff: 2048
  - # heads 8
  - training batch_size: 4096
  - Adam as learning optimizer

- Trained on 1 GPU/Tested on 1 GPU

Results CA-EU

- 200K available parallel data
- used up to 100K
- test set from the same domain as the parallel data

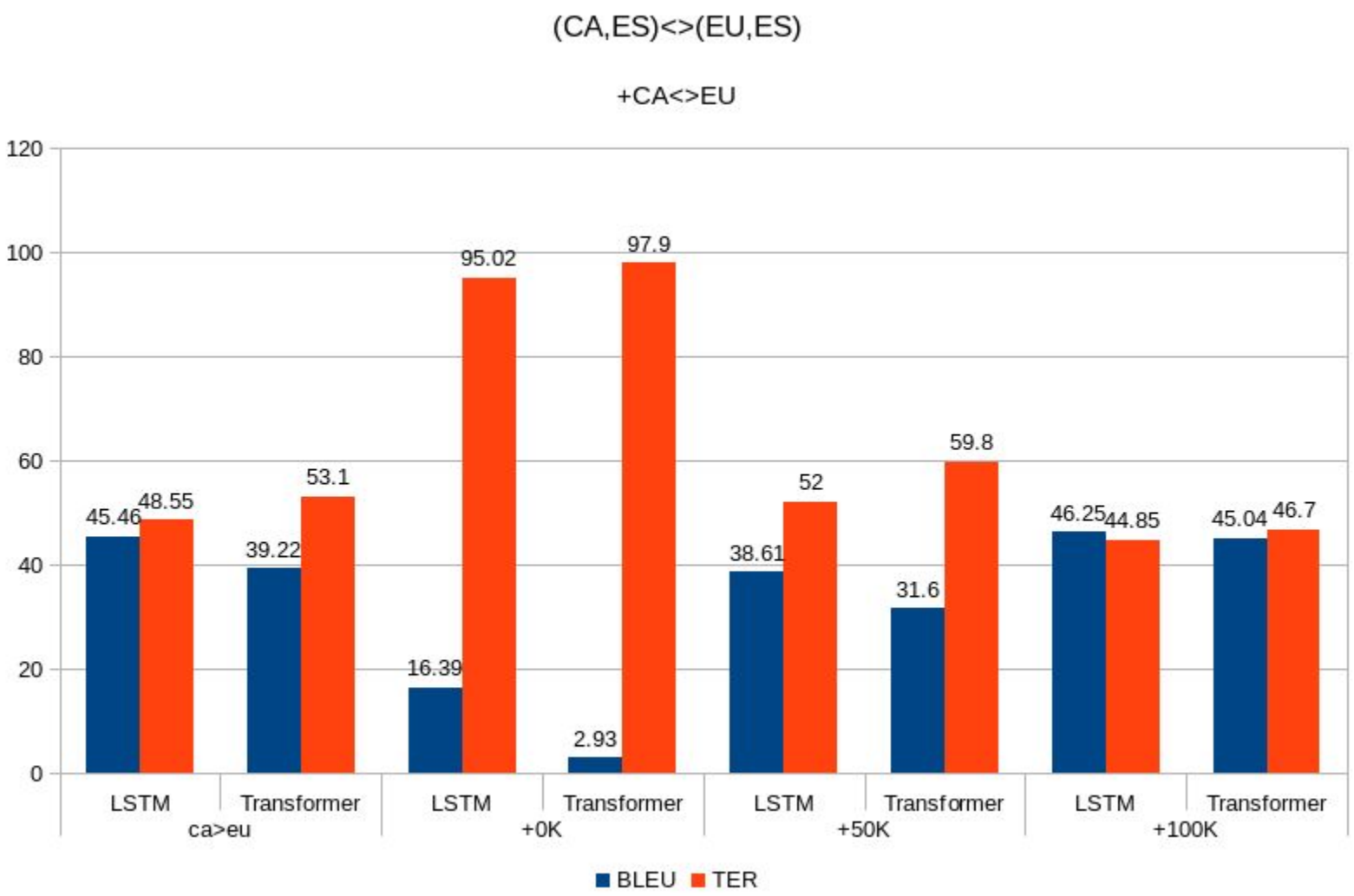
LSTM:

	ca>eu	(ca,es)<>(es,eu)+0	(ca,es)<>(es,eu)+50K	(ca,es)<>(es,eu)+80K	(ca,es)<>(es,eu)+100K
BLEU	45.46	16.39	38.61	44.24	46.25
TER	48.55	95.02	52	47.08	44.85
CHRF3	58.3945	30.8232	56.9894	61.6281	63.3835
CHRF1	63.5674	32.0881	61.0687	65.5365	67.316

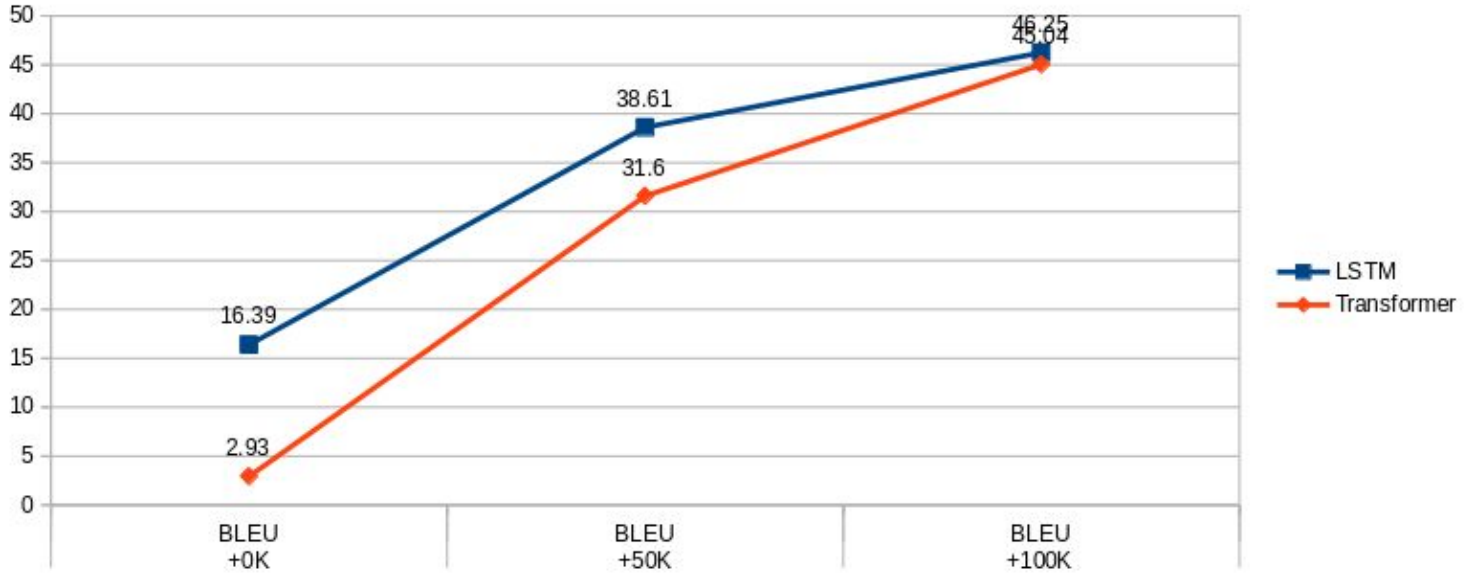
Transformer:

	ca>eu	(ca,es)<>(es,eu)+0	(ca,es)<>(es,eu)+50K	(ca,es)<>(es,eu)+80K	(ca,es)<>(es,eu)+100K
BLEU	39.22	2.93	31.6	/	45.04
TER	53.1	97.9	59.8	/	46.7

Results CA-EU



Results CA-EU



Results HU-RU

- 23K available (Books domain)
- used up to all available parallel data
- test set from the same domain as the parallel data

	hu>ru 23K	(hu,en)<>(en,ru)+0	(hu,en)<>(en,ru)+12K	(hu,en)<>(en,ru)+23K
BLEU	3	0.86	8.35	10.37
TER	95.67	97.84	80.31	74.79
CHRF3	30.3109	15.0262	37.5291	41.7277
CHRF1	33.0571	17.892	42.2992	46.7515

Results AM-TI

- 11K parallel data available (Bible)
- used up to all available parallel data
- aim was to create am>ti MT (a low resource scenario)

		Ballanced		Disballanced	
	am>ti	(am,en)<>(ti,en)	(am,en)<>(ti,en)+9K	(am,en)<>(ti,en)	(am,en)<>(ti,en)+9K
BLEU	0.57	0.67	2.03	0.01	8.01
TER	95.68	91.14	92.62	99.34	82.93
CHRF3	18.4193	24.0842	31.1435	3.7339	42.2927
CHRF1	24.2746	29.5177	34.5188	6.1146	44.9757

Presentation Outline

- Related Work
- Our work
- Data
- Experiments
- **Discussion**

Presentation Outline

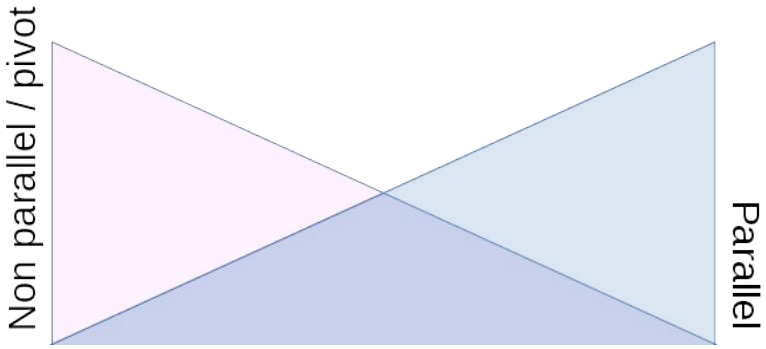
- Related Work
- Our work
- Data
- Experiments
- **Discussion**

Discussion

- No parallel is not a good solution.
- The amount of parallel data is essential.
- Combining parallel and
  non-parallel (or pivot) data

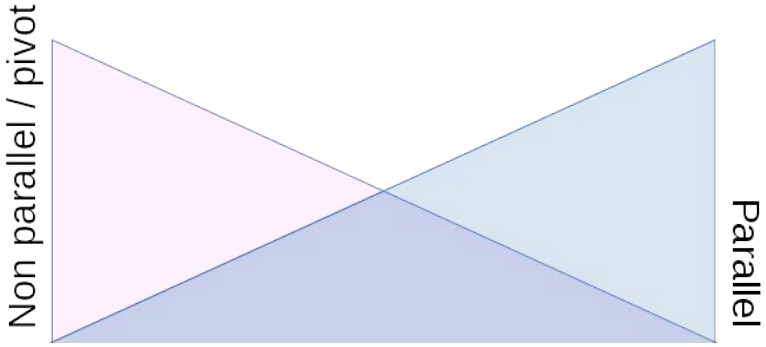
Discussion

- No parallel is not a good solution.
- The amount of parallel data is essential.
- Combining parallel and non-parallel (or pivot) data



Discussion

- No parallel is not a good solution.
- The amount of parallel data is essential.
- Combining parallel and non-parallel (or pivot) data



- Using parallel and pivot data for low resource scenarios is the best way to go.
- Transformer is more impacted than LSTM
- The quality differences are global, expressed by all metrics.

Presentation Outline

- Related Work
- Our work
- Data
- Experiments
- Discussion

Future work, future-current work