



Shallow Fusion of Weighted Finite-State Transducer and Language Model for Text Normalization

Evelina Bakhturina*, Yang Zhang*, Boris Ginsburg

NVIDIA, Santa Clara, USA

{ebakhturina, yangzhang, bginsburg}@nvidia.com

Abstract

Text normalization (TN) systems in production are largely rule-based using weighted finite-state transducers (WFST). However, WFST-based systems struggle with ambiguous input when the normalized form is context-dependent. On the other hand, neural text normalization systems can take context into account but they suffer from unrecoverable errors and require labeled normalization datasets, which are hard to collect. We propose a new hybrid approach that combines the benefits of rule-based and neural systems. First, a non-deterministic WFST outputs all normalization candidates, and then a neural language model picks the best one – similar to shallow fusion for automatic speech recognition. While the WFST prevents unrecoverable errors, the language model resolves contextual ambiguity. We show for English that the approach is effective and easy to extend. It achieves comparable or better results than existing state-of-the-art TN models.

Index Terms: text normalization, WFST, LM re-scoring, text-to-speech

1. Introduction

Text normalization (TN) converts text from the canonical written to the spoken form, and it is an essential pre-processing step of text-to-speech (TTS) systems, see Figure 1. Sentences consist of plain and special words or phrases, also known as *semiotic tokens* [1]. Semiotic tokens can be grouped into classes like cardinals, fractions, or dates. Semiotic tokens require non-trivial normalization while plain words remain the same. In the example shown in Figure 1 “1/4” is the only semiotic token and thus needs to be normalized.

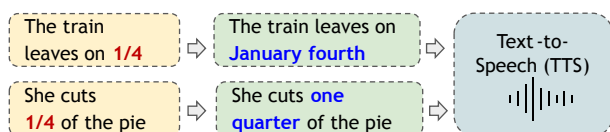


Figure 1: The normalization of the semiotic token “1/4” is context-dependent.

TN has a few characteristics that differentiate it from other natural language processing (NLP) tasks. First, the normalized texts are scarce and difficult to collect [2, 3]. To our knowledge, Google text normalization dataset [4] is the only sizeable public dataset. Second, TN applications such as mobile assistants demand very low error rates [5]. Errors are called *unrecoverable* if they alter the input semantic [5], e.g. “1/4” → “one fourteenth”. Information-sensitive industry

*Equal contribution. Order determined alphabetically.

sectors such as finance can not tolerate any chance of unrecoverable errors. Lastly, semiotic tokens make up only a small percentage of the sentence, but they stem from a huge variety of semiotic classes [5]. These make TN a challenging NLP task, and it remains an active area of research.

TN systems can be rule-based, neural or a combination of both. The traditional **rule-based** systems utilize either regular expressions or weighted finite-state transducers (WFST) [6] to define a set of language-specific rules. These systems are easy to extend, debug and reason about. However, they can not resolve contextual ambiguity well without making the rules significantly more complex. For example, it is very complicated to write a rule that resolves “1/4” to “one quarter” in one case and to “January fourth” in another context, see Figure 1. Additionally, rule-based systems can be easily customized to support Speech Synthesis Markup Language (SSML) [7], which is becoming a requirement for TTS systems in production. SSML gives users control over different verbalization options, and users can explicitly specify if “2” should be spoken as the cardinal “two” or the ordinal “second”. Recent **neural** systems treat TN as a machine translation task using Seq2Seq models [4, 8, 9]. These are good at incorporating context, but they require large datasets to scale, and suffer from catastrophic unrecoverable errors. Therefore, some **hybrid** systems propose to use a generative neural network (NN) constrained by WFST grammars [5, 10]. Nonetheless, if the WFST does not cover a particular example, the NN performs unconstrained normalization. These hybrid systems can reduce the number of unrecoverable errors, but still require supervised training data [5]. Therefore, state-of-the-art TN systems in production rely mostly on rule-based normalization [3, 6].

We propose a new hybrid TN approach that uses rule-based WFST grammars to output all possible normalization options, which are then re-scored with a pre-trained neural language model (LM) to resolve contextual ambiguities, see Figure 2. This approach has the following benefits:

- Ability to control system’s output and prevent certain unrecoverable errors, like hallucinations.
- Easy to improve normalization capability by extending the WFST grammar.
- Compatible with many off-the-shelf pre-trained LMs. Also, LMs can be easily adapted as they are trained in a self-supervised fashion without the need for labeled data, i.e. written-spoken text pairs.

The proposed WFST+LM approach is similar to LM-decoding in automatic speech recognition (ASR) or machine translation [11]. Instead of a single transcription, ASR outputs multiple options, and then a LM model re-scores these options to account for language fluency and context. To the best of our knowledge, we are the first to apply shallow fusion of

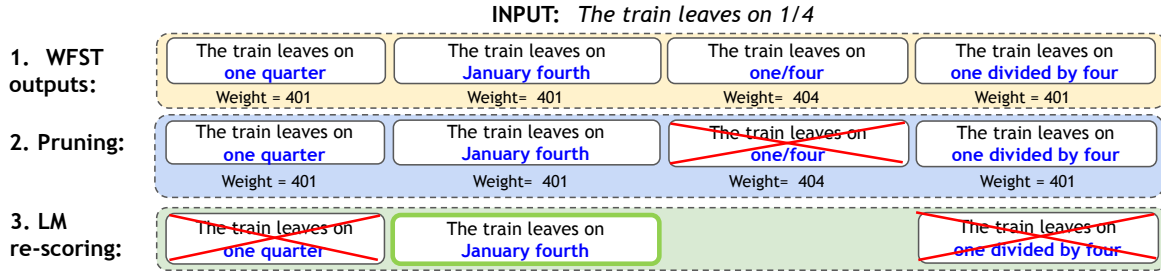


Figure 2: WFST+LM shallow fusion pipeline. 1. WFST generates all possible normalization forms and assigns weights to each option. 2. Pruning normalization options with weights above threshold “401.2”, in this example we drop “one/four”. It has higher weight since it was not fully normalized. 3. LM re-scoring picks the best among the remaining options.

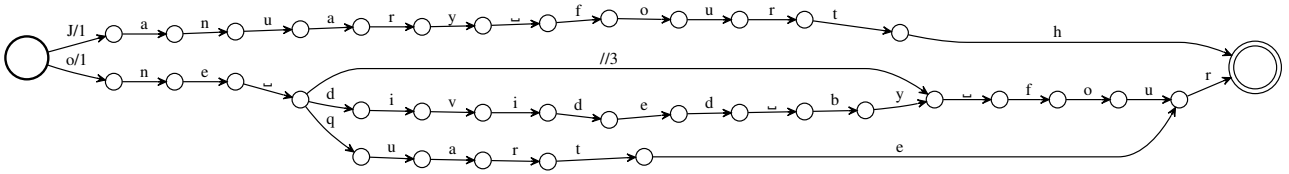


Figure 3: WFST lattice of normalization options for “1/4”. The options shown are “January fourth”(W=1), “one quarter”(W=1), “one divided by four”(W=1), “one/four”(W=4). By construction, if a semiotic token is not fully normalized, e.g. “/” in “one/four”, it has a higher weight, e.g. 4 vs. 1. This allows us to filter out these options by applying a threshold on the weights.

WFST and LM to TN. We created a hand-crafted set of context-dependent normalization test cases to demonstrate the effectiveness of our approach compared to rule-based and NN-based TN systems. The model and the set of the context-dependent test cases will be released at <https://github.com/NVIDIA/NeMo>.

2. WFST-based TN re-scoring with LM

2.1. Deterministic WFST

A path through a weighted finite-state transducer (WFST) [6] maps an input sequence to an output sequence, where each transduction, called arcs, has an associated weight w , see Figure 3. A path’s weight W is the sum of all its arcs’ weights, and a deterministic WFST outputs a single shortest path which has the minimum weight. As a starting point for our proposed system we use a deterministic WFST based on the NeMo text normalization framework [12], which does not take context into account (hereafter DetWFST). In case of ambiguous inputs such as “1/4” in “The trains leaves on 1/4”, a deterministic WFST can mistakenly choose “one quarter” over “January fourth”.

2.2. Non-deterministic WFST

We extend aforementioned DetWFST to output multiple valid normalization options, see Figure 2. We eliminate redundancy of semantically equivalent normalization forms, e.g., “January fourth” and “the fourth of January”. Beyond that, we want every semiotic token normalized, so WFST assigns higher weights to a path that keeps the token unchanged. Specifically, we picked a weight of $w=100$ for unchanged word and $w=2$ for a punctuation mark. On the other hand, all arcs that normalize a semiotic token have weights $w \in [1.0; 1.01]$. This allows us to prune all unwanted outputs by enforcing a threshold on their respective paths’ weights, see Figures 2, 3.

Let us assume our input has k semiotic tokens, and S is

the set of possible normalization options, where $w_{s,k}$ is the weight of the k -th token of the normalization option s . Then, all weights W_s for $s \in S$ satisfy the following condition:

$$W_s = C + \sum_k w_{s,k} \leq C + k \cdot 1.01$$

$$= C + k \cdot 1.0 + k \cdot 0.01$$

where C is the total weight of all non-semiotic tokens, and we assume that k is at most 20. Further,

$$C + k \cdot 1.0 + k \cdot 0.01 \leq \min_{s \in S} W_s + k \cdot 0.01$$

$$\leq \min_{s \in S} W_s + 20 \cdot 0.01$$

$$= \min_{s \in S} W_s + 0.2$$

Thus, we are keeping only outputs with weights within 0.2 of the shortest path.

2.3. Language model re-scoring

The pruned WFST output options are re-scored with pre-trained language models from the Hugging Face library [13]. For autoregressive language models, like GPT [14], we rank the normalization options based on their perplexity. For masked language models we use Masked Language Model (MLM) scoring¹ [15]. MLM is the average of pseudo-log-likelihood scores calculated by masking input tokens one by one.

We modify MLM score calculation when an input sentence contains multiple semiotic tokens that require normalization. In this case, to minimize bias, we exclude all but one semiotic token from the context, which LM considers. The goal is to have only a single semiotic option when calculating the MLM score,

¹<https://github.com/awslabs/mlm-scoring>

while the rest of the semiotic tokens in question are masked. The scores for each masked variant are averaged to get an aggregated value for the sentence. For example, to calculate the score for “What’s < one half > cup plus < two thirds > cup?”, where the angular brackets highlight semiotic spans, we first calculate weights for “What’s [MASK] cup plus two thirds cup?” and “What’s one half cup plus [MASK] cup?”. Then, we take the average to get the score for the sentence of interest.

3. Experiment setup

3.1. Evaluation datasets

We use pairs of written and normalized English texts for evaluation. Publicly available “general” datasets are large and suit as a benchmark to compare our results with related works. We additionally created a “confusing” dataset, which focuses on ambiguous examples.

Google text normalization dataset (hereafter GoogleTN) [4] is one of the few publicly available TN datasets for English with decent coverage of various semiotic classes and is widely used in the field. Normalization forms in this dataset are generated semi-automatically with Google TTS system’s Kestrel text normalization system [3]. The test set contains 7551 sentences. For our second “general” dataset, we use LibriTTS [16] derived from LibriVox audiobooks² and corresponding Gutenberg texts³. We only evaluate the subset of LibriTTS where the normalization differs from the written form, resulting in 7677 sentences.

For the “confusing” dataset (hereafter EngConf) we collected context-dependent normalization examples from an internal dataset, Google TN dataset (disjoint from aforementioned) and a conversational corpus [17]. Since the Google TN data was labeled semi-automatically there are many errors. These were manually corrected before being added to EngConf.

3.2. Models

We compare our WFST+LM approach with the deterministic WFST (DetWFST) [12] and the end-to-end neural TN model (Duplex) [9]. Both baselines achieve one of the the highest reported normalization accuracies on GoogleTN dataset [4]. Duplex comprises a distilled RoBERTa-based token classification model [18] and a generative T5-base model [19]. Every input token the classifier identifies as semiotic token is normalized by the T5 model. In our hybrid approach we used the pre-trained BERT-based-uncased [20] checkpoint from Hugging face [13], unless stated otherwise.

3.3. Metric

For quantitative analysis we calculate the sentence-level accuracy from the ground truth normalization and the model output [4]. We relax the strict string equality requirement by accepting different formats of the same phrase as equivalent. For example, “the fifth of November two thousand twenty” and “November fifth twenty” are both valid spoken forms of the date “2020/11/05” and are considered equivalent.

²<https://librivox.org/>

³<https://www.gutenberg.org/>

Table 1: Comparison of the model performance of DetWFST, Duplex and WFST+LM on three datasets (sentence accuracy, %). The EngConf contains confusing “one-to-many” test cases, the other datasets cover general TN cases.

Dataset	Number of sentences	Det WFST	Duplex	WFST + LM
EngConf	231	68.83	55.41	94.37
GoogleTN	7551	97.29	99.07	97.79
LibriTTS	7677	98.65	90.40	99.01

Table 2: Sentence accuracies on EngConf dataset using different language models for LM re-scoring.

Model	accuracy, %
distilgpt2 [21]	82.68
openai-gpt [22]	88.31
bert-base-en-uncased (MLM) [20]	94.37
bert-base-en-cased (MLM) [20]	92.64
roberta-base-en-cased (MLM) [18]	93.07

4. Results

4.1. Overall performance

Table 1 summarizes the performance results of WFST+LM, DetWFST and Duplex on the datasets EngConf, GoogleTN, and LibriTTS. For WFST+LM, we found masked language models with MLM-scoring [15] work better for than autoregressive models, e.g. GPT [14] and DistilGPT2 [21], see Table 2. BERT-base-uncased [20] results in the highest sentence accuracy.

The WFST+LM beats Duplex model on all but GoogleTN data. Duplex, which is a neural model trained on 800,000 sentences of the training split of the Google text normalization dataset, reaches an accuracy of 99.07% compared to 97.29% of DetWFST and 97.79% of WSFT+LM. To our knowledge, the accuracy of DetWFST is the highest among publicly available rule-based systems on GoogleTN and is only slightly behind the neural model. Both DetWFST and WFST+LM lack some rules for exceptional input formats, which account for some of the performance gap, e.g., “Set the thermostat to 75F” in Table 3. However, Duplex fails to generalize on LibriTTS and EngConf datasets. WFST+LM outperforms Duplex with 99.01% vs. 90.40% and 94.37% vs. 55.41% sentence accuracy, respectively. Duplex suffers from unknown input formats and domain shifts. For example, it does not recognize “1/4” as a date in “The train leaves on 1/4” and falsely produces “The train leaves on one quarter” instead. Duplex can also make mistakes when normalizing cardinals containing more than four digits, e.g. “10001” → “one hundred thousand one”. For information-sensitive industries such as finance, this type of unrecoverable error is detrimental. More failure types can be found in Table 3.

WFST+LM outperforms DetWFST on all three datasets. On the general domain datasets, LibriTTS and GoogleTN, the LM re-scoring does not provide much performance gain over the deterministic WFST. These datasets do not contain many confusing cases and thus can be mostly normalized using deterministic grammars. However, LM re-scoring offers a noticeable improvement over DetWFST on the EngConf dataset. Language model integration resolves ambiguities from context and achieves 94.37% vs. 68.83% sentence accuracy. One such

Table 3: Error patterns for DetWFST, Duplex and WFST+LM.

Error Type	Det WFST	Duplex	WFST +LM	Example errors	
Number error	not affected	affected	not affected	IN: TARGET: Duplex:	Number 10001 Number Ten thousand one Number one hundred thousand one
Unknown format	affected	affected	affected	IN: TARGET: DetWFST: Duplex: WFST+LM:	Set the thermostat to 75F Set the thermostat to seventy five degrees Fahrenheit Set the thermostat to seventy five F Set the thermostat to seventy five degrees Set the thermostat to seventy five F
Hallucination	not affected	affected	not affected	IN: TARGET: Duplex:	Josiah in the gutter! exclaimed Mrs. Pegler. Josiah in the gutter! exclaimed Misses Pegler. Josiah in the gutter! exclaimed m r e Pegler.
Omission	not affected	affected	not affected	IN: TARGET: Duplex:	Cambridgeshire, CB 10 1 SD Cambridgeshire c b one zero one s d Cambridgeshire c b one one s d
Class ambiguity	affected	less affected	less affected	IN: TARGET: DetWFST: Duplex:	The train leaves on 1/4 The train leaves on January fourth The train leaves on one quarter The train leaves on one quarter
Smart URL splitting	affected	less affected	less affected	IN: TARGET: DetWFST: Duplex: WFST+LM:	It's WeAreSC.com It's We Are S C dot com It's W e A r e s c dot com It's We Are esc dot com It's we are sc dot com

context-dependent example is “The train leaves on 1/4” shown in Figure 2.

Similar to shallow fusion [11] for speech recognition or neural machine translation, LM re-scoring can make mistakes especially if the LM was trained on a different domain. In our case, publicly available pre-trained LMs are trained on un-normalized data but used to score normalized text. For example, the LM struggles to differentiate between “born in one thousand nine hundred seventy” vs. “born in nineteen seventy” as a spoken form of “born in 1970” and “Henry III” can be mistakenly normalized to “Henry three” instead of “Henry the third”. However, while the failures of the re-scoring approach can be expected by design of the WFST output options, it is not the case for the neural model. Failing examples for each of the methods are listed in Table 3.

Neural models such as Duplex can make unexpected errors and are hard to control and debug. At the same time, the WFST-based solutions can fail to cover some cases due to a limited set of grammar rules, but they are reliable and easy to debug if a rule is present.

4.2. Additional use cases of re-scoring

LM re-scoring can help choose the right inflection form for morphology rich languages. For example, in German, ordinals change their suffix depending on the following noun.

Email and URL normalization can also benefit from the proposed approach. Without an extra magnitude of complexity, deterministic WFSTs can only convert URLs into space-separated characters. For example, “myemail@greattech.com” becomes “m y e m a i l a t g r e a t t e c h d o t c o m”. However, URLs often contain words which should be pronounced as a whole, e.g. “my email at great tech dot com”. LM re-scoring can up-vote this option easily. Duplex can sometimes

identify words in a URL correctly but it is prone to hallucinations or omissions especially when encountering unseen words, see Table 3.

5. Conclusion

We propose shallow fusion of WFST and neural language model for text normalization. The non-deterministic WFST outputs multiple normalization options for ambiguous inputs, which are re-scored by a pre-trained language model. WFST+LM provides contextual disambiguation while preventing unrecoverable errors such as hallucination and token omission that are common for neural text normalization approaches. The WFST+LM system is easy to debug and extend, and it allows users to remain in control. We find that masked language models outperform generative models in our experiments. This approach reaches an accuracy of 94.37% on an ambiguous context-dependent dataset, which outperforms a state-of-the-art neural model, and it reaches a competitive 97.79% on the Google text normalization dataset. WFST+LM achieves 99.01% accuracy on the LibriTTS corpus, showing its resilience to domain shift.

6. Acknowledgements

The authors would like to thank Vitaly Lavrukhin and Elena Rastorgueva for their review and feedback.

7. References

- [1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.
- [2] R. Sproat, “Lightly supervised learning of text normalization: Russian number names,” in *IEEE Spoken Language Technology Workshop*, 2010.
- [3] P. Ebden and R. Sproat, “The kestrel tts text normalization system,” *Natural Language Engineering*, 2015.
- [4] R. Sproat and N. Jaitly, “An RNN model of text normalization,” in *Interspeech*, 2017.
- [5] H. Zhang, R. Sproat, A. H. Ng, F. Stahlberg, X. Peng, K. Gorman, and B. Roark, “Neural models of text normalization for speech applications,” *Computational Linguistics*, 2019.
- [6] M. Mohri, *Weighted Automata Algorithms*, 2009, pp. 213–254.
- [7] “Speech Synthesis Markup Language (SSML),” <https://www.w3.org/TR/speech-synthesis>.
- [8] C. Mansfield, M. Sun, Y. Liu, A. Gandhe, and B. Hoffmeister, “Neural text normalization with subword units,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, 2019.
- [9] T. M. Lai, Y. Zhang, E. Bakhturina, B. Ginsburg, and H. Ji, “A unified transformer-based framework for duplex text normalization,” *arXiv preprint arXiv:2108.09889*, 2021.
- [10] M. Sunkara, C. Shivade, S. Bodapati, and K. Kirchhoff, “Neural inverse text normalization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
- [11] Ç. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, and Y. Bengio, “On using monolingual corpora in neural machine translation,” *CoRR*, vol. abs/1503.03535, 2015.
- [12] Y. Zhang, E. Bakhturina, and B. Ginsburg, “NeMo (Inverse) Text Normalization: From Development to Production,” in *Interspeech*, 2021.
- [13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Oct. 2020, pp. 38–45.
- [14] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018.
- [15] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, “Masked language model scoring,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- [16] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “Libritts: A corpus derived from librispeech for text-to-speech,” in *Interspeech*, 2019.
- [17] X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser, “Benchmarking natural language understanding services for building conversational agents,” in *IWSDS*, 2019.
- [18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, 2020.
- [20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.
- [21] “DistilGPT2,” <https://huggingface.co/distilgpt2>.
- [22] “OpenAI GPT,” https://huggingface.co/docs/transformers/model_doc/openai-gpt.