Examination and Extension of Strategies for Improving Personalized Language Modeling via Interpolation

LIQUN SHAO (PRESENTER), SAHITYA MANTRAVADI, TOM MANZINI, ALEJANDRO BUENDIA, MANON KNOERTZER, SOUNDAR SRINIVASAN, CHRIS QUIRK

Microsoft

Background & Objectives

The task of language modeling: to predict which words come next, given a set of context words

Language modeling can be used to aid the user during composition by suggesting words, phrases, sentences, and even paragraphs that complete the user's thoughts, which is one developing use case for Natural language interfaces (NLIs)

This paper describes methods for handing the crucial edge case of out-of-vocabulary (OOV) tokens and interpolation coefficient optimization

Model is to be used to **re-rank sentence completion sequences**

Main contributions

Evaluate several approaches to handle OOV tokens, covering edge cases not discussed in the LM personalization literature

Provide novel analysis and selections of interpolation coefficients for combining global models with user-personalized models

Experimentally analyze trade-offs and evaluate our personalization mechanisms on public data, enabling replication by the research community

Personalized Interpolation Model

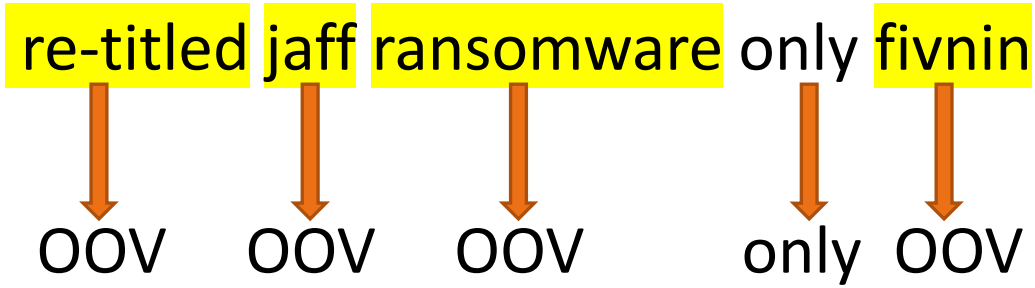
- Explore using a combination of both large-scale neural LMs and small-scale personalized n-gram LMs, previously has been studied from Chen et al., 2015
- Extend previous work by computing the perplexity of these models not by exponentiation of the cross entropy, but by explicitly predicting the probability of test sequences
- Evaluation metrics:
  - Perplexity (PP) ↓: the exponentiation of the entropy of a probability distribution
  - Lift in perplexity (PP lift) ↑:

$$PP\ lift = \frac{PP_{global} - PP_{interpolated}}{PP_{global}}$$

- Interpolation strategies:
  - $P = \alpha P_{personal} + (1 - \alpha) P_{global}$
  - α – interpolation coefficient, indicates how much personalization is added to the global model

Example of OOV issue

- Vocab = [only, …]
- Reddit User Comment:



Prediction: oov oov oov only oov

Probability: 0.5 0.5 0.5 0.4 0.5

Issue: PP is low, but the quality of predictions is poor because of high probability of OOV tokens as they occur more frequently than the tokens in the vocab

OOV Mitigation Strategies

1. PP_{base}: Do nothing, assigning OOV tokens their estimated probabilities
2. PP_{skip}: Skip the OOV tokens, scoring only those items known in the training vocabulary
3. $PP_{backoff}$: Back-off to a uniform OOV penalty, assigning a fixed probability Φ to model the likelihood of selecting the OOV token
   - Φ: hyperparameter needs to be tuned for each user case
   - In our experiments: $\emptyset = \frac{1}{Vocab\_size}$

Experiment Setup

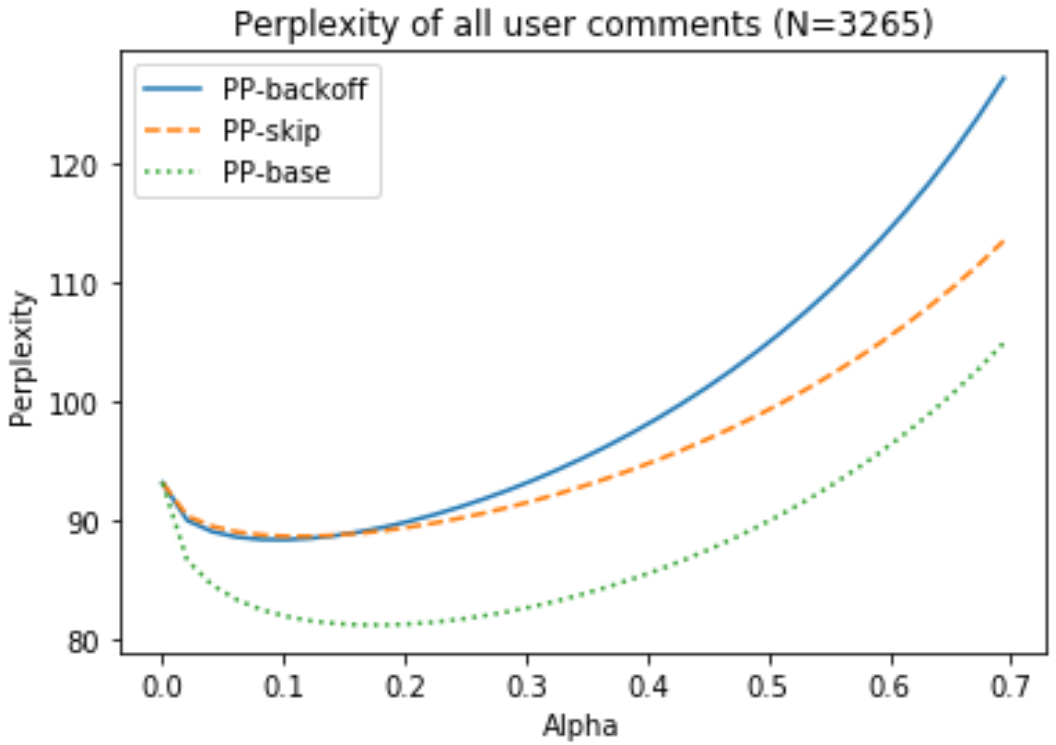
Data:

- Reddit:
  - Public, posts and comments, linked to usernames, time- and date-marked
  - High variability of vocabulary - variety of topics (subreddits), informality of language, and volume of data
- Global LSTM:
  - from 2016 to 2018, sampled Training – validation – testing (70%-20%-10%) of users
  - 10 billion tokens in training data, 29 million unique ones
- Personalized n-grams:
  - all comment data from 3265 random Reddit users

Vocab size *n*:

- Need to be tuned based on data
- Very few gains in user-level OOV rates with the vocab size from 50k to 1M
- Chose vocab size – **50k** for our experiments

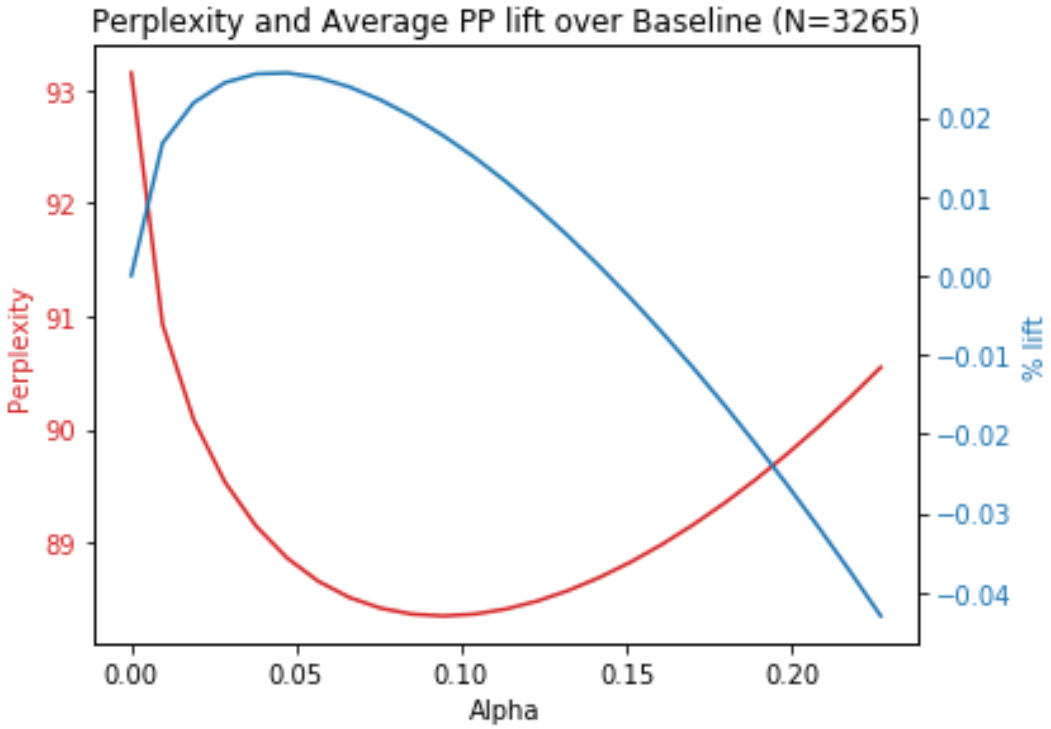
OOV Mitigation Experiment



Average of interpolated PP for all users for varied values of $\alpha < 0.7$ for each method of approaching OOV tokens

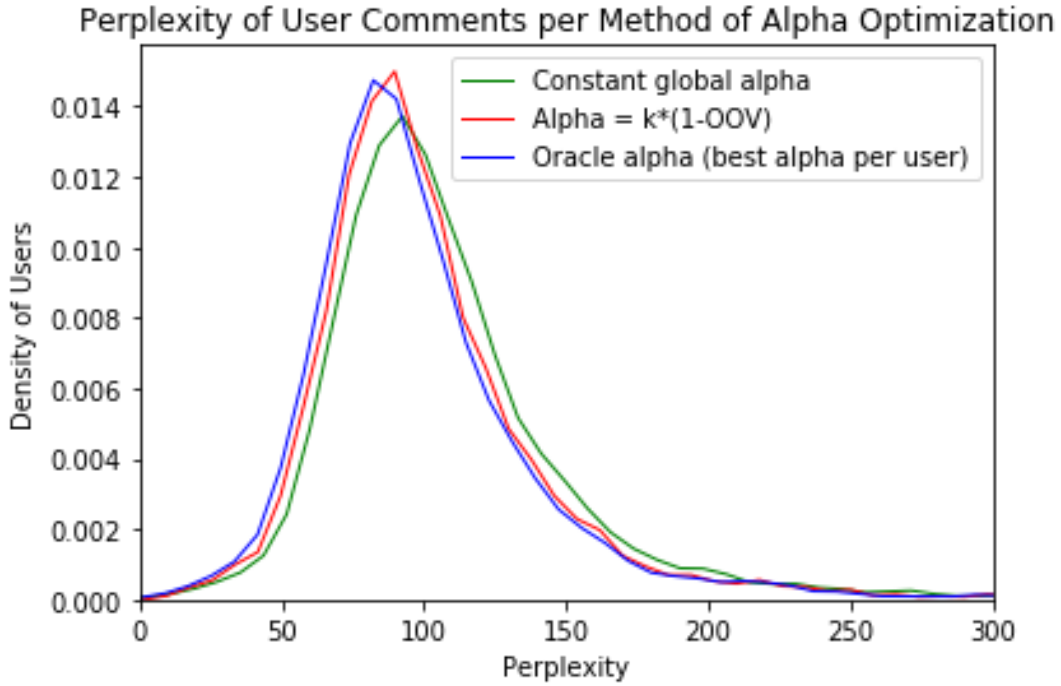
- PP_{base}: issues with low PP but poor performance, disconnected from downstream use in NLI
- PP_{skip}: mathematical issue if all tokens are OOV, PP will be infinite
- $PP_{backoff}$:
  - measures near the minima that closely aligned with PP_{skip} while also free of the mathematical and procedural issues associated with PP_{skip} and PP_{base}
  - presents the most accurate picture of model performance. Use this for following experiments

Analysis of Personalization



$PP_{backoff}$ and average $PP_{backoff}$ over baseline for various values of $\alpha < 0.22$

Improve prediction results ($PP_{backoff}$) over users (α=0.105) N = 3265	Improve some users… A lot (α=0.041) N = 3265
Improves 67.3% of users	Improves 74.2% of users
Average lift over baseline 2.5%	Average lift over baseline **2.7**%



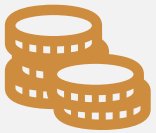
Distribution of interpolated $PP_{backoff}$ for users using each method of α optimization. The values for $\alpha = k * (1 - OOV\ rate)$ are averaged over 10 random selections

Interpolation Coefficient α Optimization

- Compute a set of oracles α values that empirically minimize interpolated perplexity, and achieve the best average $PP_{backoff}$ lift – **6.1%**
- $\alpha = k * (1 - OOV\ rate)$ to optimize k
  - $PP_{backoff}$ lift of **5.2%**, and **80.1%** of users achieve the best user improvement
  - Achieves near-oracle performance
  - Yields lower $PP_{backoff}$ for more users than using a constant α value

Conclusion

Presented new strategies for interpolating personalized LMs



Discussed strategies for handling OOV tokens to give better vision into model performance

Evaluated these strategies on public data allowing the research community to build upon these results

Acknowledgements

- Microsoft's AI Development Acceleration Program (MAIDAP)
  - Vijay Ramani
- Microsoft Search, Assistant and Intelligence (MSAI) team
  - Geisler Antony, Kalyan Ayloo, Mikhail Kulikov, Vipul Agarwal, Anton Amirov, Nick Farn and Kunho Kim
- Reviewer: T.J. Hazen

Thank You!