

Neural Search Space in Gboard Decoder

Yanxiang Zhang*, Yuanbo Zhang*, Haicheng Sun, Yun Wang, Billy Dou,
Gary Sivek, Shumin Zhai

Google Inc

zhangyx,zyb,haicsun,wyun,billydou,gsivek,zhai@google.com

Abstract

Gboard Decoder produces suggestions by looking for paths that best match input touch points on the context aware search space, which is backed by the language Finite State Transducers (FST). The language FST is currently an N-gram language model (LM). However, N-gram LMs, limited in context length, are known to have sparsity problem under device model size constraint. In this paper, we propose **Neural Search Space** which substitutes the N-gram LM with a Neural Network LM (NN-LM) and dynamically constructs the search space during decoding. Specifically, we integrate the long range context awareness of NN-LM into the search space by converting its outputs given context, into the language FST at runtime. This involves language FST structure redesign, pruning strategy tuning, and data structure optimizations. Online experiments demonstrate improved quality results, reducing Words Modified Ratio by [0.26%, 1.19%] on various locales with acceptable latency increases. This work opens new avenues for further improving keyboard decoding quality by enhancing neural LM more directly.

1 Introduction

Gboard is a statistical-decoding-based keyboard on mobile devices developed by Google. Statistical decoding is far more necessary than one might think due to the error-prone process of “fat finger” touch input on small screens. According to Azenkot and Zhai (2012), the per-letter error rate is around 8%-9% without decoding. With decoding, typos such as substitutions (due to the proximity of two keys or cognitive misspellings), omissions, insertions, and transpositions could be automatically corrected by the key-correction and (word) auto-correction functions in the Gboard decoder, leading to an error-tolerant user experience. Powered by language models (LM), the Gboard decoder also

provides rich functionalities such as word completion, post correction, next word prediction, smart compose (in-line predictions) to further save users’ physical input effort.

The decoding process involves two phases: building search space (decoder graph), and performing beam search within the space based on user touch inputs. Gboard decoder utilizes context, a lexicon and language transducers - the familiar $C \circ L \circ G$ composition (Ouyang et al., 2017; Hellsten et al., 2017) - to construct the search space. C is a bi-key key to key transducer while L is a key to word transducer, C and L are statically composed together offline since the size is small. Fig. 1-A illustrates how gesture typing and tap typing inputs are converted into bi-keys and Fig. 1-B illustrates a composed $C \circ L$ targeting four words. Before this work, G is a N-gram language FST containing 64k words for n-grams and 170k words for uni-grams. Composition between $(C \circ L)$ and G are dynamically conducted due to the large size of G . Fig. 1-C shows a simple G containing only four words, and Fig. 1-D illustrates a composed $(C \circ L) \circ G$, which is similar to $(C \circ L)$ but with weights achieved by using the look-ahead composition filters proposed by Allauzen et al. (2009, 2011).

In practice, the whole search space of $(C \circ L) \circ G$ like Fig. 1-D can’t be fully expanded due to its huge size. Only states which are close to users’ bi-key inputs will be expanded. Specifically, the states are pruned based on the combination of LM scores and spatial scores in the decoder graph while the user is typing.

In this work, the N-gram LM is replaced by the NN-LM. Due to the framework complexity brought by the rich functionalities, we propose a runtime conversion solution to minimize the framework change. We call the search space built on the NN-LM the **Neural Search Space (NSS)**, and the original one the N-gram Search Space.

Algorithm 1 and Algorithm 2 describe the

*Equal contribution.

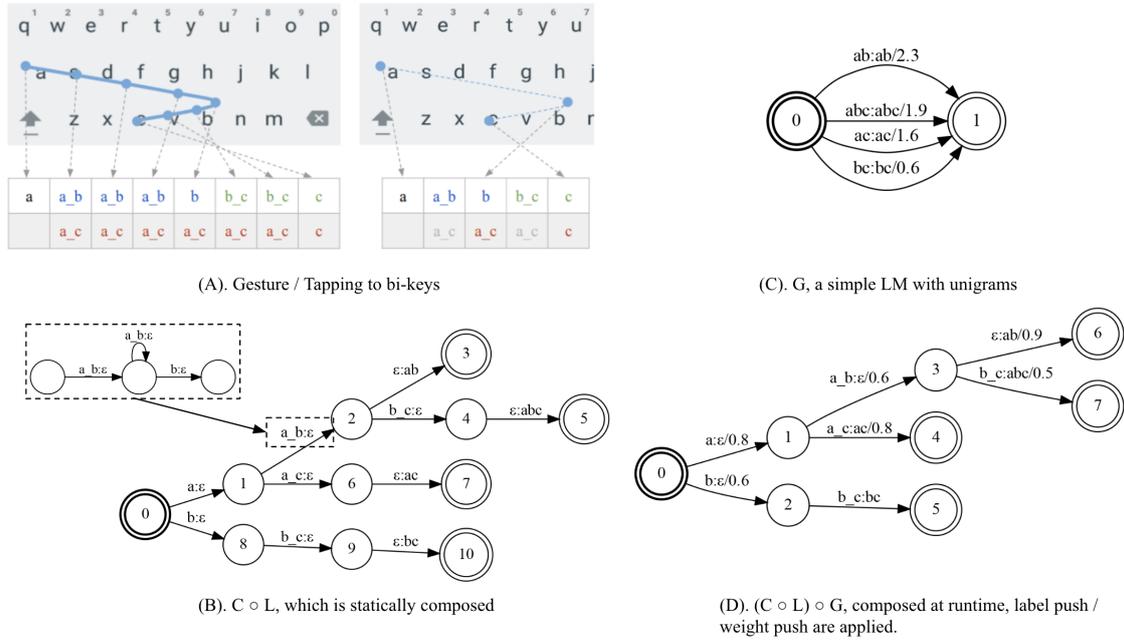


Figure 1: Build search space by composing $(C \circ L) \circ G$

changes of NSS in initializing and extending the search space at a high level, in which the codes in red are for Neural Search Space and the codes in blue are for N-gram Search Space.

Algorithm 1 Initialize Search Space

Input: C_{input} \triangleright What users have committed
Output: $State_{SS}$ \triangleright Initial states in search space
 $S_{C \circ L} \leftarrow 0$
if N-gram LM **then**
 $G \leftarrow G_{ngram}$
 $S_G \leftarrow FindState(C_{input}, G_{ngram})$
else if NN-LM **then**
 $G_{NLM} \leftarrow UpdateLM(NLM, C_{input})$
 $G \leftarrow G_{NLM}$
 $S_G \leftarrow 0$
end if
 $State_{SS} \leftarrow Compose(C \circ L, G, S_{C \circ L}, S_G)$

Algorithm 1 is called before users start to type a new word, for example, when users open the keyboard for an editor box or when users commit a word by tapping on space. Algorithm 2 is called when users are typing a word; for tap-typing, it will be called upon each key tap.

NSS has minor changes in both algorithms, In Algorithm 1, rather than finding a specific state in a static N-gram LM given context, $UpdateLM$ inserts the next words and corresponding probabilities given context at the start state of G_{NLM} as arcs, thus the start state is always 0. In Al-

gorithm 2, $ExtendLM$ additionally extends the G_{NLM} which aims to handle the multi-word problem discussed later.

Algorithm 2 Extend Search Space During typing

Input: $Bikey_{seq}$ \triangleright Bikeys of the typing word
Output: $State_{SS}$ \triangleright States during typing
 $NewState_{SS} \leftarrow \{\}$
if N-gram LM **then**
 $G \leftarrow G_{ngram}$
else if NN-LM **then**
 $G_{NLM} \leftarrow ExtendLM(NLM, C_{input})$
 $G \leftarrow G_{NLM}$
end if
for $S_{C \circ L}, S_G$ in $State_{SS}$ **do**
 $S \leftarrow Compose(C \circ L, G, S_{C \circ L}, S_G)$
 $NewState_{SS} \leftarrow NewState_{SS} + S$
end for
 $Prune(NewState_{SS}, Bikey_{seq})$
 $State_{SS} \leftarrow NewState_{SS}$

NSS presents three key challenges: handling out-of-vocabulary words (OOV) given NN-LM symbol table constraint, preventing search space exploding caused by assuming word separation at each touch frame, and controlling latency considering dynamic NN-LM inference and on-the-fly FST conversion. We address these through carefully generated FST structure design, accurate pruning strategies, and data structure optimizations.

We conducted extensive live experiments on US

English, British English, Spanish in Spain and the US, Portuguese in Portugal. Our key metrics are Words Modified Ratio (WMR), approximating word error rate by reporting the proportion of words modified by the user after their initial commit, and typing speed measured by Words Per Minute (WPM). Online experiment results demonstrated WMR improvement in [0.26%, 1.19%], at an acceptable level of latency increase [17%, 28%].

The contributions of this work can be summarized as follows:

- We propose Neural Search Space, integrating the long context representation ability of NN-LM into a carefully designed FST.
- We resolve practical problems such as OOV, word separation hypothesis, and latency problems through efficient FST structure design, accurate pruning strategies and data structure optimizations.
- We demonstrate the effectiveness of NSS under production environment over millions of users through live experiments, improving the user experience by reducing WMR and enhancing typing speed in a system optimized over decades.

2 Background

Recent advances of Neural Network LMs (NN-LM), notably projects such as GPT-4 (OpenAI, 2023), PaLM 2 (Anil et al., 2023), demonstrate their superior performance compared to N-gram LMs, particularly in capturing longer context.

Federated Learning (FL) (McMahan et al., 2017; Kairouz et al., 2021) with Differential Privacy (DP) (Dwork et al., 2006, 2014) enables Gboard to improve LM quality with user data while preserving user privacy by distributing model training across user devices instead of collecting data centrally. Prior work employed FL to train LMs for Next Word Prediction, Smart Compose, and On-The-Fly rescoring in Gboard following Hard et al. (2018); Xu et al. (2023). However, these applications either operate on first pass decoding results produced by N-gram LMs, or do not affect decoding suggestion which has the largest impact on typing experience.

To benefit from FL of NN-LM and retain decoding efficiency, previous research has explored projecting or approximating NN-LMs onto N-gram LMs (Chen et al., 2019; Suresh et al., 2019, 2021), and making the FST differentiable (Hannun et al., 2020). However, such conversions inevitably incur

losses due to limited context and back-off smoothing necessitated by sparsity (Chen and Goodman, 1999).

In this work, we replace the N-gram LM within the search space with an NN-LM trained via FL, enhancing long context capabilities. The deployed NN-LM is an LSTM / CFG model similar to those in Hard et al. (2018); Xu et al. (2023).

3 Challenges

Ideally, an NN-LM would score all known words for optimal coverage. However, vocabulary size is limited due to the high computational cost of the final dense layer. Our deployed NN-LM has a 30k-word vocabulary (top words from Federated Counting), while the full lexicon contains 170k words. Scoring the remaining 140k words in our generated FST poses a key challenge.

Missing the space key and mistyping it with the “cvbn” keys are the two common and consequential mistakes in mobile typing, turning multiple words into one single string (See Appendix A.1 for demo cases). Converting <word, probability> pairs to an FST for the current context would only provide NN-LM scores for the first word in such cases, with subsequent words receiving contextless unigram scores. This penalizes and perhaps suppresses multi-word candidates. We address this using dynamic inference in Section 4.3.

Gboard operates under strict latency constraints. Key presses should trigger visible feedback within 20ms as highlighted in Ouyang et al. (2017). NSS inevitably increases latency due to NN-LM inference and FST conversion. Dynamic inference, employed to address space substitution issues, significantly expands the search space by hypothesizing word separations at each frame, further exacerbates this challenge.

4 Methods

We detail the *UpdateLM* and *ExtendLM* described in Algorithm 1 and Algorithm 2 respectively below.

4.1 Algorithms

The pseudocode for *UpdateLM* and *ExtendLM* is provided in Algorithm 3 and Algorithm 4.

In *UpdateLM*, G_{NLM} is first set to the initial structure G_{base} (Fig. 2), either by direct reset in decoder initialization or via *ResetFST*. As G_{base} is as large as the full 170k-word vocabulary, and

the modified FST will have thousands of new states and arcs on top of that, in-place reset is more efficient than copy. We propose a more compact data structure for efficient reset in Section 4.4.3.

Algorithm 3 UpdateLM

Input: C_{input} \triangleright Committed words
Input: NLM \triangleright Neural Network LM
Output: G_{NLM} \triangleright Runtime generated FST
if $G_{NLM} = null$ **then**
 $G_{NLM} \leftarrow G_{base}$
else
 $ResetFST(G_{NLM})$
end if
 $S_{start} \leftarrow 0$
 $ModifyFST(G_{NLM}, C_{input}, S_{start}, NLM)$

Next, $ModifyFST$ inserts the NLM outputs into G_{NLM} as arcs attached on the start state 0 (Fig. 3).

Algorithm 4 ExtendLM

Input: C_{input} \triangleright Committed words
Input: NLM \triangleright Neural Network LM
Output: G_{NLM} \triangleright Runtime generated FST
 $S_{extend} \leftarrow FindStatesToExpand()$
 $DynamicInferencePruning(S_{extend})$
for S in S_{extend} **do**
 $W \leftarrow FindAdditionalContext(S)$
 $C_{extend} \leftarrow C_{inputs} + W$
 $ModifyFST(G_{NLM}, C_{extend}, S, NLM)$
end for

Similarly, $ExtendLM$ modifies G_{NLM} at other states chosen dynamically based on context and scores (discussed in Section 4.4.2). An example FST structure after $ExtendLM$ is shown in Fig. 4.

4.2 FST Structure

The initial structure of the FST in NSS is shown in Fig. 2. State 0 is the start state and state 1 is the unigram state. Unigrams are attached to the unigram state as arcs with format “word/weight”, where weight is the negative log probability. This example only has 5 unigrams. For clarity, we decouple the self-loop on the unigram state by duplicating the unigram state in the graph. Only one zero weight epsilon arc is attached to the start state before any modification.

Given new context, $UpdateLM$ inserts NLM outputs into G_{NLM} as arcs (Fig. 3). Three words

and weights are attached to the start state as arcs, each leading to a new state with an epsilon arc to the unigram state. The NN-LM contains fewer words than the total unigrams. The epsilon arc from the start state has the <UNK> probability from the NN-LM.

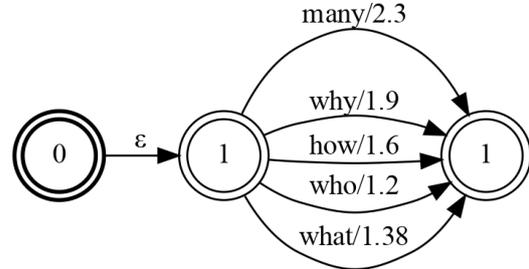


Figure 2: Initial FST Structure

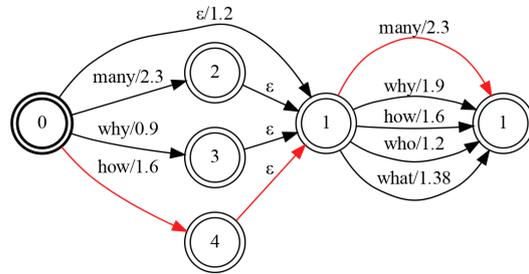


Figure 3: FST Structure after $UpdateLM$, three words are in the NN-LM vocab

The epsilon arc plays a key role in handling OOV: if a word is not in the vocabulary of the NN-LM, the search traverses the epsilon arc to the larger unigram state. Here “OOV” means words in the unigrams but not in the NN-LM; real OOV words are handled by literal decoding and dynamic models following Ouyang et al. (2017), which is the same for N-gram LMs as for NN-LM.

This structure also handles the words with space substitution errors: the first word of the contiguous multi-word candidate is scored by the NN-LM, and the rest receive unigram scores. For example, in Fig. 3, the path of “how many” from state 0 to state 4 to state 1 is highlighted in red.

To provide NN-LM scores for all words in contiguous multi-word candidates, we introduce Dynamic Inference below.

4.3 Dynamic Inference

To be able to provide NN-LM scores for all words in contiguous multi-word candidates, the FST structure is expanded dynamically based on the most likely target words users are typing. Inference will run on the concatenation of the base context and the

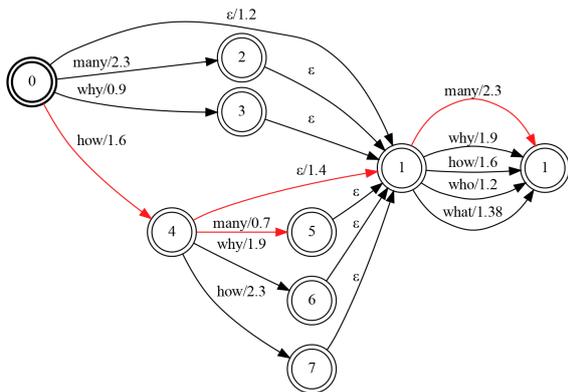


Figure 4: FST structure after one-time dynamic inference

possible target words, and the result probability distribution will be merged into the runtime generated G_{NLM} . This process is named Dynamic Inference, which is exactly the *ExtendLM* in Algorithm 4.

Fig. 4 illustrates an example of dynamic inference. Assuming *FindStatesToExpand()* returns state 4, then “how” is the target word, NN-LM inference is conducted on context + “how”, and the outputs are converted to the arcs attached to state 4, which is very similar to the operations on state 0. Dynamic inference will keep updating the FST at the newly-added states in a recursive manner.

After expansion, there are two paths for “how many”, $0 \rightarrow 4 \rightarrow 5$ and $0 \rightarrow 4 \rightarrow 1 \rightarrow 1$, the former path can provide pure NN-LM scores for the candidate while the latter still provides the mixed scores. The search phase will return the path with higher score, for this case, $0 \rightarrow 4 \rightarrow 5$ path will be returned.

Each state expansion necessitates both NN-LM inference and FST structure updates, leading to a substantial latency increase. Section 4.4.2 mitigates the number of expansions while Section 4.4.3 adapts the FST data structure to frequent modifications efficiently.

4.4 Latency Optimization

Various optimizations are explored to meet Gboard’s latency requirements. The most effective methods are listed below.

4.4.1 Arc Pruning

Each G_{NLM} modification involves inserting 30k $\langle \text{word}, \text{score} \rangle$ pairs. However, since lower scores are unlikely to survive in beam search phase, words with probabilities less than a fixed T_{arc} are omitted from the FST. T_{arc} is set to e^{-15} for *UpdateLM* and e^{-12} for *ExtendLM*. This generally keeps

only 1k to 5k words, which reduced the tap typing latency increment from +211.54% to +81.51% in offline evaluation.

Settings and detailed results of the offline evaluation can be seen in Appendix A.2 and Appendix A.3 respectively.

4.4.2 Dynamic Inference Pruning

Ideally, we should expand G_{NLM} at all states if possible, however, the time complexity is an unbearable $O(N^L)$, where N is the number of words and L is the length of the candidate.

Dynamic Inference Pruning is applied in Algorithm 4 to reduce the complexity. We adopt two rules simultaneously to decide whether a state should be expanded.

- Only states with scores larger than a threshold T_{extend} are eligible for expansion.
- Only the top N states with eligible scores may be expanded.

We explored various thresholds in offline evaluation, empirically choosing $T_{extend} = e^{-12}$ and $N = 1$, which increases tap typing latency by 79.92%.

4.4.3 Frequently Modified FST

The default mutable FST implementation we use is OpenFST (Allauzen et al., 2007), in which arcs are stored independently per state to offer flexibility to add and remove arcs and states. However, it’s inefficient when the FST is incrementally updated and frequently reset. Using reset as an example, we would need to delete the arcs of each state first and then delete the states, which is expensive.

Based on this requirement for incremental updates and frequent resets, we propose a customized FST implementation. The arcs of all states are stored in the same array, and the FST maintains a map from states to the indices of their corresponding arcs in the large array. When resetting the FST, we only need to clear the single array of arcs and then delete the map of states; the arc array can be reused instead of reallocated each time.

The FST structures are illustrated in figures in Appendix A.5.

5 Evaluation

We conducted live experiments on uniform random samples of the eligible Gboard populations (Sivek and Riley, 2022) for US English (en-US), GB English (en-GB), ES Spanish (es-ES), PT Portuguese (pt-PT) and US Spanish (es-US).

Language	Devices	WMR(%)	WPM(%)	Latency(%)	Total Users(M)
en-US	ALL	-0.26	+0.06	+24.13	14.00
es-ES	ALL	-0.77	+0.40	+23.13	3.30
pt-PT	ALL	-0.99	+0.59	+28.34	0.85
en-GB	ALL	-1.19	+0.40	+17.92	1.99
	3+ GB	-1.31	+0.48	+13.98	1.43
	6+ GB	-1.13	+0.30	+7.36	0.64
es-US	ALL	-1.03	+0.39	+17.43	14.37
	3+ GB	-1.24	+0.43	+15.14	8.86
	6+ GB	-1.24	+0.52	+12.89	1.46

Table 1: Live Experiment results for en-US, es-ES, pt-PT, en-GB and es-US.

5.1 A/B Metrics

Metrics in the A/B experiments to measure the quality and latency are:

- **Words Modified Ratio (WMR):** The ratio of words being modified during typing or after committed; improvement is shown by reduction.
- **Words Per Minute (WPM):** The number of committed words per minute.
- **Latency:** The average time for decoding.
- **Total Users:** The number of users participating in the experiments with the target languages.

5.2 Experiment Setup

There are two arms in the live experiments:

- **Control Arm:** the LM is a N-gram FST which is obtained by approximating the NN-LM trained via Federated Learning or counting from server corpus.
- **NSS Arm:** the LM is a simple FST generated by one-layer LSTM at runtime.

The NN-LM in live experiment is a one-layer LSTM model with the following configuration:

- *vocab_size* = 30k
- *embedding_dim* = 96
- *lstm_size* = 670
- *total_parameters* = 6.4M
- *training_loss*: cross entropy of next word prediction.

We hypothesize that the quality of NSS is limited by the latency. To verify this, we also report metrics restricted to high-end devices with memory larger than 3G / 6G for en-GB and es-US.

5.3 Result Analysis

The online live experiment results are listed in Table 1. All metrics other than Total Users are reported as percent changes relative to the control arm. All latency changes and all bolded WMR and WPM changes are significant at a 0.05 level (null hypothesis: the metric change is 0).

It’s observed in Table 1 that:

- The NSS arm reduces WMR with a 95% confidence interval of [0.26%, 1.19%] on various languages while increasing latency in the [17%, 28%] range.
- Higher-end devices exhibits marginally greater improvements in WMR and WPM with smaller latency increases. Specifically for example, for es-US, the latency increment on 6G+ devices is 12.89%, 5% less than the increment in ALL devices, while the WMR and WPM improvement are larger. The metrics of en-GB on 6G+ devices are not better than the other settings, which we argue is potentially caused by the small population attending the live experiments.
- WPM is consistently improved, suggesting that the latency cost does not adversely affect user experience.

From the perspective of production, as a well optimized system, WMR of Gboard decoder ranges within [4.5%, 7%] for locales studied in this paper, the relative improvement larger than 0.1% is deemed significant. Additionally, the observed latency increase fell below the sensitive threshold, allowing us to achieve both WMR/WPM optimization and latency control. Specifically, the Gboard decoder consists of a series of modules, with NSS positioned upstream. Any latency increase in NSS can potentially impact downstream modules like key correction and auto-corrections, negatively af-

fecting key metrics. Our experimental findings indicate that if the latency increase remains below a certain level, downstream modules are unaffected, and users do not perceive the latency change.

The confirmation on the hypothesis regarding higher-end devices empowers us to deploy more powerful models on devices with greater capabilities, further enhancing the user experience.

It's expected that the quality improvement on en-US is much smaller than the other locales. Due to the high product priority, the baseline of en-US is much stronger than other locales, which adopts the FST approximated from a FL trained LM (Chen et al., 2019) and a specialized N-gram model for the search domain.

6 Discussions

This work successfully bridged the gap between the long range context awareness power of NN-LMs and the efficiency requirements of Gboard's decoder. By creatively adapting NN-LMs to an FST structure and implementing latency optimizations, we deployed the Neural Search Space in production. Experiments demonstrated that NSS significantly improves decoding quality, particularly on higher-end devices, with an acceptable latency trade-off. This direct integration unlocks potential for future enhancements driven by NN-LM advancements, promising further gains in keyboard decoding and overall user experience.

Building upon NSS, we identify several promising directions for future research:

- **Integrating Transformer models:** Transformer models are known for their superior quality and training efficiency. Exploring their integration within NSS presents an exciting opportunity. However, a key challenge here is maintaining system performance, given the substantial computational resources and memory required by the Transformer model to handle long contexts. Further investigation is needed to assess the quality gains achievable with models constrained to fewer than 10 million parameters due to these system limitations.
- **Leveraging richer context:** Compared to traditional n-gram models, NN-LMs offer a more flexible framework for incorporating diverse contextual information. Integrating signals like application domain, country, time, and extended user history holds the potential to

further enhance model quality.

- **Exploring SentencePiece LMs:** The current NSS utilizes word-level LMs, which can be limited by OOV issues. Employing SentencePiece (Kudo, 2018) LMs could improve performance by providing better word coverage and a more nuanced representation of language.

Beyond enhancing NSS, another avenue for exploration is replacing the FST-based decoder with a neural decoder. While we have investigated this approach, certain challenges hinder its immediate adoption as a full replacement:

- **Quality Gap:** The current system, refined over a decade by numerous engineers, incorporates extensive prior knowledge about error patterns, which is difficult to encapsulate within a single neural model.
- **Debugging Challenges:** The current system allows for straightforward debugging and correction of errors by adjusting weights in FSTs. Transitioning to a purely neural decoder would sacrifice this flexibility..

However, we recognize the inherent advantages of neural models, such as superior semantic understanding and context capture. Therefore, we continue to experiment with end-to-end approaches. One promising avenue is to run neural models in parallel with the existing system and merge their candidate suggestions, leveraging the strengths of both FSTs and neural approaches. This hybrid approach allows us to benefit from the precision and debuggability of the FST-based system while capitalizing on the advanced contextual understanding of neural models.

References

- Cyril Allauzen, Michael Riley, and Johan Schalkwyk. 2009. A generalized composition algorithm for weighted finite-state transducers. In *Interspeech*, pages 1203–1206.
- Cyril Allauzen, Michael Riley, and Johan Schalkwyk. 2011. A filter-based algorithm for efficient composition of finite-state transducers. *International Journal of Foundations of Computer Science*, 22(08):1781–1795.
- Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. Openfst: A general and efficient weighted finite-state transducer library: (extended abstract of an invited talk). In *Implementation and Application of Automata: 12th International Conference, CIAA 2007, Prague, Czech*

- Republic, July 16-18, 2007, Revised Selected Papers 12*, pages 11–23. Springer.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Shiri Azenkot and Shumin Zhai. 2012. Touch behavior with different postures on soft smartphone keyboards. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pages 251–260.
- Mingqing Chen, Ananda Theertha Suresh, Rajiv Mathews, Adeline Wong, Cyril Allauzen, Françoise Beaufays, and Michael Riley. 2019. Federated learning of n-gram language models. *arXiv preprint arXiv:1910.03432*.
- Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.
- Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.
- Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.
- Awni Hannun, Vineel Pratap, Jacob Kahn, and Weining Hsu. 2020. Differentiable weighted finite-state transducers. *arXiv preprint arXiv:2010.01003*.
- Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.
- Lars Hellsten, Brian Roark, Prasoon Goyal, Cyril Allauzen, Françoise Beaufays, Tom Ouyang, Michael Riley, and David Rybach. 2017. Transliterated mobile keyboard input via weighted finite-state transducers. In *Proceedings of the 13th International Conference on Finite State Methods and Natural Language Processing (FSMNLP 2017)*, pages 10–19.
- Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210.
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguerre y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.
- R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.
- Tom Ouyang, David Rybach, Françoise Beaufays, and Michael Riley. 2017. Mobile keyboard input decoding with finite-state transducers. *arXiv preprint arXiv:1704.03987*.
- Gary Sivek and Michael Riley. 2022. Spatial model personalization in gboard. *Proceedings of the ACM on Human-Computer Interaction*, 6(MHCI):1–17.
- Ananda Theertha Suresh, Brian Roark, Michael Riley, and Vlad Schogol. 2019. Distilling weighted finite automata from arbitrary probabilistic models. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 87–97.
- Ananda Theertha Suresh, Brian Roark, Michael Riley, and Vlad Schogol. 2021. Approximating probabilistic models as weighted finite automata. *Computational Linguistics*, 47(2):221–254.
- Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher Choquette, Peter Kairouz, Brendan McMahan, Jesse Rosenstock, and Yuanbo Zhang. 2023. [industry] federated learning of gboard language models with differential privacy. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

A Appendix

A.1 Multi-word Demo

The two frequently multi-word typos described in Section 3 can be seen in Fig. 5.

A.2 Offline Evaluation Setting

TypingTester is the tool used for offline evaluation. It’s a testing framework for the Gboard decoder and related C++ code. It repeatedly runs the decoder over a sequence of touch points, compares the output text to expected text, and estimates metrics like word error rate, next word prediction accuracy, decode time, and more.

The dataset used for offline evaluation contains touch points for 2500 sentences, which is gathered from 50 volunteers by typing the same 50 sentences.

In this paper, typingtester is used to report the relative decoding latency change, the tests run on a workstation with 3.7Ghz, 6 core Intel CPU.

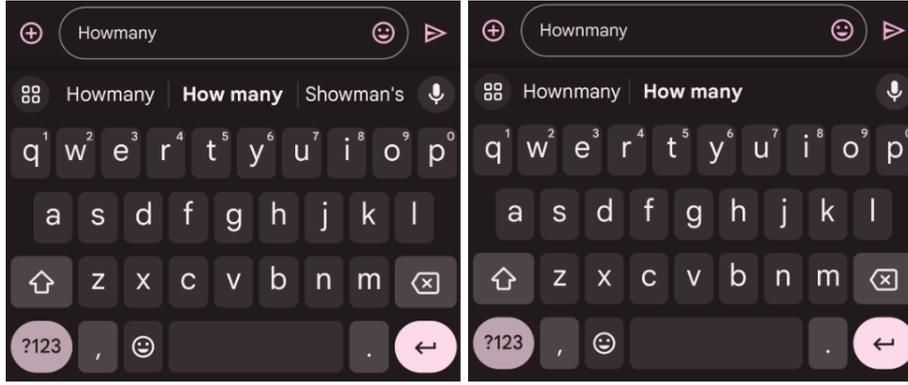


Figure 5: Word with Space Substitution Errors. Left: misses the space key. Right: mistypes the space with "n" ("cvbn")

A.3 Arc Pruning Latency Impact

The latency impact of arc pruning with different thresholds is shown in Table 2. We report the relative change of latency over the prod N-gram search space, the latencies for tap-typing and gesture typing are reported separately.

T_{Update}	T_{Extend}	Tap (%)	Gesture (%)
e^{-15}	e^{-12}	+81.51	-9.28
e^{-10}	e^{-10}	+57.96	-36.03
e^{-12}	e^{-12}	+79.85	-25.37
e^{-15}	e^{-15}	+138.36	-3.21
0	0	+211.54	+15.59

Table 2: Latency change on various arc threshold combinations

The chosen thresholds for Neural Search Space are e^{-15} for *UpdateLM* and e^{-12} for *ExtendLM*. The latency increase of tap typing is 81.51%, which is larger than the latency increment online due to the reasons listed below.

- The optimization gap between workstation and phone devices, eg: tflite inference is faster on device.
- Not all modules are involved in the offline evaluation.

The gesture latency is reduced by 9.28%, which benefits from the inherent property that gesture typing is free of the missing/mistyping space key problem, such that the dynamic inference defined in Algorithm 4 is not required.

No pruning happens when the thresholds are set to 0, and 130.03% and 24.87% latency savings on tap typing and gesture typing are observed respectively comparing to the chosen thresholds.

A.4 Dynamic Inference Pruning Latency Impact

Dynamic Inference pruning strategy contains two thresholds, N controls how many states could be expanded at most per char and T_{expand} controls whether a state is eligible to be expanded, which are set to be 1 and e^{-12} respectively after verifying on live experiments.

Table 3 displays the latency impact of various thresholds. Gesture latency is not affected significantly as it doesn't have dynamic inference. Tap latency is affected by the range from 10% to 20% when changing the thresholds. The latency increase is reduced from 79.92% to 40.86% if cancelling the dynamic inference, which defines the loose upper bound of latency increase in dynamic inference optimization.

N	T_{extend}	Tap (%)	Gesture (%)
1	e^{-12}	+79.92	-7.60
2	e^{-12}	+94.51	-8.23
3	e^{-12}	+99.31	-8.35
1	e^{-15}	+91.17	-6.19
1	e^{-10}	+69.21	-7.07
No Dynamic Inference		+40.86	-8.45

Table 3: Latency change on various dynamic inference thresholds

A.5 Customized FST structure

As described in Section 4.4.3, the default modified FST implementation in OpenFST is illustrated in Fig. 6. while the customized implementation of a frequently updated FST is illustrated in Fig. 7.

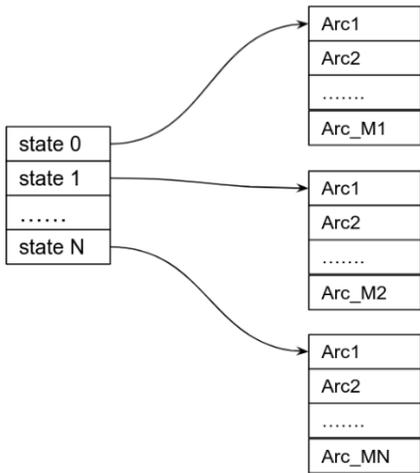


Figure 6: Default Modifiable FST

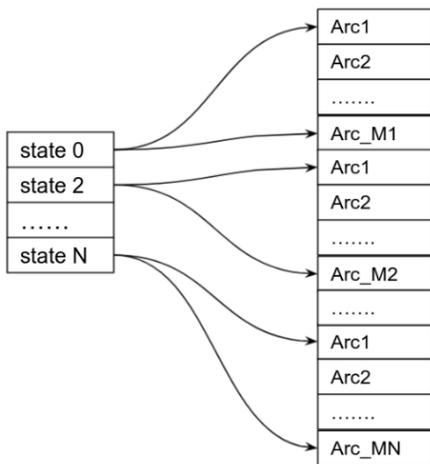


Figure 7: Optimized Modifiable FST in incremental update scenarios