# Language Agnostic Data-Driven Inverse Text Normalization

*Anonymous submission to INTERSPEECH 2023*

## Abstract

The rise of automatic speech recognition (ASR) models has created an urgent need for converting spoken language into written text to provide better user experiences. This has drawn the attention of researchers, particularly for real-time on-device ASR deployment, towards the inverse text normalization (ITN) problem. While data-driven ITN methods have shown great promise in recent studies, the lack of labeled spoken-written datasets is hindering the development for non-English data-driven ITN.

To bridge this gap, we propose a language-agnostic data-driven ITN framework that leverages data augmentation and neural machine translation specifically designed for real-time miniature models and low-resource languages. Additionally, we have developed an evaluation method for language-agnostic ITN models when only English data is available. Our empirical evaluation attests to the efficacy of this language-agnostic ITN modeling with data augmentation approach for multiple non-English languages.

**Index Terms**: Inverse text normalization, Multilingual, Language-agnostic

## 1. Introduction

The latest advancements in automatic speech recognition (ASR) technologies have made it possible to use voice as an input source for interacting and communicating with digital environments, leading to widespread adoption worldwide. Furthermore, ASR systems are now extending their language capabilities to provide a similar or superior experience to native speakers across the globe. To facilitate communication and improve the legibility of the spoken output from an ASR system, it is necessary to pair it with an inverse text normalization (ITN) system, which produces corresponding written texts for a more seamless user experience.

Converting spoken text to written text is not a straightforward task, as the same spoken phrase can have multiple written forms depending on the context. For instance, the phrase "twenty twenty" could be written as (a) *2020* to denote the year or number, (b) *20:20* for time, (c) *20/20* for eyesight/vision, (d) *20-20* for a score in game or a cricket match. Similarly, *ten to twelve* 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. These examples demonstrate that context plays a crucial role in determining the appropriate written form, which is difficult to achieve using a rule-based system. Therefore, data-driven approaches are gaining traction among researchers [1–4] to improve ITN systems, also known as data-driven ITN (DD-ITN) models.

Spoken-written text pairs are used to train neural network based DD-ITN models. However, obtaining pairs that cover diverse ITN entities, such as cardinals, ordinals, date-time, money, fractions, decimals, address, metrics, email, URL, and abbreviations, can be difficult, especially in low-resource languages. Moreover, written representations of the same entity may vary across languages, such as 3:30 pm being 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.

We have observed that separating the ITN module from the end-to-end ASR system can result in independent performance improvements, utilizing a relatively large text-only dataset. Our on-device ASR systems have achieved a substantial decrease in word-error-rate (WER) by over 10%, a reduction in real-time-factor (RTF) by over 8%, and a reduction in memory usage. As a result, we propose using the ITN as a low-latency post-processing service. In this study, our aim is to address the problem of data scarcity and expand the language capabilities of ITN models by internationalizing (i18n) and introducing a unified, language-agnostic ITN model that supports multiple languages. The main contributions of our work are as follows:

- We propose a text normalization (TN) method for English that transforms written form texts to spoken form texts. Unlike conventional TN system, our augmented TN system generates more possible variants of spoken forms; which can help build robust ITN system.
- We propose to apply neural machine translation (NMT) for internationalization of the ITN models, which can be considered as a knowledge distillation approach. We use NMT 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

In order to train a language agnostic ITN system, spoken-written text pairs are required for each language. However, publicly available resources containing spoken-written text pairs are scarce, particularly in languages other than English. Therefore, we propose a two-step data augmentation method to generate spoken-written text pairs for multiple languages. In the first step, we use highly available written text resources in English to generate spoken form texts. Next, we employ machine translation in the second step to produce text pairs for other languages. We will explain each part of the system concisely due to space limitations.

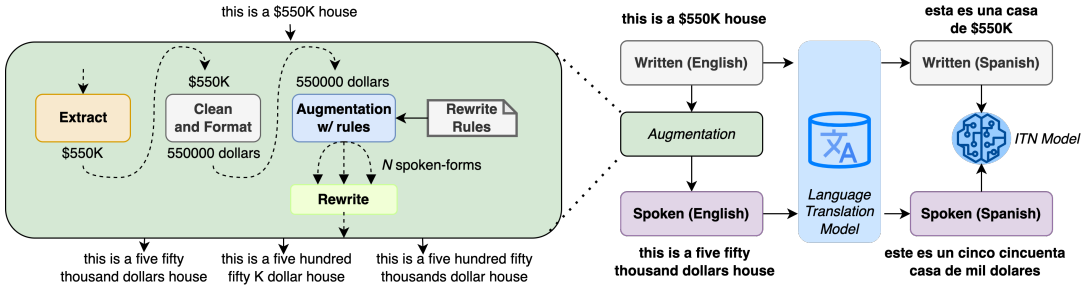


Figure 1: *Multilingual data generation using enhanced rule-based text normalization system and machine translation model.*

Table 1: *Examples of generated spoken form using conventional TN system and our enhanced TN system.*

| Written Text Input | Spoken Text from Conventional TN | Spoken Text from Enhanced TN System |
|---|---|---|
| 6:15 am | six fifteen a m | six fifteen a m<br>six fifteen in the morning<br>six fifteen<br>six past fifteen a m<br>quarter past six a m<br>quarter past six morning<br>six and quarter a m |
| $1.20 | one dollar and twenty cents | one dollar and twenty cents<br>one dollar twenty cents<br>one dollar two zero cents<br>one point two zero dollars<br>a dollar twenty cents |
| 123g | one hundred twenty three grams | one hundred twenty three grams<br>one hundred twenty three gram<br>one twenty three grams<br>one twenty three gram<br>one hundred and twenty three grams<br>one hundred and twenty three gram<br>one two three grams |

### 2.1.1. Enhanced Text Normalization

In text-to-speech (TTS) systems, traditional TN system generates a fixed variation of spoken forms, often using a rule-based approach, that conforms to the verbalizer standard for any given written text with TN-ITN entities. However, these spoken forms may lack full information about the subject, which highlights the importance of covering more spoken variations and alternatives. To address this, we developed an enhanced TN system for English that generates a diverse range of spoken forms, covering almost all possible variations for various entities, as demonstrated in Table 1. Our enhanced rule-based TN system supports various entities, including cardinal, ordinal, decimal, fraction, measures, money, date, time, and telephone entities. The system performs a series of steps for each written text input, which are illustrated in Figure 1 and described below.

a. *Pick:* From the text corpus pick up sentences with numerical ITN entities and discard other sentences.

b. *Categorize and Segment:* Extract text chunks that matches our carefully prepared regular expressions for each type of ITN entities in a predefined order. First, time and date entities segments are collected, followed by measures, currency, abbreviations. Then we find fractions, decimals, ordinals, phone numbers, and cardinal in respective order.

c. *Verbalize:* Entities matching text segment are then cleaned, formatted, and normalized with digit and spoken words to its closest verbalized form. For examples, *Time: 12:45 → 12 hours 45 minutes.*; *Date: 12/31/2022 or (31-12-2022) → 31 December 2022*; *Measures: 10K lb → 10000 lb.; 207.6 kmps → 207.6 kilometers per second.; 2 kg → two kilogram, two kilos, two kilo*, etc.

d. *Numeric Normalization:* We apply our rewrite mapping rules with core logic to generate $N$ spoken form texts for verbalized text by selecting all possible combinations of digits in order, and recursively applying them if necessary.

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*]<br>La media storica di gennaio è di trentuno gradi. [*it*]<br>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*]<br>La media storica di gennaio è di 31 gradi. [*it*]<br>La media histórica de enero es de 31 grados. [*es*] |

e. *Rewrite:* A rewrite module replaces the written-form text with $N$ generated spoken forms.

With our enhanced TN system and a wealth of written text resources at our disposal, we are now able to generates a multitude of possible spoken variations, which is statistically 22× times more diverse than a conventional TN system.

### 2.1.2. Multilingual spoken-written text pairs

We propose using NMT models to generate spoken-written text pairs in target languages for which we do not have adequate pairs. However, we have found that the outputs of the NMT model do not always meet our criteria for quality. To ensure the quality of our spoken-written text pairs, we have implemented the following measures: (a) *Spoken/Written Mismatch:* Discarding translated texts that have mismatches between written and spoken forms, (b) *Word Error Rate (WER):* Strictly adhering to the WER metric for selecting non-ITN text segments, (c) *Target Language Conformity:* Ensuring conformity between the source and target languages, and filtering out any malformed or incorrect translations with input from linguists, like *801 → eight o one [en] ≠ otto o uno [it]*, etc. We provide translation accuracy in Table 6, the pictorial position of translation module in the pipeline in Figure 1, and a few translation examples in Table 2 for reference.

## 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 re-

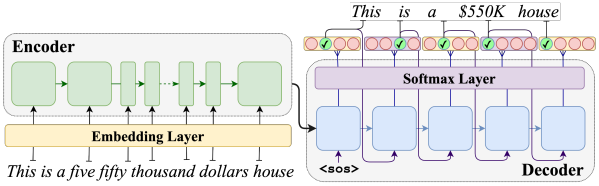


Figure 2: *Encoder-Decoder model architecture for ITN.*

current 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 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.

Table 3: *Sequence-to-Sequence ITN 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 |

# 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. We have chosen two NMT models for our experiments: (a) Opus-MT [9] which is supported by the EasyNMT library, and (b) In-House NMT, an internally developed NMT model that performs similarly to NLLB [10]. Also, unless specified, we use In-House NMT for data augmentation as it has better BLEU scores 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, these two datasets contain English/target language text pairs. Our method utilizes only the written English texts with ITN entities for each selected languages. For low-resource languages such as Kazakh [*kk*], the number of written English texts containing ITN entities can be as low as 739 sentences. The spoken-written pairs for training are generated using the pipeline in Fig. 1.

[1] https://opus.nlpl.eu/

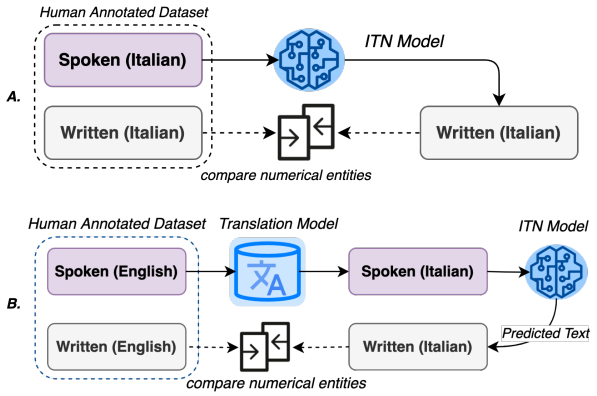


Figure 3: *ITN evaluation strategy with case A: Target language evaluation dataset is available; Case B: Target language evaluation dataset is not available. We use translation model to prepare spoken target texts. In both cases, we compare numerical entities output from the ITN model.*

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 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. In Case B, when the target language pair dataset is not available, we translate the spoken form text from the human-annotated English dataset to the target language. We then apply the trained ITN model to obtain the written form text in target language. Thereafter, we verify if 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 an incorrect one. Note that translation models may output written form in the target language for spoken source input, which we discard, and only evaluate the ITN model on correctly translated spoken form texts. We use normalized accuracy to measure the performance of our ITN model expressed as the fraction of correct prediction over all the ITN entities.

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)**. We ensure to handle special cases when comparing written English with the written form in the target language, follows a few examples:

- **12/24 hour conversion**: Use of 24 hours system in French and other target language; *1:30 p.m. [en]* vs. *13h30 [fr]*.
- **Accommodating Zeros**: Use alternate magnitudes; e.g., *24,000 [en]* vs. *24 mille [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.<br>[*fr*] J'ai découvert que **neuf** statistiques sur **dix** sont fausses. | [*en*] I found out that **9** out of **10** statistics are wrong.<br>[*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**.<br>[*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**.<br>[*fr*] La **60ème** surprise de papa a lieu ce samedi. Arrivée avant **18h**. |

- **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, with marginal improvement seen in Tamil [*ta*], which uses a different script from any other languages. 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*].

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 OpenSubtitles training set.

Finally, Table 8 shows the ITN normalized accuracy when using a different model architecture and different NMT model [9]; shows that the LSTM-based Seq2Seq model is better than the Transformer model, and In-House NMT is better than the Opus-MT for our task. In Table 9, we compared the rule-based ITN and DD-ITN systems with and without data augmentation techniques. Our evaluation on Google's text normalization dataset [13] with over $70k$ ITN entities showed that incorporating our enhanced text normalization data augmentation for English had a positive impact. Additionally, our analysis provided a detailed breakdown of the normalized accuracy for each ITN entity type, which clearly demonstrated that the performance of DD-ITN with the enhanced TN system was significantly better than that of other ITN systems. Notably, we observed considerable improvements in accuracy for ordinal, decimal, measure, and time entities.

Table 6: *Normalized accuracy of monolingual and multilingual models on the Dictation testset.*

| 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% |

‡ lang., trans., and acc. stands for languages, translation, and accuracy.
† 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%** |

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 | 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 9: *Comparison of standard rule-based ITN system and data-driven ITN with/without enhanced TN system for English in normalized accuracy.*

| Class | Size | Rule-based ITN | Data-driven ITN without enhanced TN system | Data-driven ITN with enhanced TN system |
|---|---|---|---|---|
| Cardinal | 10000 | 92.00% | 55.00% | **97.00%** |
| Ordinal | 10000 | 82.87% | 50.90% | **97.50%** |
| Decimal | 10000 | 2.20% | 53.10% | **90.80%** |
| Measure | 10000 | 46.70% | 42.79% | **88.20%** |
| Telephone | 4024 | 93.20% | 88.00% | **93.80%** |
| Digit | 5442 | 77.60% | 56.80% | **90.60%** |
| Time | 1159 | 36.10% | 35.40% | **73.60%** |
| Date | 10000 | 84.50% | 46.00% | **89.40%** |
| Money | 10000 | 53.80% | 34.90% | **55.40%** |
| **Overall** | **70625** | **63.20%** | **51.40%** | **86.20%** |

## 4. Conclusion

In this study, we examine the effectiveness of a language-agnostic data-driven ITN model. Our results show that using a single 12-language model can substantially improve the normalized accuracy of low-resource languages while maintaining reasonable performance for high-resource languages, even with a fixed model size. Additionally, we investigate the architecture and machine translation model used in our framework and find that our best-performing system utilizes a Seq2Seq model with an In-House NMT approach. For future work, we plan to focus on designing better architectures for on-device deployment, evaluating our model on clean and native language datasets, adapting our data augmentation methods to better handle native language nuances and variations, and expanding our approach to include more languages.

# 5. References


[1] E. Pusateri, B. R. Ambati, E. Brooks, O. Platek, D. McAllaster, and V. Nagesha, "A mostly data-driven approach to inverse text normalization." in *INTERSPEECH*. Stockholm, 2017, pp. 2784–2788.

[2] 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.

[3] M. Sunkara, C. Shivade, S. Bodapati, and K. 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] A. Antonova, E. Bakhturina, and B. Ginsburg, "Thutmose tagger: Single-pass neural model for inverse text normalization," *arXiv preprint arXiv:2208.00064*, 2022.

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

[6] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

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

[9] J. Tiedemann and S. 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, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, and e. a. Heafield, "No language left behind: Scaling human-centered machine translation," 2022.

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

[12] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. [Online]. Available: https://arxiv.org/abs/2004.09813

[13] K. Dataset, "Google Text Normalization Challenge," https://www.kaggle.com/datasets/google-nlu/text-normalization, 2023, [Online; accessed 20-Jan-2023].