Mitigating the Disparity for Machine Translation Quality for Low Resource Languages

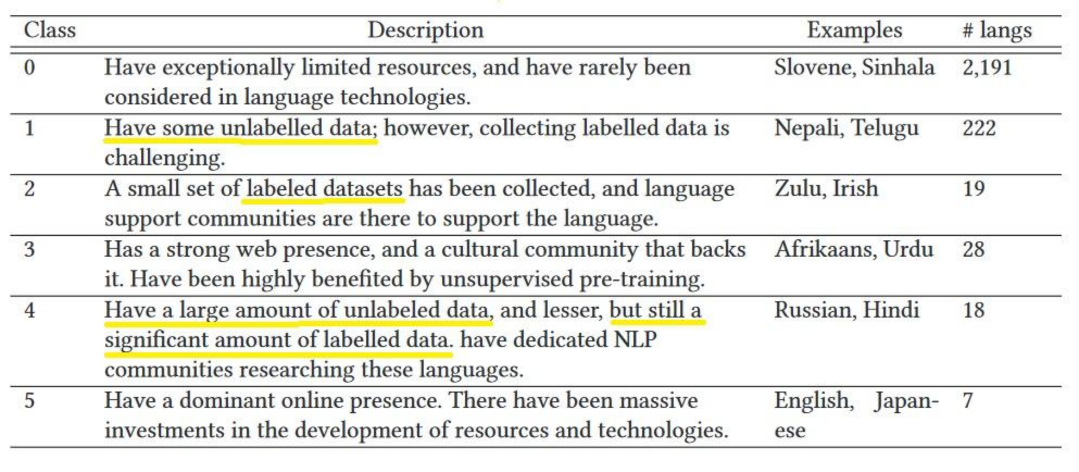
By Jeffrey Miller

Who fares better with the same technology?

What is the issue?
Why is this an issue?
How is this an issue?

Low Resource and High Resource Languages

- Monolingual and Parallel Data
- Corpora



Class	Description	Examples	# langs
0	Have exceptionally limited resources, and have rarely been considered in language technologies.	Slovene, Sinhala	2,191
1	Have some unlabelled data; however, collecting labelled data is challenging.	Nepali, Telugu	222
2	A small set of labeled datasets has been collected, and language support communities are there to support the language.	Zulu, Irish	19
3	Has a strong web presence, and a cultural community that backs it. Have been highly benefited by unsupervised pre-training.	Afrikaans, Urdu	28
4	Have a large amount of unlabeled data, and lesser, but still a significant amount of labelled data. have dedicated NLP communities researching these languages.	Russian, Hindi	18
5	Have a dominant online presence. There have been massive investments in the development of resources and technologies.	English, Japanese	7

Figure from: [4]

Machine Learning

- Machine Translation (MT)
- Statistical Machine Translation (SMT)
- Neural Machine Translation (NMT)

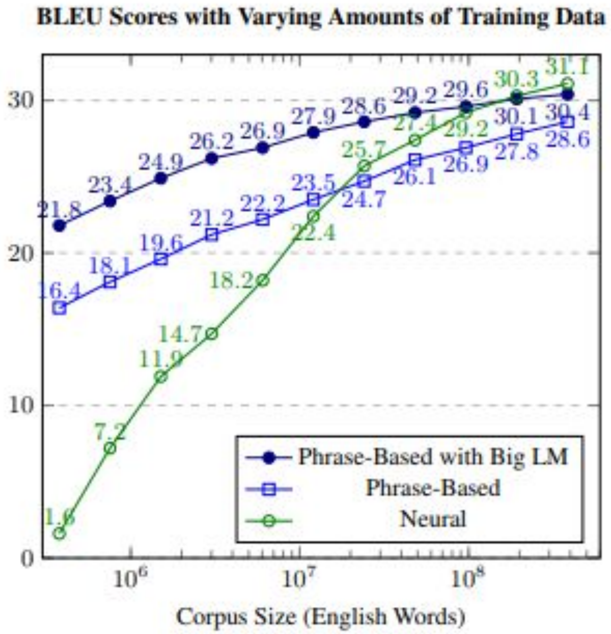
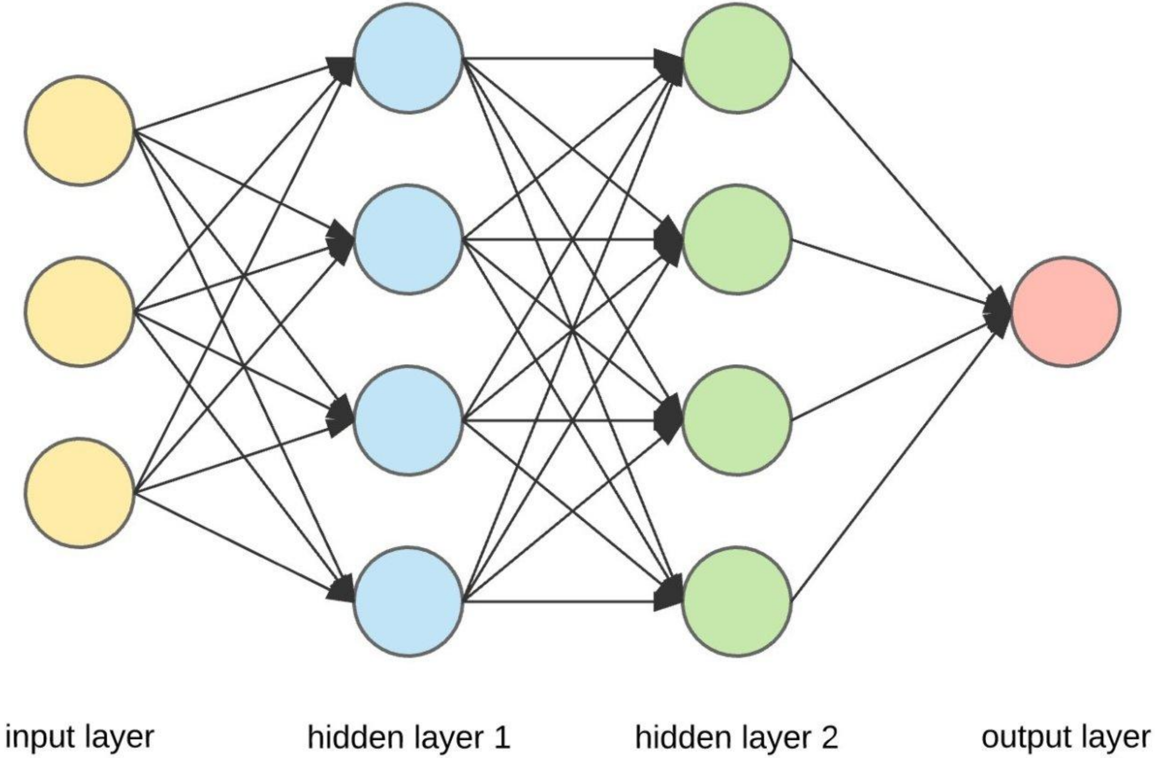


Figure from: [2]

Talking Points:

- Artificial Neural Networks (ANNs) and Encoder-Decoder
- LRL techniques
  - Data Augmentation
  - Transfer Learning
- Neural Translation Machine (NMT)
  - Semi Supervised
  - Unsupervised
- Application & Results

ANNs



Encoder Structure

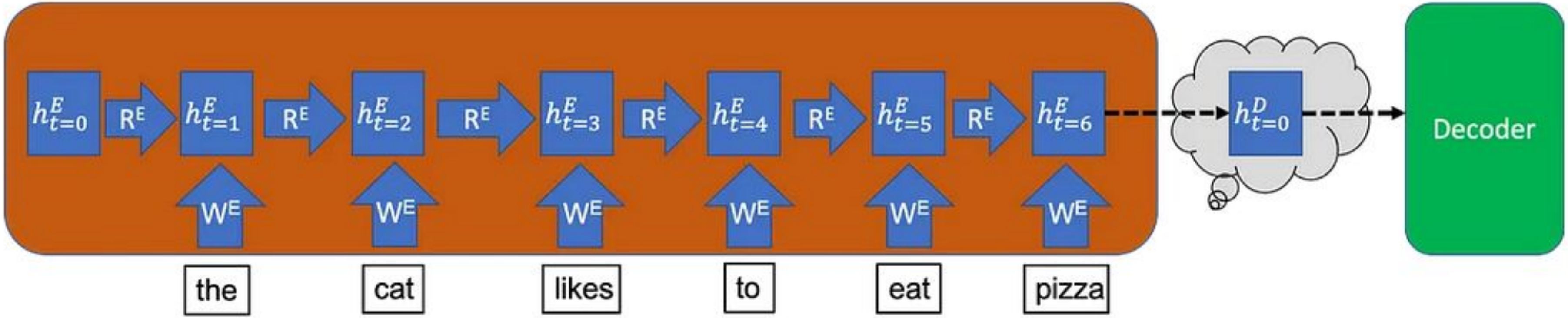


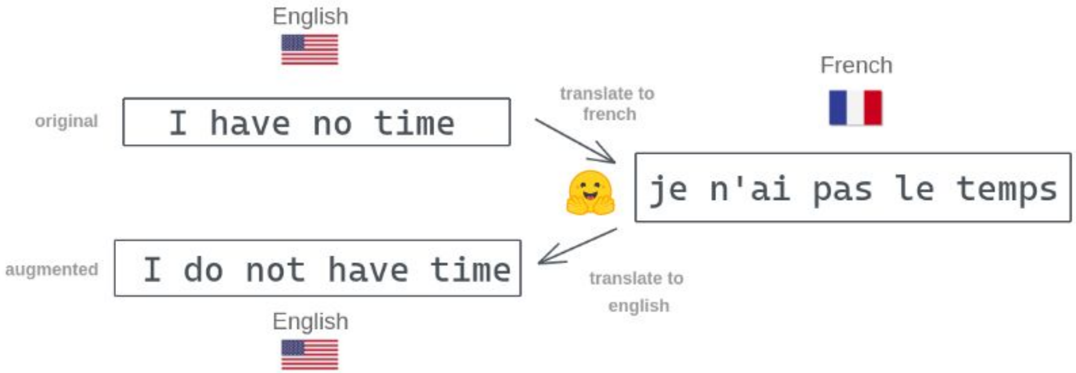
Figure from: [3]

Low Resource Techniques

Data Augmentation

Data Augmentation:

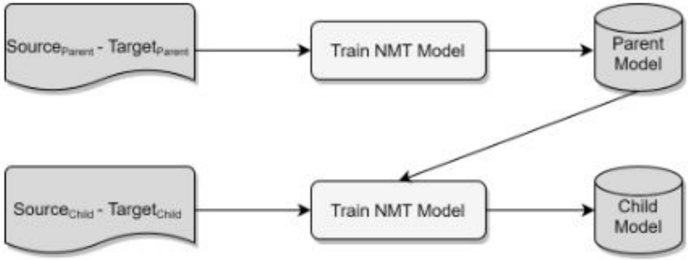
- Parallel Corpus Mining
- Back Translation
- Word/Phrase Replacement



Transfer Learning

Transfer Learning

- “Transfering” the parameters of a high-resource pair to a low resource pair
- Transfer Learning for Multi-NMT
- Transfer Protocol
  - “Freezing”

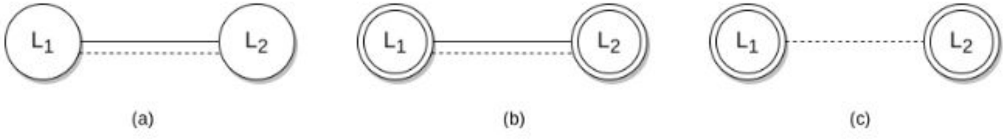


Source$_{Parent}$ - Target$_{Parent}$
Train NMT Model
Parent Model
Source$_{Child}$ - Target$_{Child}$
Train NMT Model
Child Model

Figure from: [4]

NMT Architectures

- Semi Supervised
  - Language Model
  - Multi task learning
- Unsupervised
  - Initialization
  - Recurrent Translation



(a): Supervised; (b): Semi Supervised; (c): Unsupervised

Figure from: [4]

Unsupervised

- Initialization
    - Word Embeddings
- Translation and Auto-encoding

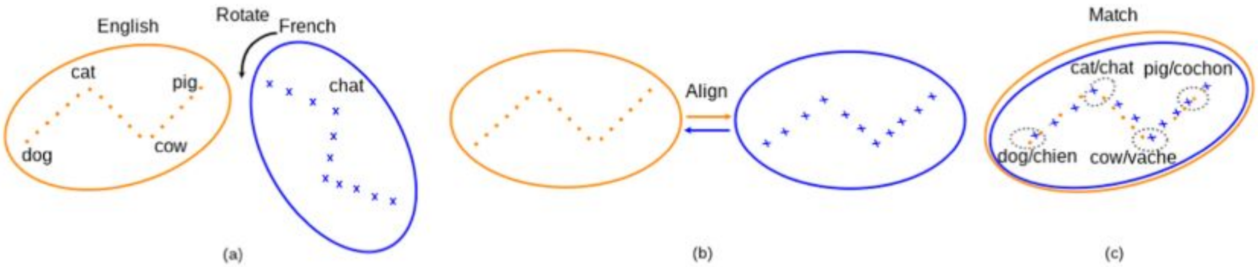
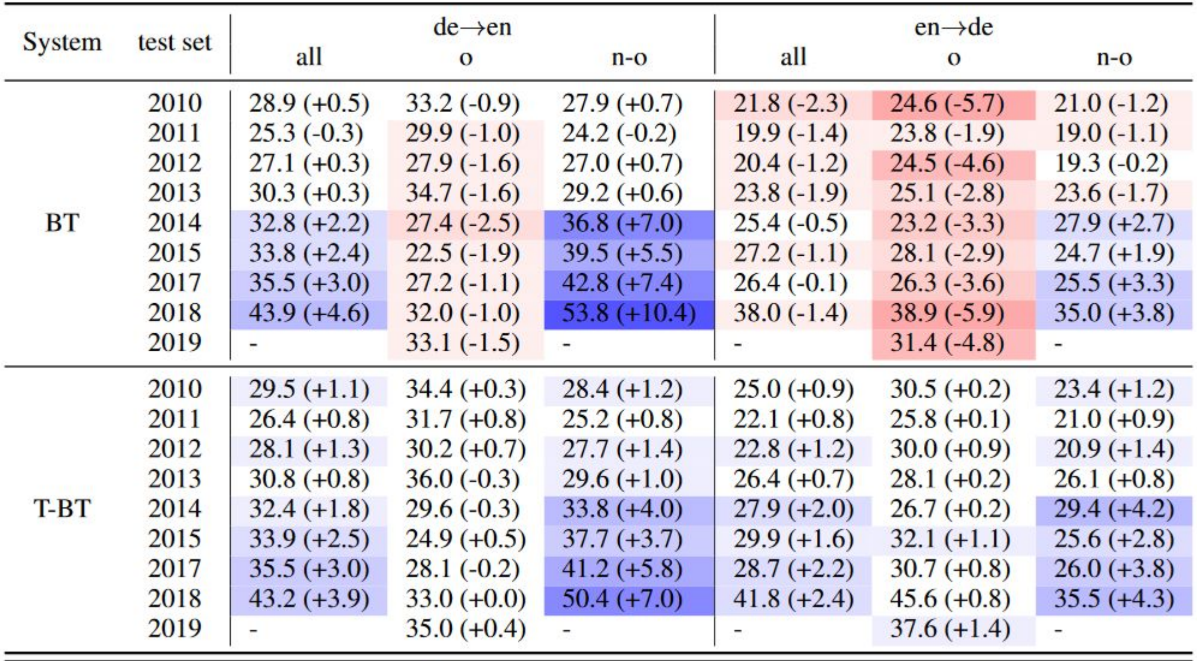


Figure from: [4]

Application and Results

- Back-Translation: Tagged and Untagged
- WMT9 German-English Corpus



System	test set	de→en			en→de		
		all	o	n-o	all	o	n-o
BT	2010	28.9 (+0.5)	33.2 (-0.9)	27.9 (+0.7)	21.8 (-2.3)	24.6 (-5.7)	21.0 (-1.2)
	2011	25.3 (-0.3)	29.9 (-1.0)	24.2 (-0.2)	19.9 (-1.4)	23.8 (-1.9)	19.0 (-1.1)
	2012	27.1 (+0.3)	27.9 (-1.6)	27.0 (+0.7)	20.4 (-1.2)	24.5 (-4.6)	19.3 (-0.2)
	2013	30.3 (+0.3)	34.7 (-1.6)	29.2 (+0.6)	23.8 (-1.9)	25.1 (-2.8)	23.6 (-1.7)
	2014	32.8 (+2.2)	27.4 (-2.5)	36.8 (+7.0)	25.4 (-0.5)	23.2 (-3.3)	27.9 (+2.7)
	2015	33.8 (+2.4)	22.5 (-1.9)	39.5 (+5.5)	27.2 (-1.1)	28.1 (-2.9)	24.7 (+1.9)
	2017	35.5 (+3.0)	27.2 (-1.1)	42.8 (+7.4)	26.4 (-0.1)	26.3 (-3.6)	25.5 (+3.3)
	2018	43.9 (+4.6)	32.0 (-1.0)	53.8 (+10.4)	38.0 (-1.4)	38.9 (-5.9)	35.0 (+3.8)
	2019	-	33.1 (-1.5)	-	-	31.4 (-4.8)	-
T-BT	2010	29.5 (+1.1)	34.4 (+0.3)	28.4 (+1.2)	25.0 (+0.9)	30.5 (+0.2)	23.4 (+1.2)
	2011	26.4 (+0.8)	31.7 (+0.8)	25.2 (+0.8)	22.1 (+0.8)	25.8 (+0.1)	21.0 (+0.9)
	2012	28.1 (+1.3)	30.2 (+0.7)	27.7 (+1.4)	22.8 (+1.2)	30.0 (+0.9)	20.9 (+1.4)
	2013	30.8 (+0.8)	36.0 (-0.3)	29.6 (+1.0)	26.4 (+0.7)	28.1 (+0.2)	26.1 (+0.8)
	2014	32.4 (+1.8)	29.6 (-0.3)	33.8 (+4.0)	27.9 (+2.0)	26.7 (+0.2)	29.4 (+4.2)
	2015	33.9 (+2.5)	24.9 (+0.5)	37.7 (+3.7)	29.9 (+1.6)	32.1 (+1.1)	25.6 (+2.8)
	2017	35.5 (+3.0)	28.1 (-0.2)	41.2 (+5.8)	28.7 (+2.2)	30.7 (+0.8)	26.0 (+3.8)
	2018	43.2 (+3.9)	33.0 (+0.0)	50.4 (+7.0)	41.8 (+2.4)	45.6 (+0.8)	35.5 (+4.3)
	2019	-	35.0 (+0.4)	-	-	37.6 (+1.4)	-

Figure from: [1]

References

[1]Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged Back-translation Revisited: Why Does It Really Work?. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 5990–5997

[2][Six Challenges for Neural Machine Translation] (Koehn & Knowles, NGT 2017)

[3]Quinn Lanners. 2019. Neural Machine Translation.

[4]Surangika Ranathunga. 2022. Neural Machine Translation for Low Resource languages: A Survey. Comput. Surveys 55 (2022), 1–37. Issue 1.

Questions?