Interspeech 2022
18-22 September 2022, Incheon, Korea

Improving Data Driven Inverse Text Normalization using Data Augmentation and Machine Translation

Debjyoti Paul, Yutong Pang, Szu-Jui Chen, Xuedong Zhang

Meta Platforms, Inc.

debjyotipaul@fb.com, yutongpang@fb.com, srayc@fb.com, xuedong@fb.com

Abstract

Inverse text normalization (ITN) is used to convert the spoken form output of an automatic speech recognition (ASR) system to a written form. Traditional handcrafted ITN rules can be complex to transcribe and maintain. Meanwhile neural modeling approaches require quality large-scale spoken-written pair examples in the same or similar domain as the ASR system (in-domain data), to train. Both these approaches require costly and complex annotation. In this paper, we present a data augmentation technique with neural machine translation that effectively generates rich spoken-written pairs for high and low resource languages effectively. We empirically demonstrate that ITN models (in target language) trained using our data augmentation with machine translation technique can achieve similar performance as ITN models (en) trained directly with in-domain language.

Index Terms: inverse text normalization, data augmentation, data-driven modeling, speech recognition, machine translation

1. Introduction

Inverse Text Normalization (ITN) is used to convert spoken form output from an automatic speech recognition (ASR) system to the corresponding written form. ITN can be challenging since multiple different spoken forms can express identical written expressions. For example, both twenty twenty (the year) and two thousand twenty (numeric) can be transcribed to '2020'. Conversely, the same spoken form can be transcribed to two or more different written expressions depending on the context. For example, twenty twenty can be transcribed to 2020 (for the year), to 20/20 (to denote eye vision), or to 20:20 (to represent time). Such a many-to-many mapping between spoken and written forms and dependence on context makes ITN an interesting and challenging problem in speech recognition.

There has been a renewed interest in the deep learning community to explore data-driven approaches to ITN. A popular ITN approach is to use a set of simple hand-written rules together with a neural model that can statistically learn how and when to apply these rules [1–4]. But creating rules for individual languages are tedious and time consuming. We explore the techniques to distillate knowledge from large natural language models to train multi-language ITN models with minimal human interventions.

Spoken form	Written form
do you like nineties music	do you like 90s music
let's meet at three thirty	let's meet at 3:30
three thirty kilos	330 kilos
he is at thirty percent of his goal	he is at 30% of his goal

Table 1: *Examples of Spoken-Written pairs.*

To generate training data for ITN models, a common approach is to use a text normalization (TN) system [5]. However, since the TN system only outputs one flawless spoken form per written input, it does not cover the variations of spoken forms that can be generated from a single written form. If we train a model with the over simplified spoken-written pairs, the model usually over-fits to the TN system, reflecting high accuracy for the curated entities, but the model can struggle to generalize to real-world use cases.

Hence, we make the following contributions in this paper:

- We propose a text normalization method for English that transforms written-form texts to spoken-form texts. Unlike conventional text normalization system, our data augmentation system generates more possible variants of spoken forms; which can help make robust ITN system.
- With the recent development and improvements of neural machine translation, we propose to use a knowledge distillation approach for internationalization of the ITN models. We apply neural machine translation on English spoken-written text pairs to generate spoken-written pairs on target languages; and it helps ITN expanding to more languages.

2. Data Driven ITN

2.1. Data Augmentation

Considering that people can speak a particular written form in multiple ways in the real world- for example, *5.0* could be verbalized as *five point zero*, *five point o*, *five dot zero*, etc. - we have developed a specialized data augmentation method for data-driven ITN modeling.

Unlike a conventional written to spoken TN system, our ITN augmentation system is capable of generating diversified spoken forms by introducing almost all possible spoken variations to the written forms as shown in Table 2; it produces 22x factor more variations than conventional TN.

Our augmentation system performs a series of steps, such as, (i) extraction of ITN entities (ii) reformat and clean ITN entities, (iii) apply augmentation with rewrite rules (iv) rewrite the input sentence producing spoken forms; step-by-step procedure depicted in Figure 1. Table 2 present a few examples of input and outputs from the data augmentation system.

Written Text Input	Spoken Text from Conventional TN	Spoken Text from Data Augmentation System
$123	one hundred twenty three dollars	one hundred twenty three dollars one hundred twenty three dollar one twenty three dollars one twenty three dollar one hundred and twenty three dollars one hundred and twenty three dollar one two three dollars one two three dollar
6:15 am	six fifteen a m	six thirty a m six fifteen in the morning six fifteen six past fifteen a m quarter past six a m quarter past six morning six past quarter morning

Table 2: *Examples of generated spoken form using conventional TN system and our data augmentation system*

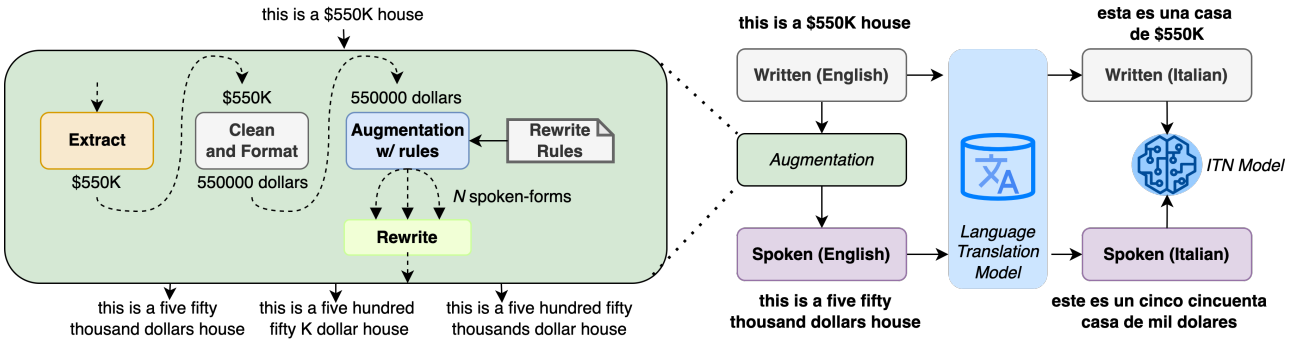

Figure 1: *Data augmentation with machine translation system for data-driven ITN modeling on target languages (e.g., Italian)*

2.2. Augmentation with Machine Translation

Recent advances in natural language translation with large-scale neural machine translation (NMT) models has empowered us to distillate ITN knowledge from such models and use it for training ITN models supporting more languages. Spoken-written text pairs in English generated from data augmentation system is translated to respective target language with NMT models as shown in Figure 1 and Table 3. We used multiple open-sourced NMT models and a home-grown NMT model to evaluate our approach. One of the challenge using NMT models for translation is unpredictable text normalization behavior, such as, conversion of spoken form source language to written text in target language and vice-versa. For example, *I have thirty dollars* $\rightarrow$ *Ho 30 dollari* (in Italian); i.e., *thirty* $\rightarrow$ *30* conversion is unintended. Moreover, NMT models are still far from generating perfect translation and errors from translation models can propagate to ITN models. We try to eliminate these drawbacks by filtering out translated spoken-written pairs where text normalization forms are not intact and choosing NMT models with reasonable BLEU score on target languages over ITN texts. With both the data augmentation system and translation pipeline in place, we can generate a significant number of spoken-written pairs synthetically for training high and low resource languages.

Form	Text in English	Translated text
spoken	Historical average for January is thirty one degrees.	La moyenne historique de janvier est de trente et un degrés [Fr] La media storica di gennaio è di trentuno gradi. [It] La media histórica de enero es de treinta y un grados. [Es]
written	Historical average for January is 31 degrees.	La moyenne historique pour janvier est de 31 degrés. [Fr] La media storica di gennaio è di 31 gradi. [It] La media histórica de enero es de 31 grados. [Es]

Table 3: *Examples of data augmentation with machine translation models for [Fr]ench, [It]alian, Spanish [Es].*

2.3. Model and Evaluation

We trained a single Seq2seq bidirectional-LSTM architecture model with our generated dataset for four languages i.e., English, French, Italian, Spanish. We use an ITN focused human annotated dataset in English, and use NMT models to translate the spoken form to the target language, and then apply the trained ITN model (in target language) on the spoken form to get the written form in target language, in this way, we just need to verify the digit in the written form is the same as in the original English dataset. This is an end2end way we proposed to measure the NMT and ITN model quality; explained in Figure 2.

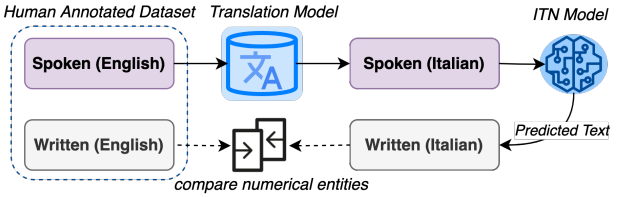

Figure 2: *ITN model evaluation strategy for languages where human annotation spoken-written pairs are not available.*

3. Experiments

From the experiment result with human-annotated dataset, we find models trained with synthetic data generated with NMT augmentation performed equally well on other languages, such as, French, Italian and Spanish for numerical entities.

Language	Augmentation Accuracy %
English	76.8
French	76.3
Italy	79.0
Spanish	79.3

Table 4: *Accuracy performance comparison of ITN model with augmentation vs baseline.*

4. Conclusion

In this demo paper, we introduce a robust data augmentation methodology for ITN that can generate rich and variant spoken-written pairs from out-of-domain textual (written) data, and then use machine translation to scale the data driven ITN to more language domains. We also proposed a new end2end way to evaluate the model in foreign language when lack of evaluation data. We empirically demonstrate that our technique significantly improves ITN model and shows that this methodology can be particularly helpful for ITN models expanding to more languages with translation where both training and evaluation data is not readily available.

5. References

[1] M. Ihori, A. Takashima, and R. Masumura, "Large-context pointer-generator networks for spoken-to-written style conversion," in *ICASSP*, 2020.

[2] S. Pramanik and A. Hussain, "Text normalization using memory augmented neural networks," *Speech Communication*, 2019.

[3] M. Shugrina, "Formatting time-aligned asr transcripts for readability," in *HLT*, 2010.

[4] H. Sak, Y.-h. Sung, F. Beaufays, and C. Allauzen, "Written-domain language modeling for automatic speech recognition," 2013.

[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.