Natural Language Understanding, Generation, and Machine Translation

Lecture 20: Low-resource Machine Translation

Alexandra Birch

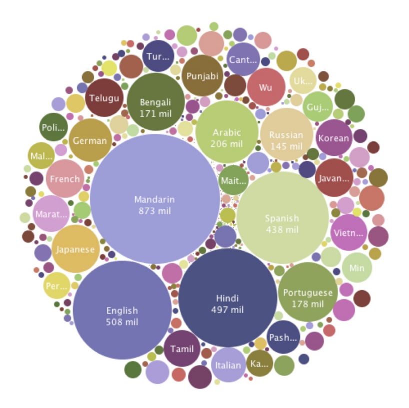
10 March 2022 (week 7)

School of Informatics
University of Edinburgh
a.birch@ed.ac.uk

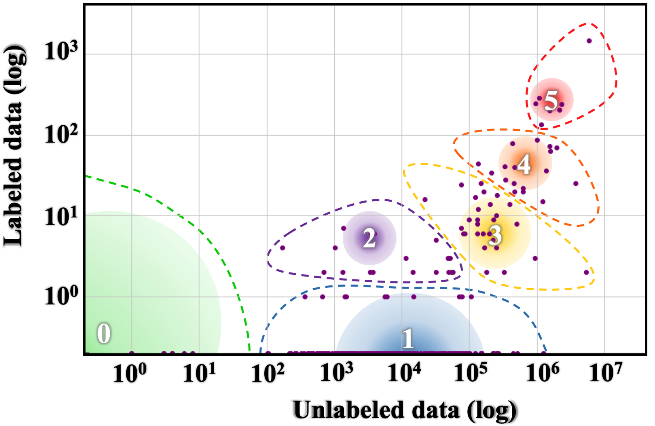
with content from Barry Haddow

Low-Resource MT

Diversity of Languages



What is low-resource?



The State and Fate of Linguistic Diversity and Inclusion in the NLP World [Joshi et al., 2020]

What is low-resource?

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

The State and Fate of Linguistic Diversity and Inclusion in the NLP World [Joshi et al., 2020]

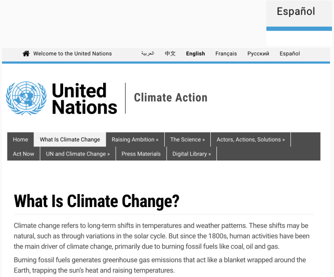
What is low-resource?

“Low-resourced”-ness is a complex problem going beyond data availability and reflects systemic problems in society.

Masakhane [Nekoto et al., 2020]

Corpus Creation

Web Crawling



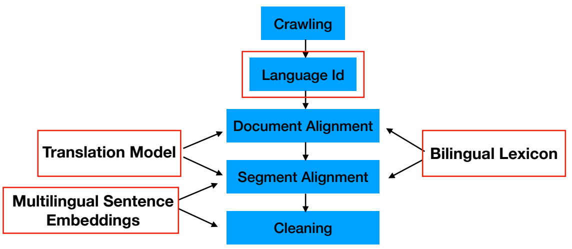
- Extract text from websites identified as multilingual
- Align documents then sentences
- Collate, deduplicate and filter

Extraction from Monolingual Data

Large collections of monolingual data contain parallel sentences: Common Crawl, Internet Archive

- How to detect these?
  - Map sentences into a common embedding space using eg. LASER
  - Nearest neighbours to find parallel sentences
- Paracrawl, WikiMatrix, CCMatrix, Samanantar

Problems with large scale extraction



- Tools for low-resource languages are poor
- False positives can dominate
- What if there is not much text at all?

Quality of Crawled Data

	Parallel		
	CCAligned	**ParaCrawl v7.1**	**WikiMatrix**
#languages	137	41	85
Source	CC 2013–2020	selected websites	Wikipedia
Filtering level	document	sentence	sentence
Langid	FastText	CLD2	FastText
Alignment	LASER	Vec/Hun/BLEU-Align	LASER
Evaluation	TED-6	WMT-5	TED-45

Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets [Kreutzer et al., 2021]

Quality of Crawled Data

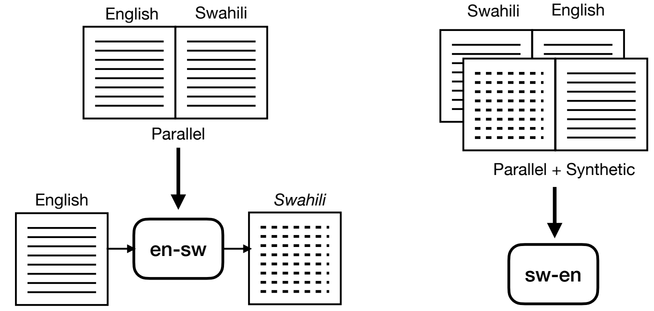
		Parallel		
		CCAligned	ParaCrawl v7.1	WikiMatrix
#langs audited / total		65 / 119	21 / 38	20 / 78
%langs audited		54.62%	55.26%	25.64%
#sents audited / total		8037 / 907M	2214 / 521M	1997 / 95M
%sents audited		0.00089%	0.00043%	0.00211%
macro	C	29.25%	76.14%	23.74%
	X	29.46%	19.17%	68.18%
	WL	9.44%	3.43%	6.08%
	NL	31.42%	1.13%	1.60%
	offensive	0.01%	0.00%	0.00%
	porn	5.30%	0.63%	0.00%

Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets [Kreutzer et al., 2021]

C - correct, X - incorrect, WL - wrong language, NL - Not a language

Using Monolingual Data

Synthetic Parallel Data

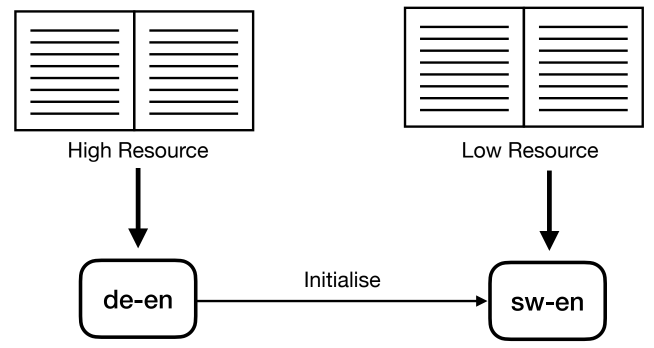


Improving Neural Machine Translation Models with Monolingual Data [Sennrich et al., 2016]

- Back translation still most popular and effective method
- Iterated back translation: 2-3 iterations sufficient
- Can fail if the initial system is too weak

Using Multilingual Data

Transfer Learning Using Parallel Data



- Initial work showed this working for Turkic languages

  [Zoph et al., 2016]
- Parent and Child do not need to be related [Kocmi and Bojar, 2018]
- Extensive investigation of choice of parents [Lin et al., 2019]
  - Data set size and lexical overlap important

Transfer learning from Many Monolingual Corpora

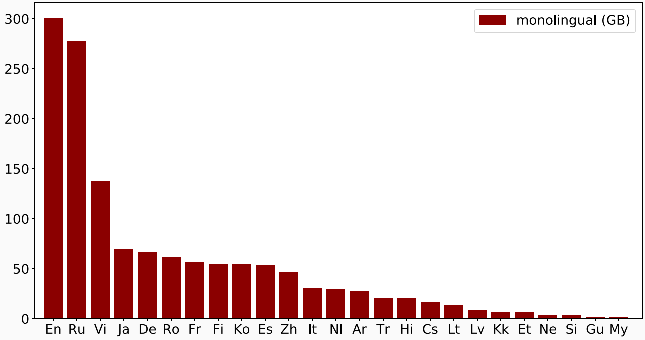
Early 2020: Large pretrained models had little influence of machine translation - why?

- MT is a very highly-resourced task for the most-studied language pairs
- MT models are encoder-decoders while most pretrained models at the time consists of only an encoder
- These models are very large and their computation time during inference can be prohibitive

This all changed with mBART

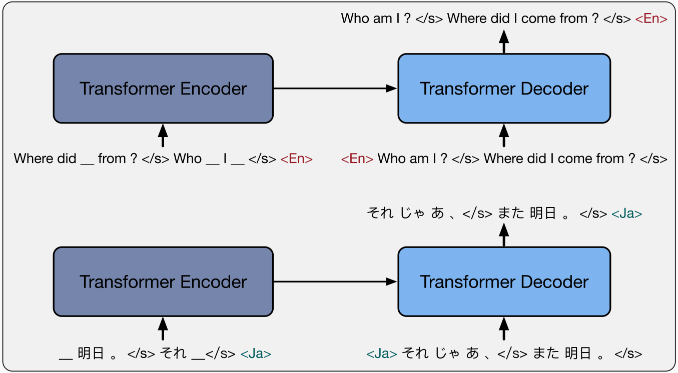
mBART

- Multilingual Denoising Pre-Training for NMT (mBART) [Liu et al., 2020]
- Pre-train massive monolinugal corpus: 25 languages from Common Crawl
- Then fine-tune parallel data separately for each translation direction



mBART

Pretrain on multiple monolingual data

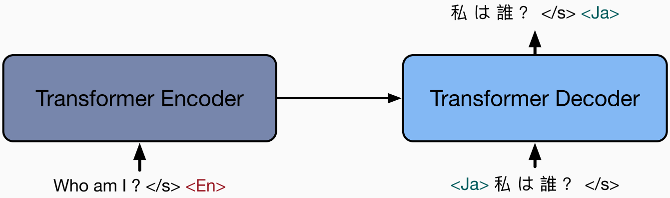


Multilingual Denoising Pre-Training (mBART)

from [Liu et al., 2020]

mBART

Fine-tune on parallel data



from [Liu et al., 2020]

mBART

- Encoder-Decoder architecture
- Objective: loss over full text reconstruction (not just over masked spans)
- Two kinds of noise:
    - mask spans of text: 35% of words
    - permute the order of sentences
- Language token for both source and target language
- Massive computational cost: trained for 2.5 weeks on 256 Nvidia V100 GPUs

mBART

Languages	En-Gu		En-Kk		En-Vi		En-Tr		En-Ja		En-Ko	
Data Source	**WMT19**		**WMT19**		**IWSLT15**		**WMT17**		**IWSLT17**		**IWSLT17**	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	**0.3**	**0.1**	**7.4**	**2.5**	**36.1**	**35.4**	**22.5**	**17.8**	**19.1**	**19.4**	**24.6**	**22.6**

Consistent improvement over low- and medium resourced language pairs

Languages	Cs	Es	Zh	De	Ru	Fr
Size	11M	15M	25M	28M	29M	41M
RANDOM	16.5	33.2	**35.0**	**30.9**	**31.5**	**41.4**
mBART25	**18.0**	**34.0**	33.3	30.5	31.3	41.0

Does not improve over random baseline for language pairs with large number of translated sentences

mBART

- mBART50 [Tang et al., 2021] offers two main extensions:
  - Extension to 50 languages
  - Fine-tuning on parallel data to give many-to-many translation

Data size	Languages
10M+	German, Czech, French, Japanese, Spanish, Russian, Polish, Chinese
1M - 10M	Finnish, Latvian, Lithuanian, Hindi, Estonian
100k to 1M	Tamil, Romanian, Pashto, Sinhala, Malayalam, Dutch, Nepali, Italian, Arabic, Korean, Hebrew, Turkish, Khmer, Farsi, Vietnamese, Croatian, Ukrainian
10K to 100K	Thai, Indonesian, Swedish, Portuguese, Xhosa, Afrikaans, Kazakh, Urdu, Macedonian, Telugu, Slovenian, Burmese, Georgia
10K-	Marathi, Gujarati, Mongolian, Azerbaijani, Bengali

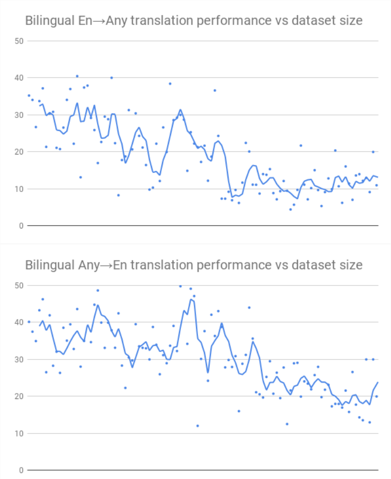
- Both mBART and mBART50 available in HuggingFace
- Basis of much practical work on low-resource MT

Multilingual Models

Idea: Handle all N by N translation directions with a single model (instead of $O(N^2)$)

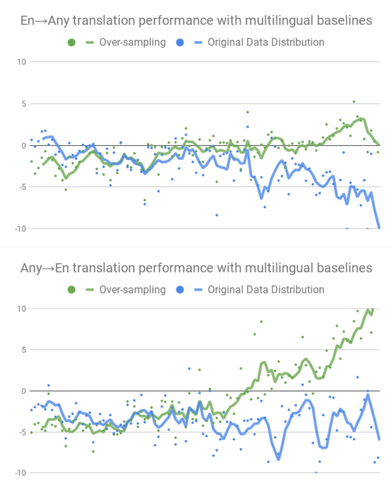
- Usually 1-n or n-1
- Use a small number of related langauges [Mueller et al., 2020]
- Or go big: 103 languages Massively Multilingual Neural Machine Translation in the Wild [Arivazhagan et al., 2019]
- There is a trade-off:
  - Transfer: benefit from addition of other languages
  - Interferance: performance is degraded due to having to also learn to translate other languages
- Benefits are more noticeable for the many-to-English and low-resource pairs
- High-resource pairs tend to be harmed
- Massive systems require capacity

Multilingual Models



BLEU score for langauge pairs ordered from most training data on left to least on the right

Multilingual Models



Difference in BLEU score from bilingual baseline. Blue: original data distribution, Green: equal sampling from all languages

Multilingual Models WMT2021

[Yang et al., 2021]

MMT	Model	cs-en	de-en	ha-en	is-en	ja-en	ru-en	zh-en	Avg	Incremental Δ
✗	**Bilingual**	28.9	41.5	15.9	30.3	19.7	40.2	34.8	30.2	—
✗	**+ Backtranslation**	28.3	38.0	28.3	34.5	21.1	38.0	30.8	31.3	+1.1
✗	**+ Finetuning**	30.4	42.8	30.3	35.5	24.6	39.5	36.2	34.2	+2.9
✓	**+ Multilingual**	32.1	43.8	36.1	39.4	26.7	40.6	36.9	36.5	+2.3
✓	**+ Ensemble**	32.3	44.5	37.2	39.9	27.2	40.9	37.8	37.1	+0.6
✓	**+ Reranking**	32.7	44.4	38.2	40.5	27.8	41.4	38.0	37.6	+0.5

Facebook AI's WMT21 News Translation Task Submission [Tran et al., 2021]

- First place: cs, ha, is

Evaluation

Evaluation of Low-resource MT

- Evaluation of MT is hard anyway
- Is automatic evaluation of low-resource languages harder?
  - Metrics are designed with high-resource langauges in mind
  - Metrics are less reliable on poor systems
  - Lack of good test sets and human evaluations for training metrics
- Human evaluation is preferable
  - Researchers need to connect to language communities

Summary

Where are we now?

- Much progress on low-resource MT
- Much more data for some languages e.g. English–Hindi now has 10M sentence pairs

However:

- NMT models are very data-inefficient
- Lack good techniques for incorporating knowledge
- Vast majority of world's languages not supported
- Need to work with language communities: Masekhane, AmericasNLP

Summary

- Collect more data
- Monolingual Data
- Multilingual Data

Next: Laura Perez NLG: semantic parsing, paraphrasing, data-to-text generation

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G. F., Cherry, C., et al. (2019).
Massively multilingual neural machine translation in the wild: Findings and challenges.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020).
The state and fate of linguistic diversity and inclusion in the nlp world.
In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293.

Kocmi, T. and Bojar, O. (2018).
Trivial transfer learning for low-resource neural machine translation.
In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., et al. (2021).
Quality at a glance: An audit of web-crawled multilingual datasets.
arXiv preprint arXiv:2103.12028.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., et al. (2019).
Choosing transfer languages for cross-lingual learning.
In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3125–3135.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020).
Multilingual denoising pre-training for neural machine translation.
Transactions of the Association for Computational Linguistics, 8:726–742.

Mueller, A., Nicolai, G., McCarthy, A. D., Lewis, D., Wu, W., and Yarowsky, D. (2020).
An analysis of massively multilingual neural machine translation for low-resource languages.
In Proceedings of The 12th language resources and evaluation conference, pages 3710–3718.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabenamualu, S. K., Osei, S., Sackey, F., et al. (2020).
Participatory research for low-resourced machine translation: A case study in african languages.

In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2144–2160.

Sennrich, R., Haddow, B., and Birch, A. (2016).
Improving Neural Machine Translation Models with Monolingual Data.
In
Proceedings of the 54th Annual Meeting of the Association for Compu
pages 86–96, Berlin, Germany.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021).
Multilingual translation from denoising pre-training.
In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3450–3466.

Tran, C., Bhosale, S., Cross, J., Koehn, P., Edunov, S., and Fan, A. (2021).
Facebook ai's wmt21 news translation task submission.
In Proceedings of the Sixth Conference on Machine Translation, pages 205–215.

Yang, J., Ma, S., Huang, H., Zhang, D., Dong, L., Huang, S., Muzio, A., Singhal, S., Hassan, H., Song, X., and Wei, F. (2021).
Multilingual machine translation systems from Microsoft for WMT21 shared task.
In Proceedings of the Sixth Conference on Machine Translation, pages 446–455, Online. Association for Computational Linguistics.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016).
Transfer learning for low-resource neural machine translation.
In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568–1575.