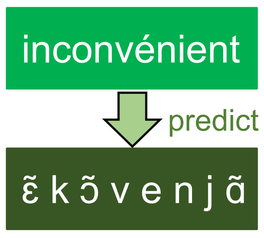
Low-Resource G2P and P2G Conversion with Synthetic Training Data

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Greg Kondrak

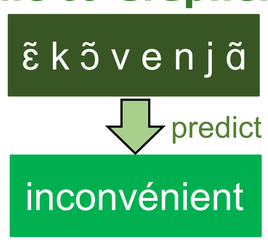
UNIVERSITY OF ALBERTA

Introduction

Grapheme to Phoneme (G2P)



Phoneme to Grapheme (P2G)



→ Alberta participated in SIGMORPHON tasks in 2016, 2017, 2018, and 2019
→ 2020 (Task 1) : **Multilingual Grapheme-to-Phoneme (G2P) conversion**
→ Default setting: 3600 training instances
→ We define a **low-resource** setting with **100 training instances**
→ We also perform **Phoneme-to-Grapheme (P2G) conversion**
→ Key ideas: **Augment training data** using a combination of diverse models

Tools & Data

Tools

→ **DTLM** (Nicolai et al, 2018): Combines discriminative transduction with character and word language models.
→ **M2M+** (Jiampojamarn, 2007): Performs high-precision many-to-many alignment by handling insertions.

Data

→ Grapheme-phoneme pairs in 15 languages. **LR setting** : 100 random pairs per language, **HR setting** : Full Dataset.
→ Augmentation data : Unannotated Wikipedia corpus.

Baselines

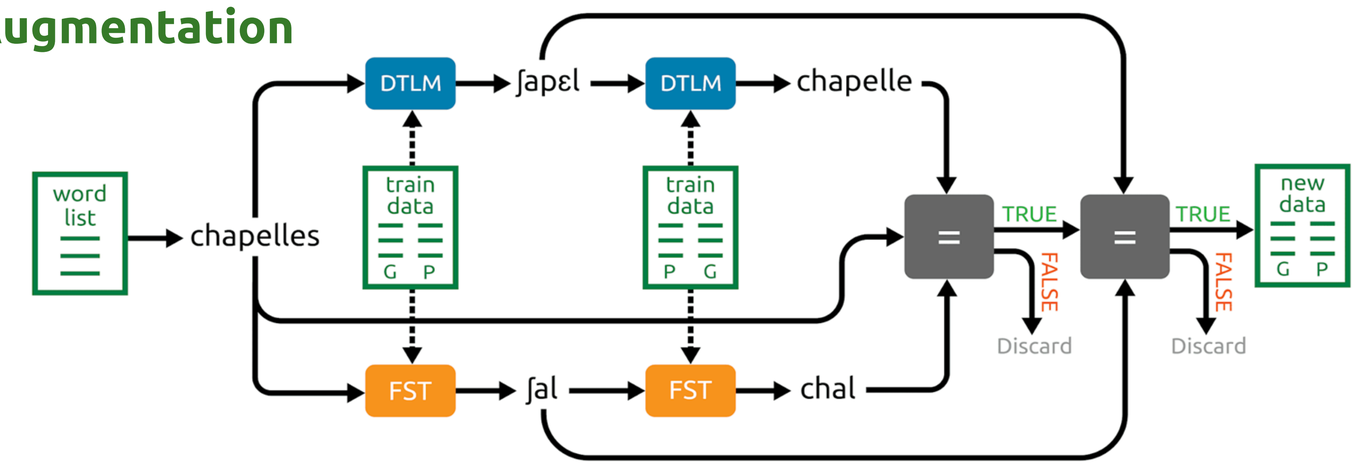
→ **FST:** Finite State Transducer, tuned on the size of n-grams.
→ **Transformer:** Encoder-decoder architecture with self-attention, which requires extensive parameter tuning.

Our Methods

Discriminative String Transduction (DTLM)

- Dynamic programming core using a set of feature templates.
- Feature set includes context. transition and joint-n gram features.
- Depends on a high-precision one to many alignment by M2M+ aligner.
- Accuracy enhanced by target character and language models

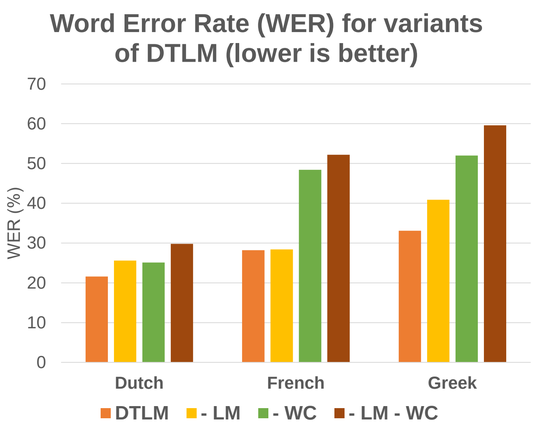
Data Augmentation



→ We introduce a method to synthesize additional training data from unannotated text.
1. Train FST and DTLM on for both G2P and P2G tasks.
2. Provide a word to the G2P models to produce phonemes
3. Provide the phonemes to P2G models to produce graphemes
4. Include the grapheme-phoneme pair in the new data if:
   a. The resulting graphemes for both FST and DTLM, match the initial word.
   b. The phonemes produced by FST and DTLM are the same.

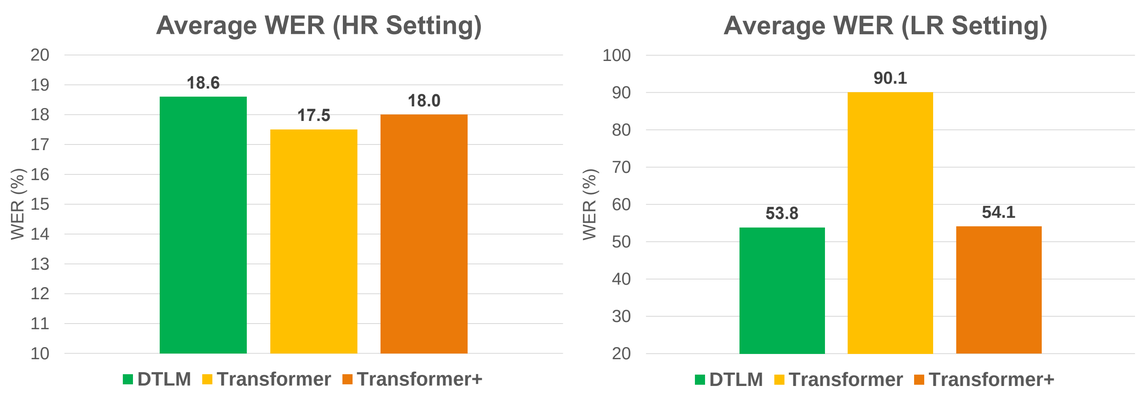
Results

DTLM Ablation Results (Phoneme-to-Grapheme)

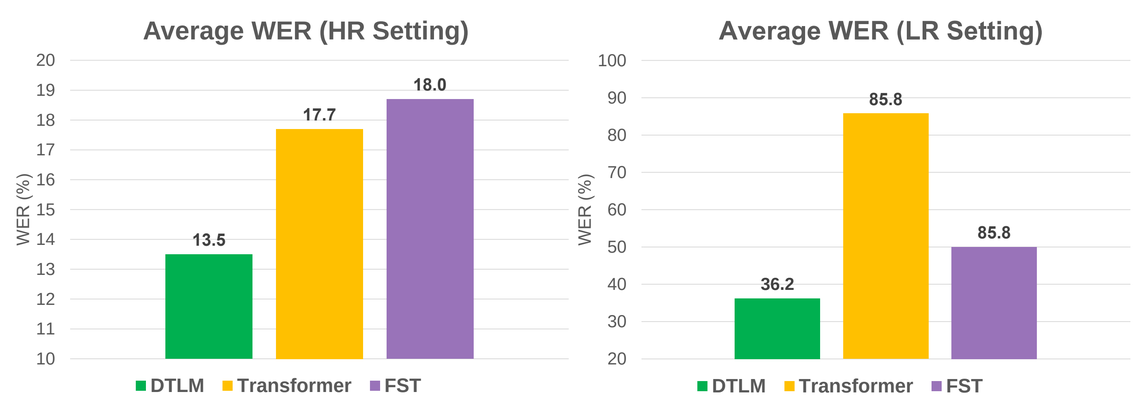


- Demonstrates impact of Word Counts (WC) and character language models (LM) on P2G.
- WC helps more than target LMs.
- Without these two components, DTLM results are in the same range as baselines.

Grapheme-to-Phoneme Test Results



Phoneme-to-Grapheme Test Results



Discussion

- Transformer fails completely in LR setting without synthetic training data, on both the G2P and P2G tasks.
- Transformer+, which takes advantage of synthetic data produced by our method, shows substantial improvement (>35%) in the LR setting of the G2P task, demonstrating the utility of our method.
- Synthetic training data can add information.
- In the HR G2P task, Transformer+ performed better than Transformer (without additional synthetic data) in our development experiments. Unfortunately, this is not reflected in the test results. We suspect this is due to tuning and hyperparameter issues.
- On the P2G task, DTLM obtains. substantially lower error rates than the other two systems, in both the HR and LR settings. DTLM remains the state of the art for P2G.
- Overall: Strong proof-of-concept for our data augmentation approach in LR settings.

Conclusion

- We proposed a novel data augmentation approach combining multiple string transduction methods.
- We explored both G2P and P2G tasks in both high-resource and low-resource settings.
- Our results demonstrate that the weakness of neural systems in low-resource settings can be mitigated through data augmentation.

Acknowledgments

- We thank Garrett Nicolai for the assistance with DTLM.
- We thank the shared task organizers for their effort.
- This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).