LANGUAGE AGNOSTIC DATA-DRIVEN INVERSE TEXT NORMALIZATION

Szu-Jui Chen[1][*][†], Debjyoti Paul[2][†],
Yutong Pang[2], Peng Su[2], Xuedong Zhang[2]

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, USA[1],
Meta AI, USA[2]

ABSTRACT

With the emergence of automatic speech recognition (ASR) models, converting the spoken form text (from ASR) to the written form is in urgent need. This inverse text normalization (ITN) problem attracts the attention of researchers from various fields. Recently, several works show that data-driven ITN methods can output high-quality written form text. Due to the scarcity of labeled spoken-written datasets, the studies on non-English data-driven ITN are quite limited. In this work, we propose a language-agnostic data-driven ITN framework to fill this gap. Specifically, we leverage the data augmentation in conjunction with neural machine translated data for low resource languages. Moreover, we design an evaluation method for language agnostic ITN model when only English data is available. Our empirical evaluation shows this language agnostic modeling approach is effective for low resource languages while preserving the performance for high resource languages.

Index Terms— Inverse text normalization, Multilingual, Language-agnostic

1. INTRODUCTION

Recent progress in automatic speech recognition (ASR) technologies brings world-wide adoption to use voice as a source of input to interact and communicate with digital environments. Moreover, ASR systems are now expanding language horizons to bring same or better experience to native language speaker around the world. For communication and better readability of spoken form output from an ASR system, it needs to be paired with an inverse text normalization (ITN) system to produce corresponding written form texts.

Converting spoken form texts to written form is not an easy task. For example, the spoken form *twenty twenty* can be written as (a) *2020* to express the year/number, (b) *20:20* for time, (c) *20/20* for eye-vision, (d) 20-20 for a score in game or a cricket match. Taking another example of *ten to twelve*, which can be written as (a) *10-12* as a cardinal number range, (b) 10:00-12:00 as time range, (c) *11:50 am/pm* as a time instance, etc. It is evident from these examples that the

speaker contexts could help in determining the correct written form which is hard to abide by a rule-based system. Therefore data-driven based systems are gaining attention of researchers [1–4] to improve ITN systems, dub as DD-ITN models.

Dataset of spoken-written text pairs are used to train these neural network based DD-ITN models. Obtaining spoken-written text pairs that covers diverse ITN entities such as cardinals, ordinals, date-time, money, fractions, decimals, address, metrics, email, URL, abbreviation etc. is hard but doable for high-resource languages such as English. In addition, the written representation of same entity can vary across languages, e.g., 3:30 pm is commonly represented as 15h30 in French. The challenge grows multi-fold while expanding to more languages, especially low-resource ones where it is hard to obtain spoken-written text pairs dataset.

In this work, we take a step towards internationalization (i18n) and expanding language horizon for ITN models by alleviating data scarcity problem and presenting an unified language agnostic ITN model for supporting multiple languages. The main contributions of this work are as follows:

- 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 build robust ITN system.

- We propose to apply neural machine translation for internationalization of the ITN models, which can be considered as a knowledge distillation approach. We use 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.

- We present a language agnostic data-driven ITN model that supports inverse normalization of spoken form texts for 12 languages. We also present a study of system design choices in our experiment section.
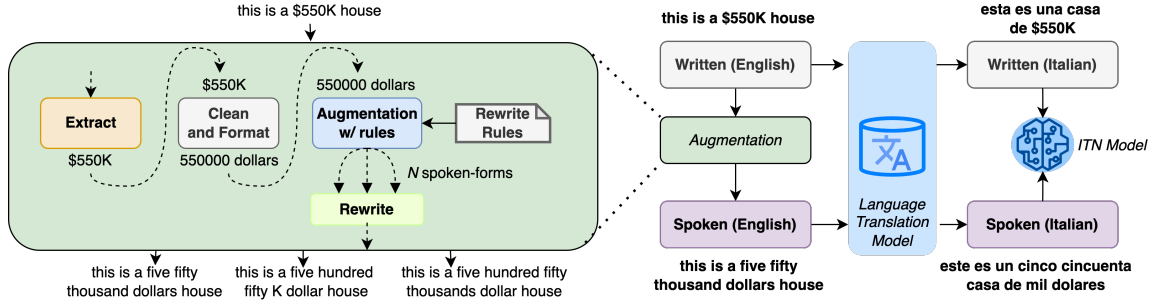
2. METHODOLOGY

2.1. Data Augmentation

To train a language agnostic ITN system, we need spoken-written text pairs for every language. However, there are not much public available resources having spoken-written text

---

Fig. 1. Multilingual data generation using rule based text normalization system and machine translation model.

pairs in English, not to mention other languages. Hence, we propose a two-step data augmentation method to generate spoken-written text pairs for multiple languages. First, we generate spoken form texts from highly available written text resources in English. Then, we use machine translation to generate text pairs for other languages.

2.1.1. Enhanced Text Normalization

Traditional text normalization (TN) used in text-to-speech (TTS) systems generates fixed variation of spoken forms (often with rule-based approach) that conforms to the verbalizer standard for any given written text with TN-ITN entities. As mentioned previously, spoken forms lack full information about the subject; hence, it is important to cover more spoken variations and alternatives. With abundant written text resources in hand, we apply our specialized text normalization augmentation system that generates almost all possible spoken variations. Table 1 shows an example of conventional TN output and augmented TN results.

Table 1. Examples of generated spoken form using conventional TN system and our data augmentation system

Written Text Input	Spoken Text from Conventional TN	Spoken Text from Enhanced TN System
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

2.1.2. Multilingual spoken-written text pairs

For languages we do not have adequate spoken-written text pairs, we propose to use the neural machine translation (NMT) model to generate the spoken-written text pairs in

Table 2. Examples of data augmentation with machine translation models for French [*fr*], Italian [*it*], Spanish [*es*].

| 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*]

target language from English pairs. We notice that the outputs of NMT can not meet our criteria; (a) spoken/written text forms mismatch during translation (b) non-conformity in source and target language; e.g. $801 \rightarrow$ eight o one [*en*] $\neq$ otto o uno [*it*], etc. Hence, we apply multiple approaches to filter out malformed/wrong translations. We limit our discussions on filtration process due to space limitations; though we provide translation accuracy in Table 6 as a reference. We present a pictorial representation of our data augmentation pipeline in Fig. 1 and provide some examples in Table 2.

2.2. Model Architecture

For our ITN task, it can be seen as a sequence-to-sequence (Seq2Seq) problem, which turns a sequence in one domain (e.g., spoken domain) to a sequence in another domain (e.g., written domain). For this Seq2Seq problem, we employ the Encoder-Decoder architecture (Fig. 2) to solve it. More specifically, two types of Encoder-Decoder model are investigated in this work: the LSTM-based Seq2Seq model and the Transformer model.

With the sequence data, a natural idea is to use the recurrent neural network (RNN), and Long Short-Term Memory (LSTM) [5] is the first choice among Recurrent Neural Networks (RNNs). In our LSTM-based Seq2Seq model, we utilize the LSTM in both the encoder and decoder. In order to encode the input sequence better, we employ bi-directional encoder to consider both the left context and right context in the sequence. Also, in the decoder side, we use attention mechanism to derive a context vector that captures the relevant source-side (encoder-side) information to help prediction. Initially, this type of Seq2Seq model is used for translation task. For more details about the model, we refer readers to paper [6].

For the Transformer model, we employ the original model from [7]. Its encoder and decoder are composed of stacked
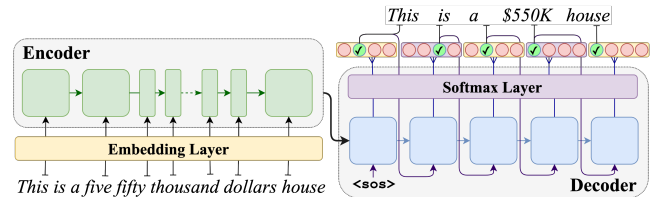
Fig. 2. Encoder-Decoder model architecture for ITN.

Table 3. Model Parameters.

	LSTM	Transformer
No. of Parameter	19.98M	19.67M
Encoder Layer	2	4
Decoder Layer	2	4
Hidden Size	256	256
Attention Head	n/a	8

modules (or layers), and each module mainly consists of multi-head attention and feed forward networks. The attention mechanism will take the whole sequence into account by learning weights for input tokens in the encoder. In the decoder, the masked attention mechanism is applied to predict the next token based on the previous tokens.

For fair comparison, we experiment with both models with similar parameter sizes. The details of the model parameter can be found in Table 3.

3. EXPERIMENTS

We select 12 languages based on their richness of resources and their writing script, as shown in Table 4. For example, we see Russian and Kazakh share the same Cyrillic script where the latter is a low-resource language. For the model input/output text tokenization, we use SentencePiece tokenizer model (SPM) [8] with a vocabulary size of 20,000. Our ablation study with varying SPM vocabulary sizes (omitted due to space constraints) shows little performance improvement beyond that. To note, unless specified in experiment results, LSTM based Seq2Seq architecture is our default ITN model as described in Section 2.2. As choice of NMT models for our experiments, we use (a) Opus-MT [9] with EasyNMT library support, (b) MetaNMT, an in-house version of NMT with comparable performance to NLLB [10]. Also, unless specified, we use MetaNMT for data augmentation as it has better BLEU scores (see respective papers) and ITN performance impact (see Table 8).

Table 4. 12 languages selected for experiments.

Resources	Latin Script	Non-Latin Script
High	Italian [*it*], French [*fr*], Spanish [*es*], English [*en*], Turkish [*tr*], German [*de*]	Russian [*ru*], Greek [*el*]
Low	Icelandic [*is*], Afrikaans [*af*]	Tamil [*ta*], Kazakh [*kk*]

3.1. Dataset

We use the OpenSubtitles [11] and TED2020 [12] datasets from OPUS[1] as our training data. To be specific, only the written English texts from OPUS are used in our method. The spoken-written pairs for training are generated using the pipeline in Fig. 1.

We evaluate our models on two datasets: (a) Dictation testset: Human annotated 6,810 spoken-written conversational text pairs in English containing diverse ITN entities in mixed proportions. We apply the Case B strategy described in Fig 3 and Section 3.2 to generate evaluation data for non-English languages. (b) Caption testset: Mostly containing mathematical expressions, measures, metrics, phone numbers collected from *audios with uploaded caption*. This dataset
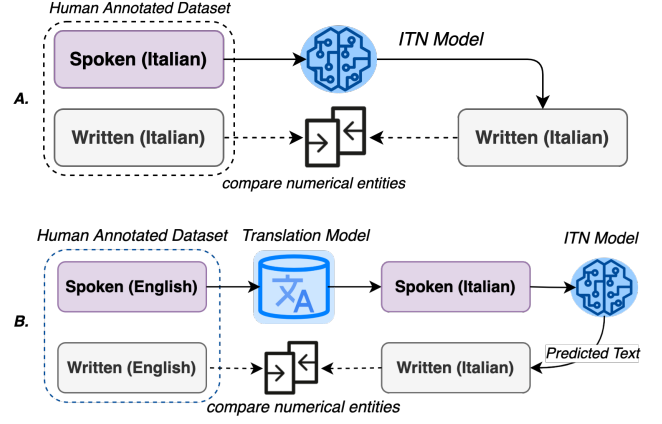
Fig. 3. ITN evaluation strategy with case A: Target language evaluation dataset is available; Case B: Target language evaluation dataset is not available. We use translation model to prepare spoken target texts. In both cases, we compare numerical entities output from the ITN model.

has [*en*]:22332, [*es*]:21216, [*fr*]:27300, [*it*]:14939, [*de*]:5960 spoken-written text pairs for respective languages.

3.2. ITN Evaluation

For the evaluation of our ITN model, we mainly focus on the numerical entities in the text. The evaluation is straightforward when ground-truth target language spoken-written pairs are available. But the lacking of proper human annotated spoken-written pairs for mid/low resource languages motivates us to propose an approach to measure the model performance on the numerical entities shown in Fig. 3.

3.2.1. ITN normalized accuracy

In Fig. 3, Case A shows the evaluation approach when target language spoken-written pair dataset is available. We apply Case B where target language pair dataset is not available, we translate the spoken form text from English human annotated dataset to target language. We then apply the trained ITN model to obtain written form text in target language. Thereafter, we verify the digit in the written form is the same as the original one in English. If they are the same, we count it as a correct instance, if not, we count it as a incorrect one. To note, translation models may output written form in target language for spoken source input, we discard them and only evaluate the ITN model on correctly translated spoken form texts. We use normalized accuracy to measure the performance of our ITN model with Eq. 1.

$$Normalized\ Accuracy = \frac{correct\ count}{correct\ count + error\ count}. \quad (1)$$

Table 5 shows ITN entities in **bold**, the correct spoken and written form translated entities are in **blue** and errors are colored with **red** with correct form in **(parenthesis)**. When comparing written form English and written form in target language, we take care of special cases; for example see the following examples:

- **12/24 hour conversion**: Use of 24 hours system in French and other target language; *1:30 p.m. [en]* vs. *13h30 [fr]*.

Table 5. Examples of errors for ITN evaluation. The second row is an example of ITN error while the third row is an example of NMT error.

Spoken	Written
[*en*] I found out that **nine** out of **ten** statistics are wrong.	[*en*] I found out that **9** out of **10** statistics are wrong.
[*fr*] J'ai découvert que **neuf** statistiques sur **dix** sont fausses.	[*fr*] J'ai récemment découvert que **neuf** (**9**) statistiques sur **10** sont fausses
[*en*] Dad's surprise **sixtieth** is on this Saturday. Arrive before **six PM**.	[*en*] Dad's surprise **60th** is on this Saturday. Arrive before **6 PM**.
[*fr*] La **soixantième** surprise de papa a lieu ce samedi. Arrivée avant **18h** (**dix-huit heures**).	[*fr*] La **60ème** surprise de papa a lieu ce samedi. Arrivée avant **18h**.

- **Accommodating Zeros**: Use alternate magnitudes; e.g., *24,000 [en]* vs. *24 mille [fr]*.

- **Small cardinal**: Disambiguate use of spoken/written form for small cardinals (1-9) based on target language preferences; e.g., *two children [en]* vs. *2 enfants [fr]*.

- **Separators**: Some languages like French use comma(,) for decimal point. Also, there are languages that use 2 or 3 digit separators with either comma, space or period; e.g., *25,000.00 [en]* vs. *25 000,00 [fr]* vs. *25.000,00 [de]*.

3.2.2. ITN entity translation accuracy

In order to evaluate the NMT model translation accuracy on ITN entities, we propose the ITN entity translation accuracy. Here we compare whether the numerical entities from translations produces consistent spoken/written form output w.r.t. input text form. An example is provided in the last row of Table 5. Furthermore, we apply back-translation strategy with the same/other NMT and compare it with original text ensuring quality dataset from translation.

3.3. Results and Analysis

We show the performance of monolingual baseline models and multilingual models on the Dictation testset in Table 6. Since the multilingual models have the same size with the monolingual models, these results are showing the impact of training data. For the high-resource languages, the performance gain of multilingual model is quite limited, and even worse performance is obtained (e.g., [*es*]). On the other hand, we observe that the multilingual models significantly improve the performance on low-resource languages (except Tamil [*ta*] since it is using a different script). For example, the normalized accuracy of monolingual model on Kazakh [*kk*] is only 0.03%, and the multilingual models could achieve the performance of 37.69%. Thus, we could say that adding training data from the same script can improve the model performance on low-resource languages. As a reference, we also provide the ITN entities translation accuracy in the last column of Table 6. The higher the accuracy, the more ITN entities are evaluated in the Dictation testset.

Similarly, Table 7 shows the performance of monolingual baseline models and the 12-language model on the Caption Testset. This is computed by the case A strategy in Fig. 3. The 12-language ITN model is able to perform better accuracy on [*en, es, de*] while maintain competitive accuracy on [*fr, it*].

Table 6. Normalized accuracy of monolingual and multilingual models on the Dictation testset. The lang., trans., and acc. stands for languages, translation, and accuracy.

Lang.	Monolingual	3-lang.	6-lang.	12-lang.	ITN trans. acc.
es	**79.15%**	78.09%	76.80%	75.17%	91.34%
fr	62.35%	**62.99%**	60.98%	60.07%	62.81%
it	70.71%	**71.42%**	69.96%	69.87%	76.02%
en	71.73%	-	**72.75%**	71.96%	-
ru	**68.39%**	-	64.66%	66.33%	82.86%
kk[†]	0.03%	-	**37.69%**	32.41%	99.63%
tr	**60.07%**	-	-	53.95%	46.19%
de	**68.24%**	-	-	63.74%	61.67%
el	**66.84%**	-	-	65.29%	64.64%
is[†]	48.50%	-	-	**61.75%**	99.36%
af[†]	29.21%	-	-	**50.51%**	96.45%
ta[†]	25.63%	-	-	**27.30%**	99.74%

[†] low resource languages

Table 7. Normalized accuracy of monolingual and 12-language model on the Caption testset.

Language	*en*	*es*	*fr*	*it*	*de*
Monolingual	63.70%	64.51%	**55.24%**	**57.57%**	48.10%
12-language	**64.74%**	**65.58%**	54.90%	56.77%	**50.19%**

From these two tables, we can see that the normalized accuracy is much higher on the Dictation testset than on the Caption Testset. We find that there is a domain mismatch in training and testset, as we find a lot of complex math expressions in the Caption testset (e.g., $2 + 4 + 2 + 7 + 4 = 19$) while none is found in the open subtitles training set.

Finally, Table 8 shows the ITN normalized accuracy when using a different model architecture and different machine translation model [9]. From the table, we can see that the LSTM-based Seq2Seq model is better than the Transformer model, and MetaNMT is better than the Opus-MT for our task.

Table 8. Normalized accuracy when comparing of architecture and translation tools on 3-langs ([*es*], [*fr*], [*it*]) model and SPM token size of 20,000 on the Dictation testset.

Arch.	NMT	*es*	*fr*	*it*
Seq2Seq	MetaNMT	**78.09%**	**62.99%**	**71.42%**
Seq2Seq	Opus-MT	71.11%	60.03%	55.89%
Transformer	MetaNMT	72.55%	57.27%	64.76%

4. CONCLUSION

This work investigates the effectiveness of language agnostic data-driven ITN model. With the same model size, a 12-languages ITN model can significantly improves the normalized accuracy of low resource languages while maintain reasonable performance for high resource languages. Also, we explore the architecture and the machine translation model used in our framework and found that Seq2Seq model with MetaNMT is our best system. Our future work will include designing better architecture for on-device deployment, evaluation with clean and native language dataset, adapt data augmentation to support native language nuances/variation, and expanding to more languages.

5. REFERENCES

[1] Ernest Pusateri, Bharat Ram Ambati, Elizabeth Brooks, Ondrej Platek, Donald McAllaster, and Venki Nagesha, "A mostly data-driven approach to inverse text normalization.," in *INTERSPEECH*. Stockholm, 2017, pp. 2784–2788.

[2] Tuan Manh Lai, Yang Zhang, Evelina Bakhturina, Boris Ginsburg, and Heng Ji, "A unified transformer-based framework for duplex text normalization," *arXiv preprint arXiv:2108.09889*, 2021.

[3] Monica Sunkara, Chaitanya Shivade, Sravan Bodapati, and Katrin Kirchhoff, "Neural inverse text normalization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7573–7577.

[4] Alexandra Antonova, Evelina Bakhturina, and Boris Ginsburg, "Thutmose tagger: Single-pass neural model for inverse text normalization," *arXiv preprint arXiv:2208.00064*, 2022.

[5] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] Minh-Thang Luong, Hieu Pham, and Christopher D Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] Taku Kudo and John Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[9] Jörg Tiedemann and Santhosh Thottingal, "OPUS-MT — Building open translation services for the World," in *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

[10] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang, "No language left behind: Scaling human-centered machine translation," 2022.

[11] Pierre Lison and Jörg Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," 2016.

[12] Nils Reimers and Iryna Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 11 2020, Association for Computational Linguistics.