Evaluating Pre-training Objectives for Low-Resource Translation into Morphologically Rich Languages

Prajit Dhar , Arianna Bisazza, Gertjan van Noord

Center for Language and Cognition
University of Groningen

May 10, 2022

rijksuniversiteit groningen

Low Resource Languages

- Lack of linguistic resources
  - Parallel sentences for Machine Translation
  - Annotated sentences
- Huge disparity between available resources[1] and number of L1 speakers[2]
  - Tamil (**10**M sentences vs. **75**M speakers)
  - Turkish (**75**M sentences vs. **80**M speakers)
  - German (**350**M sentences vs. **95**M speakers)
- Many low resource languages are morphologically rich
  - Dravidian languages (Tamil, Telugu, etc.)
  - Bantu languages (Swahili, Xhosa, etc.)

[1] Number of parallel English→Target Language sentences. Source: OPUS

[2] Source: Wikipedia

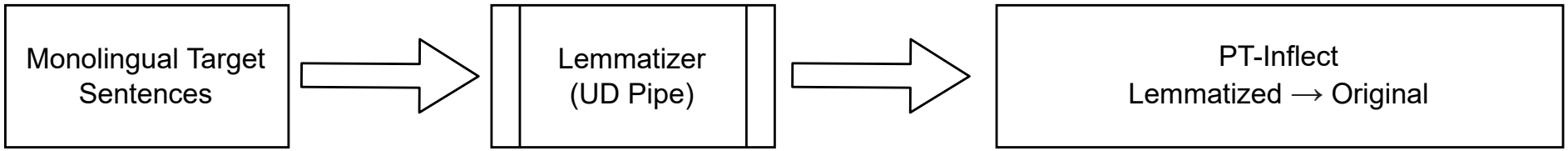
Existing Approaches

- Multilingual NMT
  - Gains seen if a closely related language present
  - Training massively multilingual NMT very expensive
- Pre-training models on monolingual data
  - MASS [Song et al., 2019], mBart [Liu et al., 2020]
  - Source side tokens are randomly replaced by a token $\rightarrow$ Input to the NMT
- Data augmentation using monolingual data
  - Back-Translation [Sennrich et al., 2016]
  - Use backward NMT system to translate target sent.$\rightarrow$source sent.
  - Train new system on Generated source sent.$\rightarrow$Original target sent.

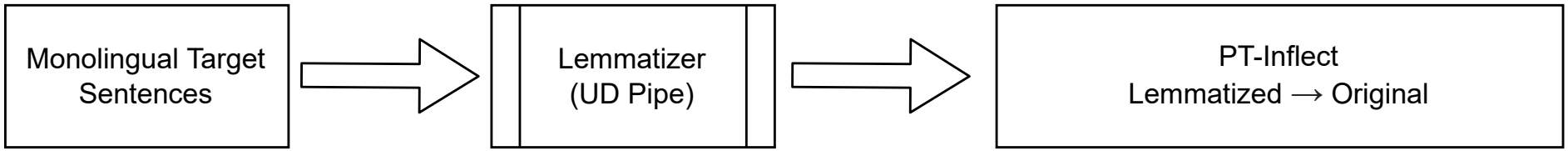
Our Target Languages

Language	Tokens(k)	EN/Trg Token Ratio	Morphology Counting Complexity	Type/Token Ratio
Estonian	72	1.39	110	0.340
Lithuanian	81	1.23	123	0.383
Tamil	26	3.81	201	0.422
Turkish	83	1.21	140	0.307
German	95	1.05	38	0.266
English	100	-	6	0.174

PT-Inflect


Monolingual Target Sentences
Lemmatizer (UD Pipe)
PT-Inflect
Lemmatized → Original

PT-Inflect



- Data Augmentation
  - Run lemmatizer on monolingual target sentences
  - Use lemmatized sentences (T_{lemmas}) as input
  - Use original sentences ($T_{original}$) as output
  - Pre-train NMT on ($T_{lemmas} \rightarrow T_{original}$) dataset

PT-Inflect

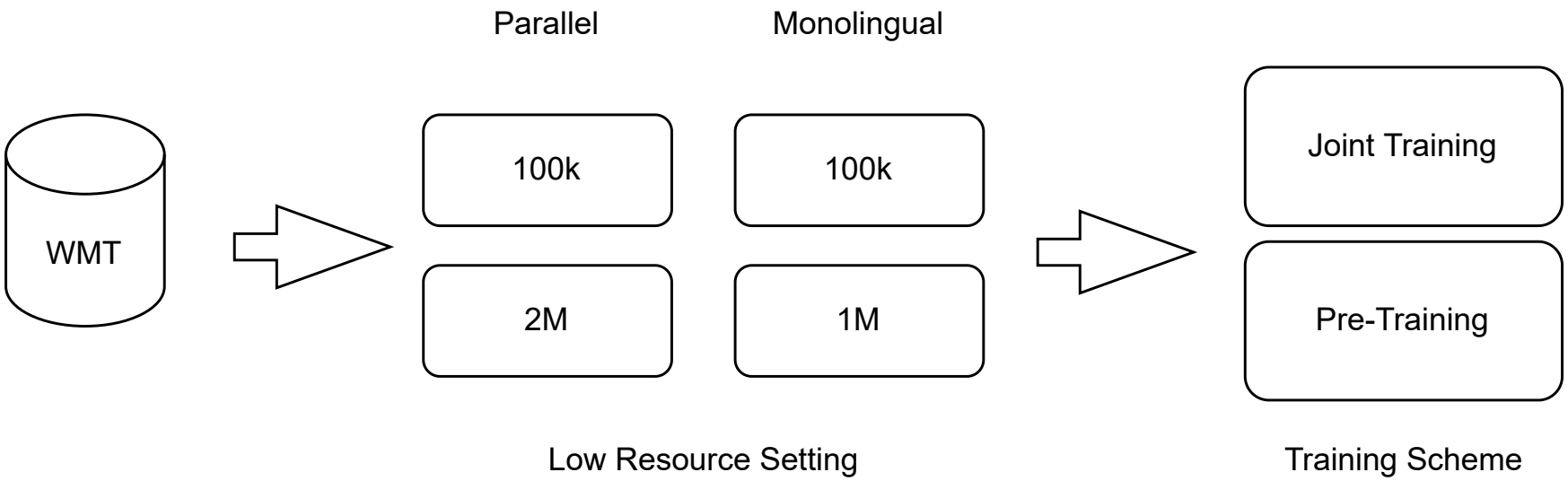
	Lemmatized sentence T_{lemmas} (with EN glosses)		Inflected sentence $T_{original}$ (with EN translation)
ET	Farish ei olema esitama üks süüdistus *Farish no to be brought single charge*	→	Farishile ei olnud esitatud ühtegi süüdistust *No charges had been brought against Mr Farish*
LT	kelias m8 po policija incidentas vėl atidaryti *way m8 after police incident again open*	→	Kelias M8 po policijos incidento vėl atidarytas *M8 fully reopens after police incident*
TR	bu ol *this to be*	→	Bu olmayacak *That's not going to happen*
TA	kaṭṭāyam avar nam iḻivupaṭuttu *viṭakkūṭu(m) *force he us humiliate may*	→	kaṭṭāyam avarkaḷ nammai iḻivupaṭutta viṭakkūṭātu *We must not let them humiliate us*
DE	dass man er|es|sie selbst treu bleiben müssen *that one it self true to remain must*	→	Dass man sich selbst treu bleiben muss *That you have to be true to who you are*

ET: Estonian, LT: Lithuanian, TR: Turkish, TA: Tamil and DE: German

PT-Inflect

- Intuition
  - Learning morphology internally is difficult for NMTs
    - Need to be provided as a learning task
- Advantage
  - Cheaper than Back-translation
    - No need to train an additional NMT

System Overview



Comparison

- Baseline (training only on Parallel Data)
- PT-Inflect
- Back-Translation [Sennrich et al., 2016]
- StupidBT [Burlot and Yvon, 2018]
- Random Masking (15 or 50) [Raffel et al., 2020]

Comparison

- Baseline (training only on Parallel Data)
- PT-Inflect
- Back-Translation [Sennrich et al., 2016]
- StupidBT [Burlot and Yvon, 2018]
- Random Masking (15 or 50) [Raffel et al., 2020]

Objective	Source	Target
PT-Inflect	şerif ofi ve 911 ara	Şerifin ofisini ve 911'i araması
Back-translation	Sheriff calls her office and 911	Şerifin ofisini ve 911'i araması
StupidBT	S_Şerifin S_ofisini S_ve S_911'i S_araması	Şerifin ofisini ve 911'i araması
RandomMask50	<MASK>ofisini ve <MASK>araması	Şerifin ofisini ve 911'i araması

Metrics

- CHRF++ [Popović, 2015]
  - our main metric
  - N-grams of character matches vs. full word matches with BLEU
  - Correlates better with human evaluation, when target language has rich morphology [Bojar et al., 2016]
- BLEU computed by SacreBLEU [Post, 2018]
  - less informative for morphologically rich languages
  - see scores in the paper

Results: Baseline vs. others

Results for 100k parallel sentences and 1M monolingual sentences

	Model Name	CHRF
ET	BASELINE	22.3
	RandMask15	24.0
	StupidBT	25.2
	BT	**27.2**
	PT-Inflect	27.1
LT	BASELINE	27.8
	RandMask15	27.7
	StupidBT	28.7
	BT	**32.1**
	PT-Inflect	29.9
TA	BASELINE	20.2
	RandMask15	24.8
	StupidBT	24.3
	BT	**27.1**
	PT-Inflect	26.8

	Model Name	CHRF
TR	BASELINE	19.9
	RandMask15	20.2
	StupidBT	20.5
	BT	**26.6**
	PT-Inflect	21.0
DE	BASELINE	34.5
	RandMask15	38.1
	StupidBT	40.2
	BT	42.9
	PT-Inflect	**43.0**

Combine

- PT-Inflect outperforms Baseline
- Gains of PT-Inflect do not correlate to morphological complexity
  - Compare German and Turkish (+8.5 vs +1.1)
  - Other factors at play

- However BT better than PT-Inflect for most settings
  - Would the combination of both techniques improve translation?

- Combine
  - 1M sentence pairs from each BT and PT-Inflect objective
  - NMT trained first on this joint data, then on parallel sentences

Results - Combine vs. others

	Model Name	CHRF
ET	PT-Inflect	27.1
	BT	**27.2**
	Combine	27.0
LT	PT-Inflect	29.9
	BT	32.1
	Combine	**33.2**
TA	PT-Inflect	26.8
	BT	**27.1**
	Combine	26.3

	Model Name	CHRF
TR	PT-Inflect	21.0
	BT	**26.6**
	Combine	24.8
DE	PT-Inflect	43.0
	BT	42.9
	Combine	**44.4**

Conclusions

- PT-Inflect outperforms Parallel (Baseline) systems
- But BT remains the better option
- Combination of PT-Inflect and BT better for some languages

References I

Bojar, O., Graham, Y., Kamran, A., and Stanojević, M. (2016).
Results of the WMT16 metrics shared task.
In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Burlot, F. and Yvon, F. (2018).
Using monolingual data in neural machine translation: a systematic study.
In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

References II

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.

Popović, M. (2015).
chrF: character n-gram F-score for automatic MT evaluation.
In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

References III

Post, M. (2018).
A call for clarity in reporting BLEU scores.
In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020).
Exploring the limits of transfer learning with a unified text-to-text transformer.
Journal of Machine Learning Research, 21(140):1–67.

References IV

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 Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019).
Mass: Masked sequence to sequence pre-training for language generation.
In *International Conference on Machine Learning*, pages 5926–5936.