Ensemble Self-Training for Low-Resource Languages: Grapheme-to-Phoneme Conversion and Morphological Inflection

Xiang Yu　　Ngoc Thang Vu　　Jonas Kuhn

IMS, University of Stuttgart, Germany

{xiangyu,thangvu,jonaskuhn}@ims.uni-stuttgart.de

Summary

- Iterative ensemble optimization and data augmentation
- Based on large amount of diverse simple models
- Effective for low-resource scenarios
- 1st in the grapheme-to-phoneme conversion task
- 4th in the morphological inflection task

Ensemble Self-Training

General Workflow

function ENSEMBLESELFTRAINING(L, U, T)
Require: labeled data L, unlabeled data U, model types T
- Initial data $L_0 = L$
- Model pool $M = \emptyset$
- **for** $n : 0 \ldots N$ **do** ▷ each iteration
  - **for** $t^k \in T$ **do** ▷ each model type
    - $m_n^k = \text{TRAIN}(t^k, L_n)$ ▷ train new models
    - $M = M \cup \{m_n^k\}$ ▷ add to model pool
  - **end for**
  - $E = \text{SEARCHENSEMBLE}(M)$ ▷ find optimal ensemble
  - Sample $u \sim U$ ▷ sample unlabeled data
  - $l = \text{SELECTDATA}(E, u)$ ▷ select reliable prediction
  - $L_{n+1} = \text{AGGREGATEDATA}(L_n, l)$ ▷ add as labeled data
  - $U = U - l$
- **end for**
- **return** E, L_k

end function

Ensemble Search

- Search for the optimal ensemble with genetic algorithms
- A binary code to represent an ensemble, e.g. 0101001110
- Fitness of an ensemble is the accuracy on the dev set
- Use selection, crossover, and mutation to evolve the ensemble

Data Selection

- Use the ensemble to predict a batch of unlabeled data
- Select the data with high agreement in the ensemble
- Add as new training data for the next iteration

Grapheme-to-Phoneme

Task & Data

- Task 1 of SIGMORPHON Shared Task [Gorman et al. 2020]
- Map a sequence of graphemes to a sequence of phonemes
  e.g. excuser → ɛkskyze
- Unlabeled data: word lists mostly extracted from OpenSubtittles

Models

- 4 types of models:
  (1) baseline pair n-gram model [Lee et al. 2020]
  (2) seq2seq model with soft attention [Luong et al. 2015]
  (3) seq2seq model with hard monotonic attention [Aharoni and Goldberg 2017]
  (4) hybrid seq2seq/tagging model: predict a short sequence for each input character
- Paired with l2r and r2l generation directions and 2 random seeds

Results

Model	WER	PER
IMS	13.81	2.76
CLUZH	14.13	2.82
DeepSPIN-3	14.15	2.92
CU-1	14.52	3.24
Pair n-gram	22.00	4.92
LSTM	16.84	3.99
Transformer	17.51	4.30

Table: Average word error rates (WER) and phone error rates (PER) on test set.

- IMS ranks 1st among the participants
- Outperforms all baselines
- How much contribution comes from
  - ensemble of simple models?
  - diversity of model types?
  - data augmentation?

Analysis

	average	ensemble
default	17.6	10.7
-diversity	16.2	11.2
-augment	18.1	10.1

Table: WER of the model average and the ensemble on dev set.

- Analyze the contribution of each factor:
  ✓ Ensemble much better than single models
  ✓ Lower model diversity (only hybrid model) leads to lower ensemble performance despite higher average model performance
  ✗ Worse performance with data augmentation

	average	ensemble
default	35.5	25.2
-augment	53.4	29.2

Table: WER in low-resource scenario.

- Simulate low-resource scenario:
  - 200 training instances for each language
  ✓ Better performance with data augmentation

Morphological Inflection

Task & Data

- Task 0 of SIGMORPHON Shared Task [Vylomova et al. 2020]
- Generate inflected word form from lemma and morphological features
  e.g. jagen + V;SBJV;PL;3;PST → jagten
- Unlabeled data: recombine the lemma and morphological features

Models

- 2 Types of models:
  (1) seq2seq model with soft attention [Luong et al. 2015]
  (2) seq2seq model with hard monotonic attention [Aharoni and Goldberg 2017]
- Paired with l2r and r2l generation directions and 2 random seeds

Results

Model	Accuracy
CULing-01-0	0.912
DeepSPIN-02-1	0.909
UIUC-01-0	0.905
IMS-00-0	0.892
LSTM	0.858
LSTM+Aug	0.888
Transformer	0.901
Transformer+Aug	0.903

Table: Average accuracy on test set.

- IMS ranks 4th among the participants
- Outperforms LSTM baselines but not Transformer baselines
- Training data size varies from 10^2 to 10^5, how well do the models perform with different data sizes?

Analysis

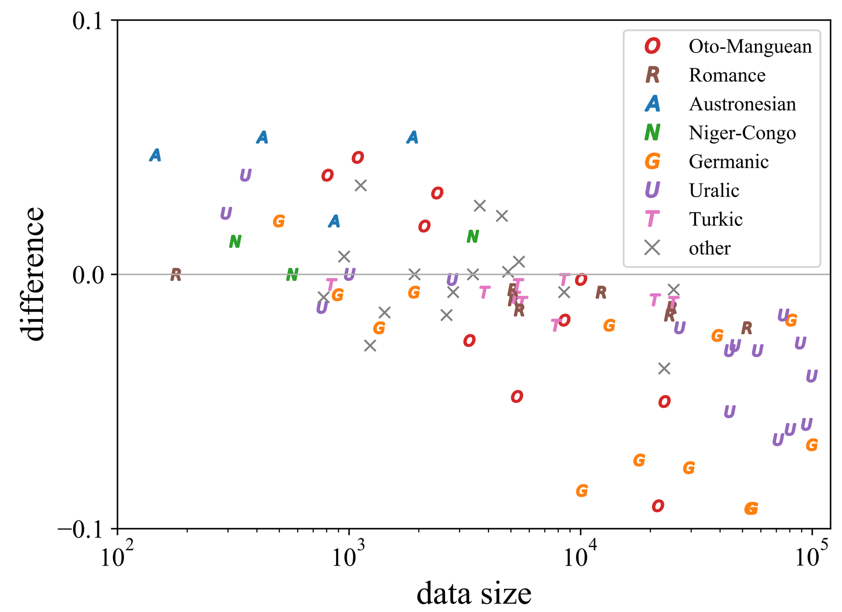

Figure: Performance difference between our system and Transformer+Aug wrt. training data size.

- Our system performs relatively better in low-resource scenarios
- No clear relation between performance and language family

Acknowledgments

This work was in part supported by funding from the Ministry of Science, Research and the Arts of the State of Baden-Württemberg (MWK), within the CLARIN-D research project.

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *SIGMORPHON*.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation mining with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4216–4221, Marseille.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Ponti, Rowan Hall Maudslay, Ran Zmigrod, Joseph Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrej Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. The SIGMORPHON 2020 Shared Task 0: Typologically diverse morphological inflection. In *SIGMORPHON*.