A Survey On Low-Resource Machine Translation

Outline

- Language-independent challenges in low-resource setting
  - Insufficient parallel data
  - Insufficient monolingual data
  - Lack of computational linguistic studies
- Language-specific challenges in low-resource setting
  - Complex morphological system
  - Lack of standard orthography
- Solutions
  - Data augmentation (data-wise)
  - Hyperparameter tuning, transfer learning (model-wise)
  - Build computational linguistic tools from scratch (data-and-model-wise)

Language-Independent Challenges in Low-Resource Setting

- Insufficient parallel data
    - hard to build translation model
- Insufficient monolingual data
    - hard to build language model
        - hard to train good semantic representation (embedding)

Insufficient Parallel Data

Language	Code	Main location	Speakers
Aymara	aym	Bolivia	1,677,100
Asháninka	cni	Peru	35,200
Bribri	bzd	Costa Rica	7,000
Guarani	gn	Paraguay	6,652,790
Hñähñu	oto	Mexico	88,500
Nahuatl	nah	Mexico	410,000
Quechua	quy	Peru	7,384,920
Rarámuri	tar	Mexico	9,230
Shipibo-Konibo	shp	Peru	22,500
Wixarika	hch	Mexico	52,500

Languages	Train	Dev	Test
es-aym	6,531	996	1,003
es-cni	3,883	883	1,003
es-bzd	7,506	996	1,003
es-gn	26,032	995	1,003
es-oto	4,889	599	1,003
es-nah	16,145	672	1,003
es-quy	125,008	996	1,003
es-tar	14,720	995	1,003
es-shp	14,592	996	1,003
es-hch	8,966	994	1,003

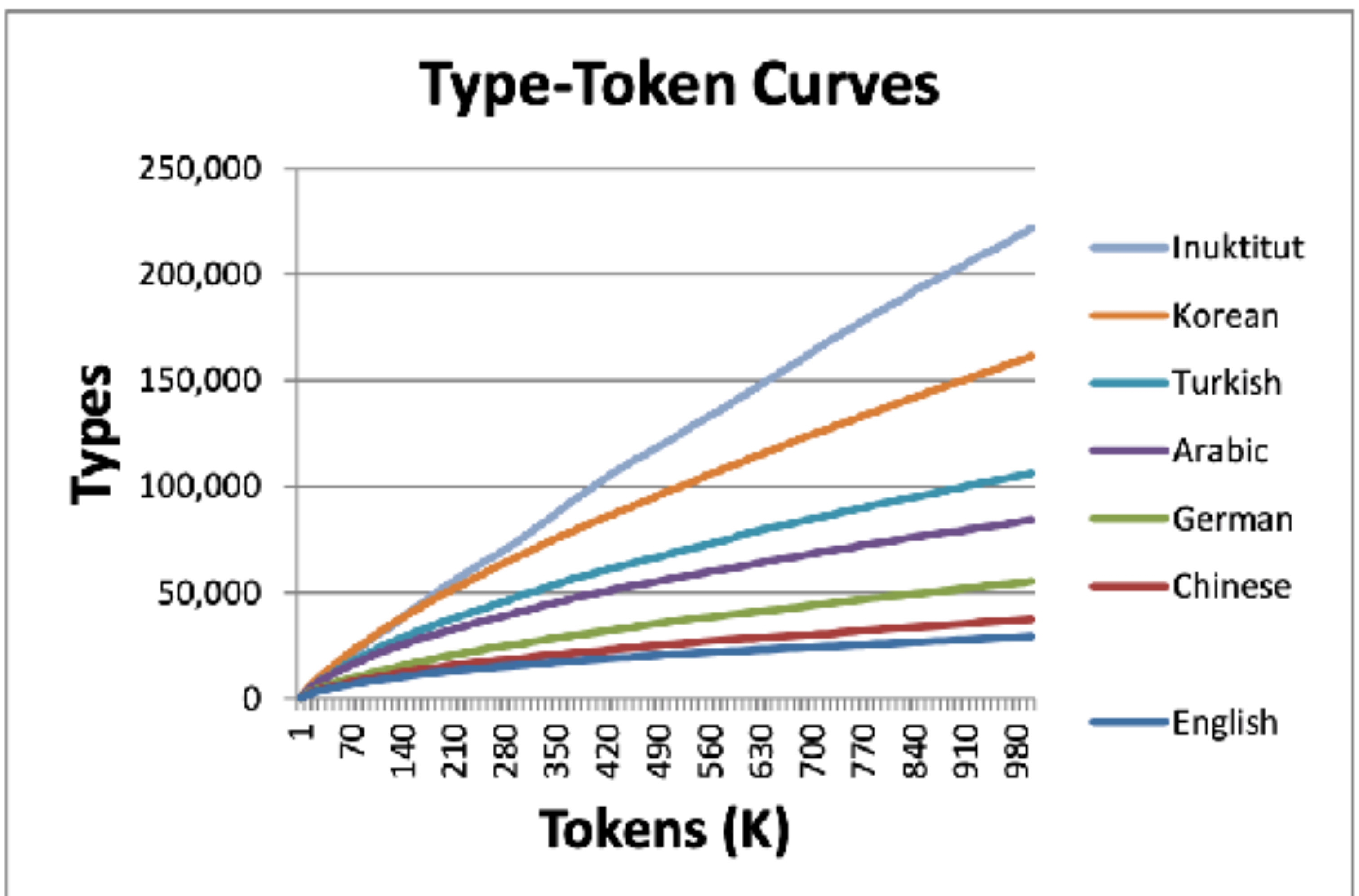
Insufficient Monolingual Data

Target language	Wikipedia		Bible	
	Size (MB)	Sentences	Size (MB)	Sentences
Hñähñu	-	-	1.4	7.5K
Wixarika	-	-	1.3	7.5K
Nahuatl	5.8	61.1K	1.5	7.5K
Guarani	3.7	28.2K	1.3	7.5K
Bribri	-	-	1.5	7.5K
Rarámuri	-	-	1.9	7.5K
Quechua	5.9	97.3K	4.9	31.1K
Aymara	1.7	32.9K	5	30.7K
Shipibo-Konibo	-	-	1	7.9K
Asháninka	-	-	1.4	7.8K
Spanish	1.13K	5M	-	-
Total	1.15K	5.22M	19.8	125.3K

Language-Specific Challenges in Low-Resource Setting

- Complex morphological system
  - Sentence-word (Jeffrey C. Micher, 2018)

Qanniqlaunngikkalauqtuqlu
qanniq-lak-uq-nngit-galauq-tuq-lu
snow-a_little-frequently-NOT-although-3.IND.S-and
"*And even though it's not snowing a great deal,*"



Language-Specific Challenges in Low-Resource Setting

- Lack of standard orthography ((Jeffrey C. Micher, 2018)
  - Same word, various spellings

Haamalaujunut
Haamlaujunut
Hamalakkunnit
Hammakkut
Hammalakkunnut
Hammalat
Hmlatni

Solutions

- Data-wise
  - Back translation
  - Iterative back translation
- Model-wise
  - Hyperparameter tuning
    - Empirical hyperparameters selection
  - Transfer learning
    - Transfer knowledge from high-source to low-resource
- Data-model-wise
  - Build computational linguistic tools from scratch

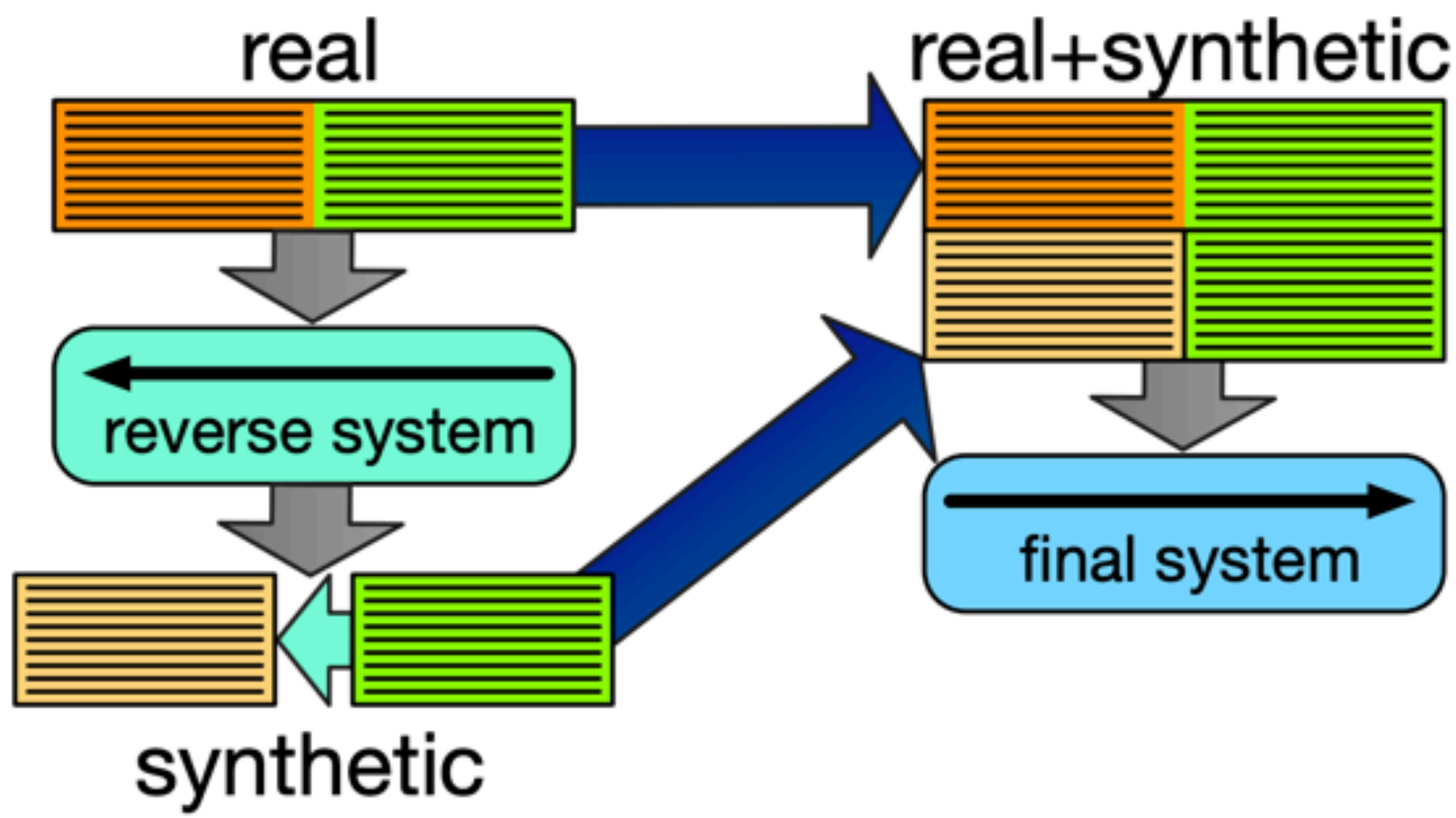
Back Translation

- A method to generate silver (synthetic) parallel data
- Let src-tgt be a language pair (where src is low-resource)
  - train a tgt-src MT model T0 with gold parallel data
  - collect extra monolingual data of tgt (mono_tgt)
  - translate mono_tgt to src with T0 (silver data gotten)
  - train a src-tgt MT model with gold and silver data

- If tgt is high-resource language -> much silver data

Back Translation

- (Hoang et al. 2018)



Iterative Back Translation

- (Hoang et al. 2018) (Feldman et al. 2020)

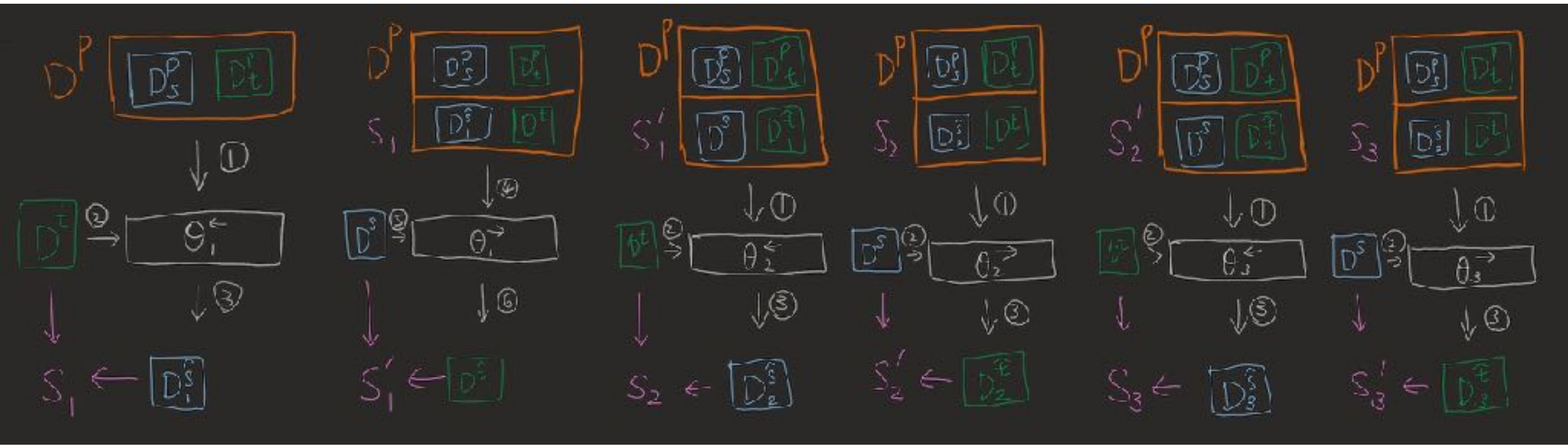
Algorithm 1 Iterative Back-Translation

Input: parallel data D^p, monolingual source, D^s, and target D^t text

```
Let T← = D^p
repeat
    Train target-to-source model Θ← on T←
    Use Θ← to create S = {(ŝ, t)}, for t ∈ D^t
    Let T→ = D^p ∪ S
    Train source-to-target model Θ→ on T→
    Use Θ→ to create S′ = {(s, t̂)}, for s ∈ D^s
    Let T← = D^p ∪ S′
until convergence condition reached
```

Output: newly-updated models $\Theta_{\leftarrow}$ and $\Theta_{\rightarrow}$

Iterative Back Translation



Hyperparameter Tuning

- (Sennrich et al. 2019)
  - Model with lower capacity (fewer layers)
  - Smaller vocabulary for BPE
    - Higher frequency threshold for subword units
    - Lower pre-set max vocabulary size
  - Smaller batch size
  - Higher dropout rate

Transfer Learning

- (Zoph et al. 2016)
- Let a low-resource language be L and the real desired MT model to be L-English.
  - collect large parallel data for example: French and English.
  - train a French-English MT model M0 with large data
  - initialize a new model M1 with same weights of M0
  - English embeddings are retained and frozen
  - L's tokens are randomly mapped to French embeddings
  - jointly train L's embedding and M1

Computational Linguistic Tools

- Inuktitut Morphological Analyzer
  - The Uqailaut Project (Farley, 2009)
  - A rule-based morphological segmentation model
  - piqujivungaarutiksanut ->pi ^ qu ^ ji ^ vungaa ^ ruti ^ ksa ^ nut

Inuktitut Computing

The UQAILAUT Project

Thank You!