Low-Resource Speech-to-Text Translation

Sameer Bansal[1], Herman Kamper[2], Karen Livescu[3], Adam Lopez[1], Sharon Goldwater[1]

[1]School of Informatics, University of Edinburgh, UK
[2]Stellenbosch University, South Africa
[3]Toyota Technological Institute at Chicago, USA
{sameer.bansal, sgwater, alopez}@inf.ed.ac.uk, kamperh@sun.ac.za, klivescu@ttic.edu

Big picture

- Speech-to-text translation has many potential applications for low-resource languages.
- But is available for a tiny fraction of the world's spoken languages as most are zero or low resource.
- Recent work has found that neural encoder-decoder models can learn to directly translate foreign speech in high-resource scenarios.
- Will this work in settings where both data and computation are limited?
- Beyond translations, word-level precision/recall results indicate that models can be useful for keyword-spotting or topic-modeling.

Experimental Setup

- Fisher Spanish speech dataset: a multispeaker corpus of telephone calls in a variety of Spanish dialects recorded in realistic noise conditions.
- Crowdsourced English translations.
- Train models using as little as 20hrs of labeled data.
- Adapt model from state-of-the-art architecture of Weiss et al. [1].
- Simplified to fit on a single GPU: word-level decoder, no model replicas, reduced dim speech features.

Evaluation

We evaluate with several metrics:

- BLEU. Measures up to 4-gram precision.
- METEOR. Allows inexact matches in predictions.
- Word-level unigram precision.

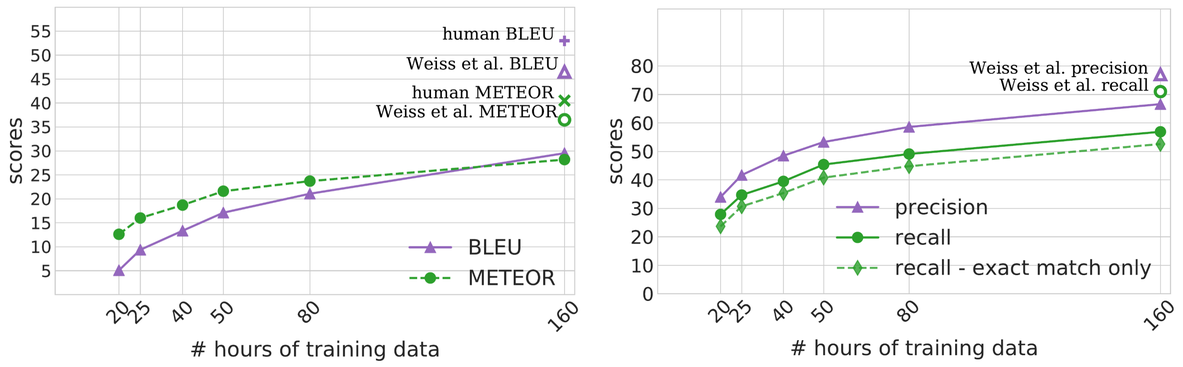

Figure 1: Fisher dev set BLEU, METEOR, precision and recall results.

model	translations
Ref	so no yes but there are people who do get bothered a lot
Weiss et al.	so no yes there are people that do bother a lot
160h	so no if people are bothering a lot
50h	so no yes that's why it bothers me a lot
25h	so i don't know if people who are bother me much
20h	so if you have a car you can do it a lot
Ref	greetings ah my name is jenny and i'm calling from new york
Weiss et al.	hi ah my name is jenny i'm calling from new york
160h	hi ah my name is jenny i'm calling from new york
50h	good ah my name is jenny calling from new york
25h	good ah my name is peruvian i'm calling from new york
20h	good ah my name is jenny

Table 1: Example translations of Weiss et al.'s model and our models on dev set utterances.

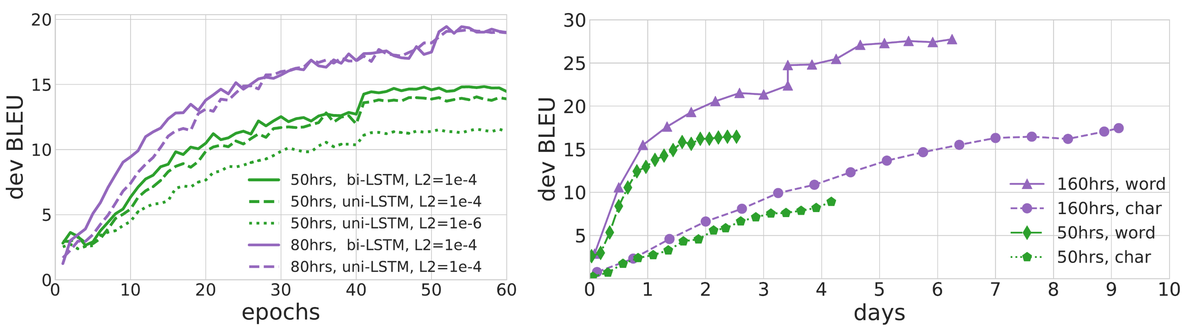

Figure 2: Left: uni-directional vs. bidirectional encoders, L2 loss penalties. Right: word vs char level model training speed

Takeaways

- With only 50hrs of labeled speech data, models achieve high precision and recall—around 50%.
- Models can be trained on a single GPU (Titan X equivalent), and converge in ∼3 days.
- Regularization parameters are critical to model performance.

Future work

- Use sub-word unit modeling to strike balance between speed of word-level decoder, and generalization capacity of character-level decoder.
- Build and evaluate models for cross-lingual keyword-spotting.

Preview of the improved scores:

model	BLEU
Weiss et al.	47.3
20h word	5
20h bpe+noise*	10.8
50h word	17.1
50h bpe+noise*	23.3

Table 2: BLEU scores before and after training improvements.
*Yet to be published.

References

[1] "Sequence-to-sequence models can directly transcribe foreign speech", R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen. In Proc. Interspeech, 2017.

Acknowledgments

We thank Ron Weiss and Jan Chorowski for sharing their translation output, and Kenneth Heafield for giving access to GPUs.