Language Agnostic Data-Driven Inverse Text Normalization

Szu-Jui Chen[1*], Debjyoti Paul[2*], Yutong Pang[2], Peng Su[2], Xuedong Zhang[2]
szujui.chen@utdallas.edu[1], debjyotipaul, yutongpang @meta.com[2]
Meta AI and University of Texas at Dallas
CRSS

INTERSPEECH 2023

Abstract

We propose a language-agnostic data-driven ITN framework using data augmentation and neural machine translation for real-time miniature models and low-resource languages. Our approach addresses the lack of labeled spoken-written datasets for non-English languages. Empirical evaluation confirms the effectiveness of our approach for multiple non-English languages, even when using only English data.

Motivations

- ↪ The same spoken phrase can have multiple written forms depending on the context.
- ↪ Difficulty in obtaining pairs covering diverse ITN entities like cardinals, ordinals, date-time, money, fractions, decimals, address, metrics, email, URL, and abbreviations.
- ↪ Variations in written-form of the same entity across languages, e.g., 3:30 pm represented as 15h30 in French.

Italian	un quarto → $\frac{1}{4}$ or 1:15 cinquecento dollari → $500
French	quatre-vingt six → 86 dix-huit trente → 1830 or 18:30
Spanish	unocento por ciento → 100% veinte veinte → 2020 or 20:20
German	zweihundertzweiundzwanzig → 2022 viertel vor zwanzig → 19:45 or $\frac{1}{4}$ vor 20

Table 1. Tricky ITN examples for a few language.

Objectives

- ↪ We propose an augmented text normalization (TN) method for English that transforms written form texts to spoken form texts, generating more spoken-variants.
- ↪ We propose using neural machine translation (NMT) for internationalizing ITN models.
- ↪ We present a language-agnostic data-driven ITN model for inverse normalization of spoken form texts in 12 languages. Additionally, we study system design choices in our experiment section.

Enhanced Text Normalization

- ↪ Traditional TN systems generate fixed variations of spoken forms using rule-based approaches.
- ↪ Spoken forms by traditional methods may lack full information about the subject.
- ↪ We developed **enhanced TN system** for English that generates a wide range of spoken forms for various entities.

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 enhanced TN system.

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 French [*fr*], Italian [*it*], Spanish [*es*]

Methodology

We propose using NMT models to generate spoken-written text pairs in target languages for which we do not have adequate pairs.

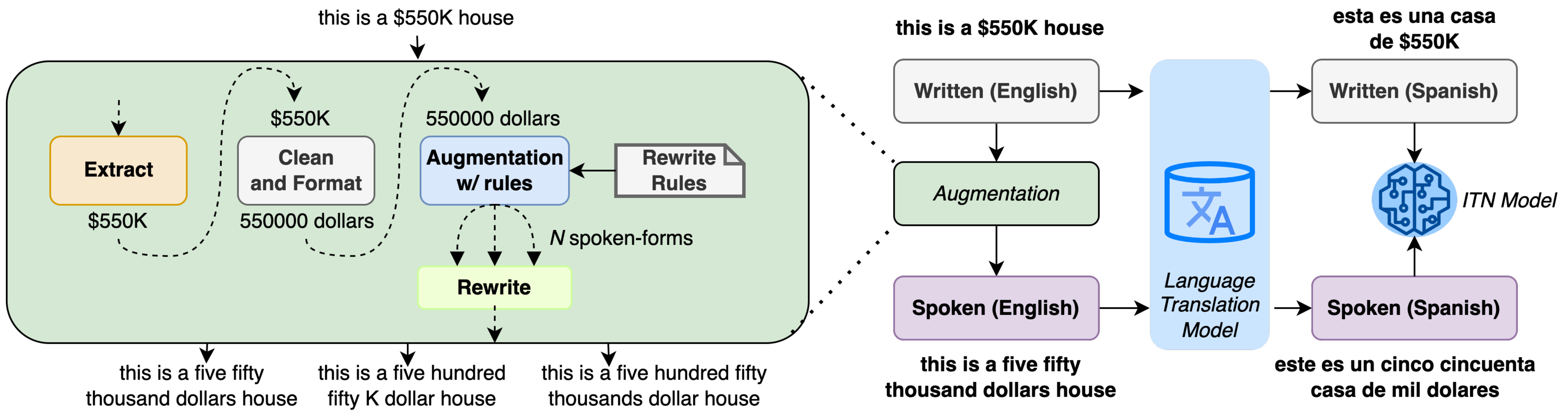

Fig 1. Multilingual data generation using enhanced rule-based text normalization system and machine translation model.

Model Architecture

Two types of Encoder-Decoder model are investigated in this work: the LSTM-based Seq2Seq model and the Transformer model.

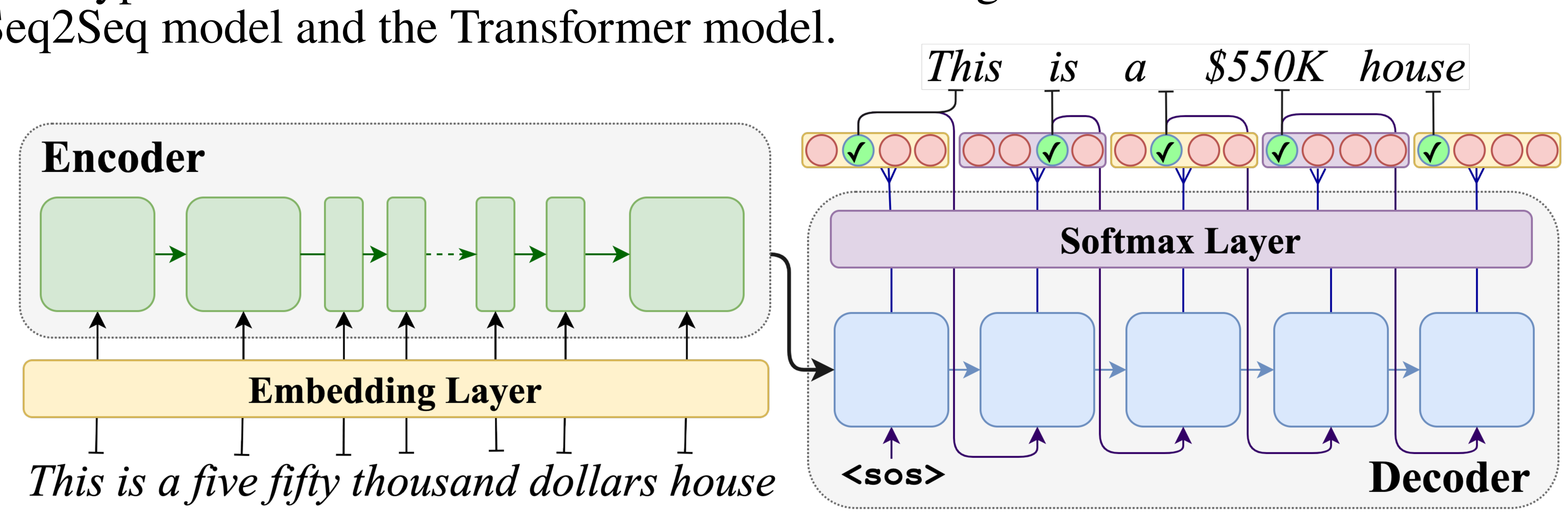

Fig 2. Encoder-Decoder model architecture for ITN.

Experiment Results

Evaluated our models on two in-house datasets.

- ↪ (a) Dictation test set: Human annotated 6,810 spoken-written conversational text pairs in English containing diverse ITN entities in mixed proportions.
- ↪ (b) Caption test set: Consisting of mathematical expressions, measures, metrics, phone-numbers. This dataset has [*en*]:22332, [*es*]:21216, [*fr*]:27300, [*it*]:14939, [*de*]:5960 spoken-written text pairs for respective languages.

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

Table 4. 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.

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%**

Table 5. Normalized accuracy of monolingual and 12-language model on the Caption testset.

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

Table 6. Normalized accuracy with architecture and translation tools on 3-langs ([*es*], [*fr*], [*it*]) model and SPM token size of 20,000 on the Dictation testset.

Language	Monolingual	3-languages	6-languages	12-languages	ITN entity translation accuracy.
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 multilingual models on the Dictation testset.

Conclusions

- ↪ A single 12-language model can substantially improve the normalized accuracy of low-resource languages while maintaining good performance for high-resource languages.
- ↪ Adding training data from the same script can improve the model performance on low-resource languages.

* **Authors with equal contributions.**