Combining Automatic Speech Recognition and Neural Machine Translation as *speech2texto*

Katie Brandegee and Nick Hinke

Table of Contents

- Project Overview
- Detailed Approach
- Results
- Next Steps
- Conclusions
- References

Project Overview: Background

Two historically significant problems within the NLP community:

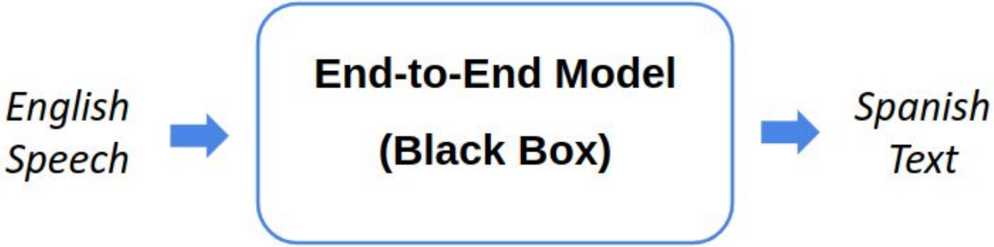
“Speech recognition, a.k.a ***Automatic Speech Recognition*** (ASR), computer speech recognition, or speech-to-text, is a capability which enables a program to process human speech into a written format.”

“***Neural Machine Translation*** (NMT) is an approach to automated translation that uses machine learning to translate text from one language into another.”

https://www.ibm.com/cloud/learn/speech-recognition https://biglanguage.com/blog/everything-to-know-about-neural-machine-translation/https://www.reliancenetworks.co.uk/wp-content/uploads/2018/04/icon-1968243_1280.png
https://softline.si/uploads/resizer/images/2f24d6/0f7da6/82f2ad/origin-mode1-760x362.jpg

Project Overview: Problem Statement

Project Goal: Convert audio samples of natural language (English) into translated text (Spanish)



English Speech

End-to-End Model

(Black Box)

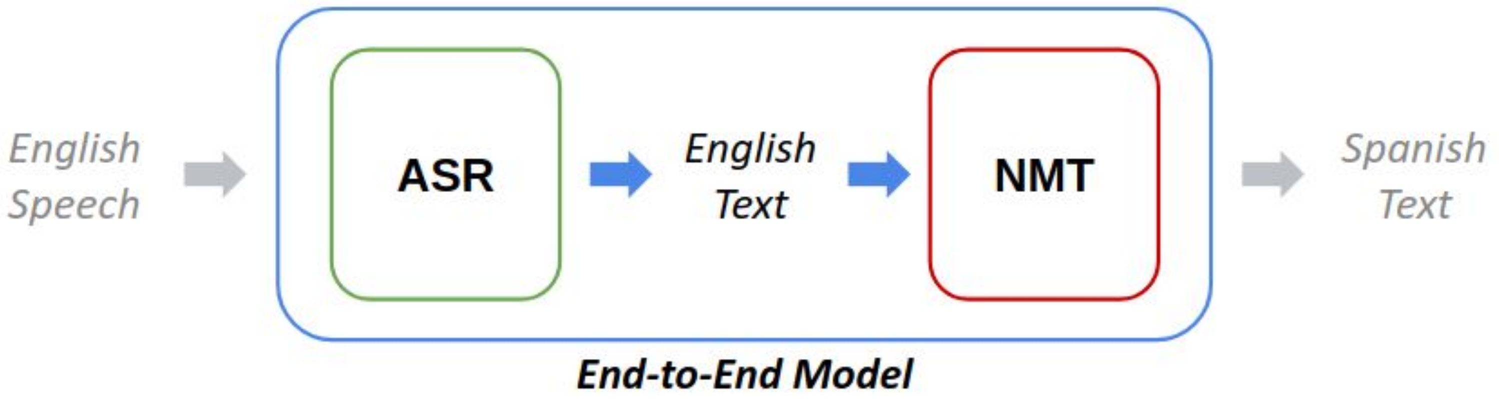
Spanish Text

Table of Contents

- Project Overview
- Detailed Approach
- Results
- Next Steps
- Conclusions
- References

Detailed Approach: ASR Stage

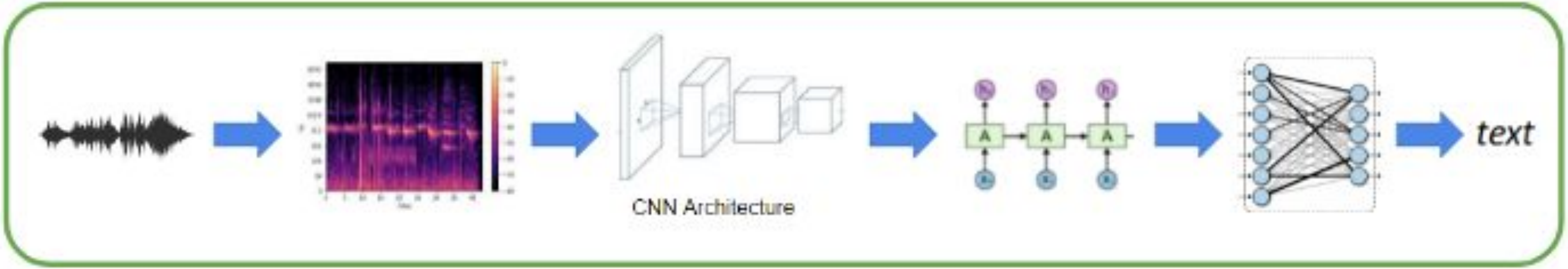
- ➢ *Stage 1:* Automatic Speech Recognition (ASR)
- ➢ *Stage 2:* Neural Machine Translation (NMT)



Detailed Approach: ASR Stage

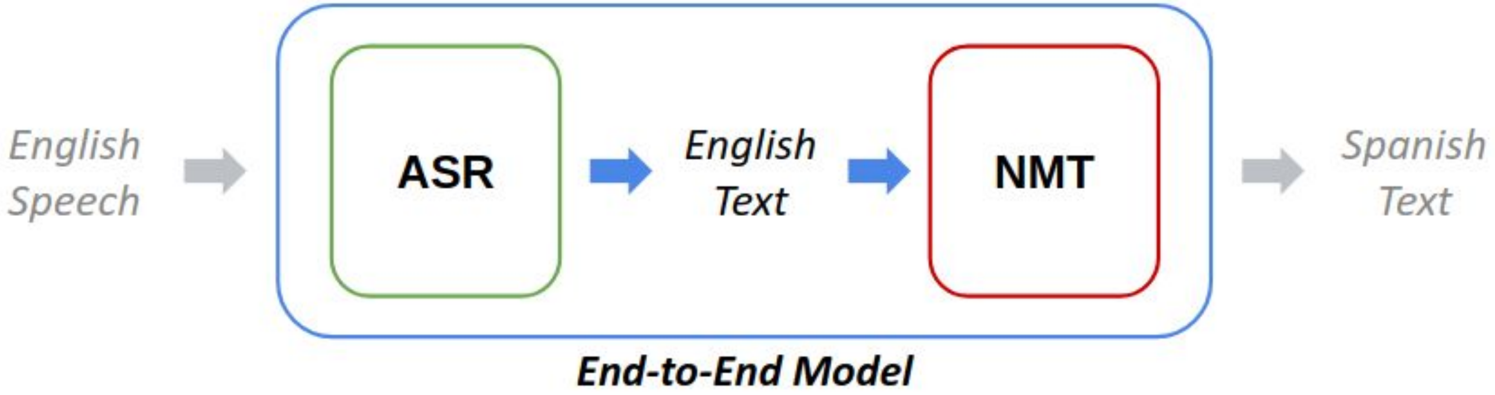
➢ *Stage 1:* Automatic Speech Recognition (ASR)

Audio --> Spectrogram --> CNN --> Bi-dir. LSTM --> FC with Softmax --> Text



Detailed Approach: NMT Stage

- ➢ *Stage 1:* Automatic Speech Recognition (ASR)
- ➢ *Stage 2:* Neural Machine Translation (NMT)

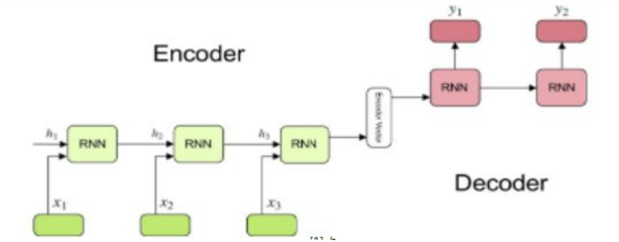


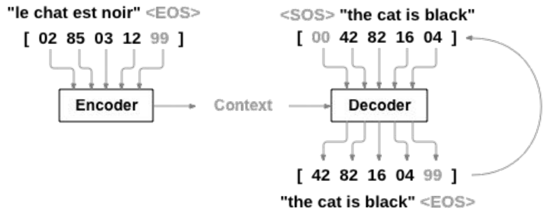
Detailed Approach: NMT Stage

➢ *Stage 2:* Neural Machine Translation (NMT)

Sequence-to-Sequence Model

English Text





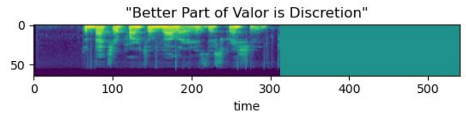
Spanish Text

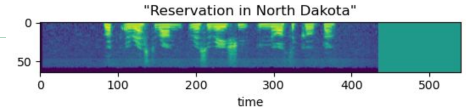
Encoder/Decoder with Attention

https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346

Table of Contents

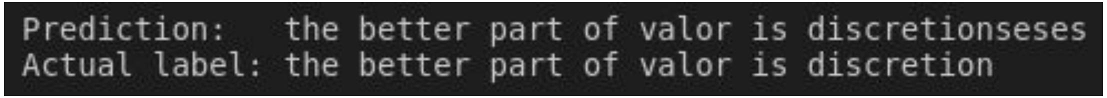
- Project Overview
- Detailed Approach
- Results
- Next Steps
- Conclusions
- References


"Better Part of Valor is Discretion"

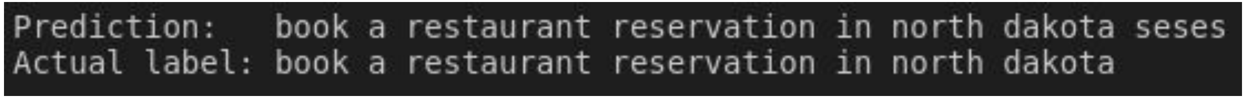

"Reservation in North Dakota"

Results: ASR Performance Evaluation

Overfit on single batch:

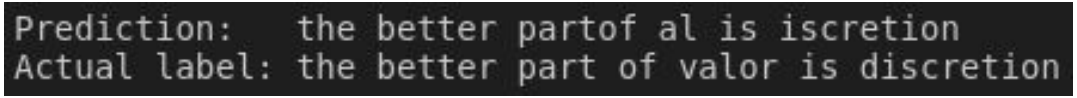


```
Prediction:   the better part of valor is discretionseses
Actual label: the better part of valor is discretion
```

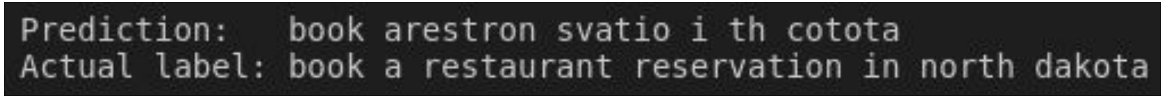


```
Prediction:   book a restaurant reservation in north dakota seses
Actual label: book a restaurant reservation in north dakota
```

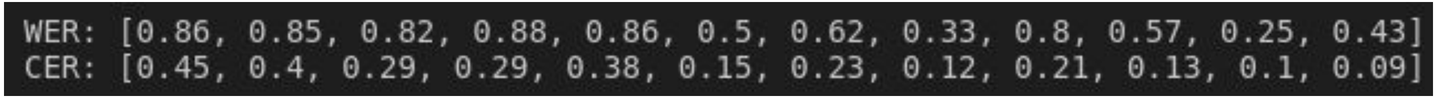
Properly trained on whole dataset:



```
Prediction:   the better partof al is iscretion
Actual label: the better part of valor is discretion
```



```
Prediction:   book arestron svatio i th cotota
Actual label: book a restaurant reservation in north dakota
```



```
WER: [0.86, 0.85, 0.82, 0.88, 0.86, 0.5, 0.62, 0.33, 0.8, 0.57, 0.25, 0.43]
CER: [0.45, 0.4, 0.29, 0.29, 0.38, 0.15, 0.23, 0.12, 0.21, 0.13, 0.1, 0.09]
```

Results: NMT Performance Evaluation

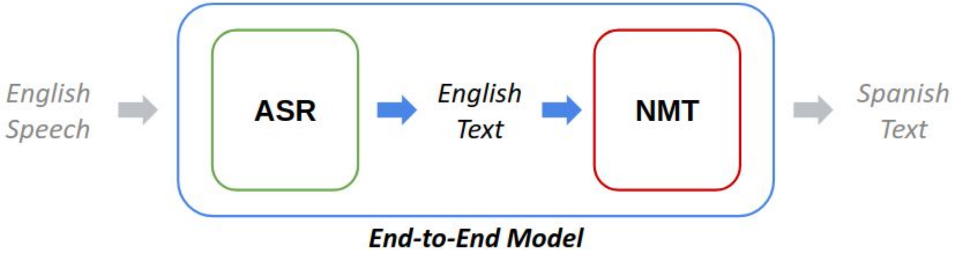
NMT:

	Good Example	Medium Example	Bad Example
Reference/Label	'Él fijó sus ojos en el techo.'	'Ella se cayó y se lastimó la rodilla.'	'¿Qué debería hacer con los libros en la mesa?'
Prediction	'Él fijó sus ojos en el techo.'	'Ella se cayó y lastimó su rodilla.'	'¿Me podrías despertar mañana a la misma hora?'
Reference Translation	He fixed his eyes on the ceiling.*	She fell and hurt her knee.*	What should I do with the books on the table?*
Prediction Translation	He fixed his eyes on the ceiling.*	She fell and hurt her knee.*	Could you wake me up at the same time tomorrow?*

*According to Google Translate

Results: Combining the Models

Practical difficulty with evaluating the combined model:



The English text outputs from the ASR stage are not the same as the English text inputs to the NMT stage.

Therefore, we cannot effectively evaluate the performance of the end-to-end model.

Table of Contents

- Project Overview
- Detailed Approach
- Results
- Next Steps
- Conclusions
- References

Next Steps: Remaining Project Work

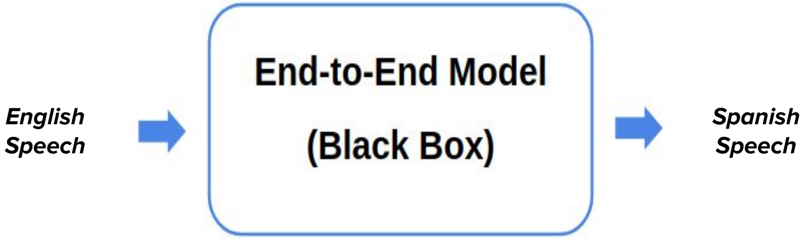
1. Continue training the ASR and NMT models with higher epoch numbers
2. Record and finetune ASR model on our own speech as time permits
3. Finish implementing language model to improve word error rate between the two stages of the model
4. Finish our own implementation of the NMT model and compare its performance to the fine-tuned model
5. Combine the ASR and NMT models for smooth, end-to-end audio to translated text conversion

Table of Contents

- Project Overview
- Detailed Approach
- Performance Evaluation
- Next Steps
- Conclusions
- References

Conclusions: Potential Future Directions

- Expand to more languages
- Create a real-time demo engine
- Create a proper GUI for easy user operation
- Add a third speech synthesis stage to the model to produce a translated audio signal from the existing translated text output


English Speech

End-to-End Model

(Black Box)

Spanish Speech

Table of Contents

- Project Overview
- Detailed Approach
- Performance Evaluation
- Next Steps
- Conclusions
- References

References

- https://huggingface.co/docs/transformers/training
- https://medium.com/@tskumar1320/how-to-fine-tune-pre-trained-language-translation-model-3e8a6aace9f
- https://mastertcloc.unistra.fr/2019/04/29/methods-of-the-machine-translation-evaluation/
- https://neptune.ai/blog/hugging-face-pre-trained-models-find-the-best
- https://towardsdatascience.com/audio-deep-learning-made-simple-part-2-why-mel-spectrograms-perform-better-aad889a93505
- https://towardsdatascience.com/audio-deep-learning-made-simple-automatic-speech-recognition-asr-how-it-works-716cfce4c706
- https://towardsdatascience.com/foundations-of-nlp-explained-visually-beam-search-how-it-works-1586b9849a24
- https://arxiv.org/pdf/1512.02595v1.pdf