Fluent Translations from Disfluent Speech in End-to-End Speech Translation

Elizabeth Salesky[1], Matthias Sperber[2], and Alex Waibel[1,2]

[1]Carnegie Mellon University, Pittsburgh PA, U.S.A.
[2]Karlsruhe Institute of Technology, Karlsruhe, Germany
elizabeth.salesky@gmail.com

Abstract

- Spoken language translation applications for speech suffer due to conversational speech phenomena, particularly the presence of disfluencies.
- With the rise of end-to-end speech translation models, processing steps such as disfluency removal that were previously an intermediate step between speech recognition and machine translation need to be incorporated into model architectures.
- We use a sequence-to-sequence model to translate from noisy, disfluent speech to fluent text with disfluencies removed using the recently collected 'copy-edited' references for the Fisher Spanish-English dataset.
- We directly generate fluent translations and introduce considerations about how to evaluate success on this task.
- We provide a baseline for a new task, the translation of conversational speech with joint removal of disfluencies.

Challenges:

- Fillers are the most frequent vocab items and are easy to translate
- The original Spanish-English data is mostly one-to-one and monotonic. Clean targets create more challenging alignments.
- Utterances go from short to shorter: down from 11.3 to 8.2 tokens. Single mistake has higher consequences for BLEU.

Takeaways:

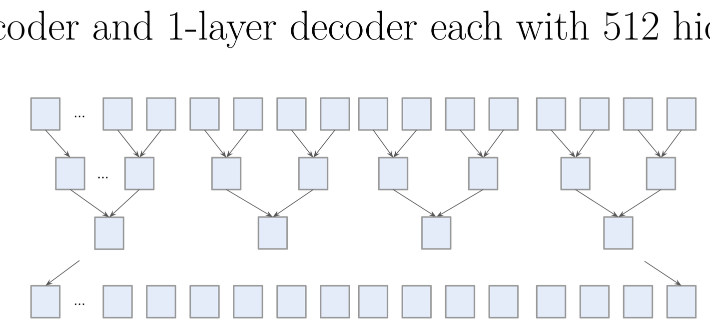
- Can maintain semantic meaning while removing disfluencies (👍)
- End-to-end model performs better than post-processing step
- Provides a baseline for future work to reduce labeled data requirements, e.g. through pre-training or LM multi-tasking
- Evaluation requires care using existing metrics

Model

Initial work on the Fisher-Spanish dataset used HMM-GMM ASR models linked with phrase-based MT using lattices.

Recently, Weiss et al. (2017); Bansal et al. (2018) showed that end-to-end SLT models perform competitively.

- We use an encoder-decoder with attention in **xnmt** with a 3-layer BiLSTM encoder and 1-layer decoder each with 512 hidden units.



- Like Bansal et al. (2018) this is a modified version of Weiss et al. (2017) – all models train in <5 days on 1 GPU
- We do not use convolutional layers to downsample, but instead use network-in-network (NiN) projections from N to N/2
  - Gives the same total 4× downsampling in time
  - Benefit of added depth with fewer parameters
- We use 40-dimensional mel filterbank features with per-speaker mean and variance normalization (Povey et al., 2011).
- We translate to target characters, as opposed to words
- All models use the same preprocessing as previous work on this dataset: lowercasing and removing punctuation except apostrophes.

Contact Information

- **Data:** https://github.com/isl-mt/fluent-fisher
- **Email:** elizabeth.salesky@gmail.com

Data

We use the Fisher Spanish-English dataset which consists of ∼160 hours of ☎ speech and 138k utterances.

The data is conversational and disfluent. Disfluencies can be **filler words** and **hesitations** (um, eh), **discourse markers** (you know, well, mm), **repetitions**, **corrections** and **false starts**, etc.

Original (**ORG**) English translations faithfully translate disfluencies in the source speech. New fluent (**FLT**) references (Salesky et al., 2018) rewrite utterances without disfluencies.

SRC	eh, eh, eh, um, yo pienso que es así
ORG	uh, uh, uh, um, i think it's like that
FLT	i think it's like that
SRC	también tengo um eh estoy tomando una clase ..
ORG	i also have um eh i'm taking a marketing class ..
FLT	i'm also taking a marketing class
SRC	porque qué va, mja ya te acuerda que ..
ORG	because what is, mhm do you recall now that ..
FLT	do you recall now that ..
SRC	y entonces am es entonces la universidad donde yo estoy es university of pennsylvania
ORG	and so am and so the university where i am it's the university of pennsylvania
FLT	i am at the university of pennsylvania

Table 1: Examples of disfluencies in Spanish source (SRC), original (ORG) and fluent (FLT) English translations

- Most common utterances in dataset are 1-2 token backchanneling
- 10.5% of all utterances marked only disfluencies

Output

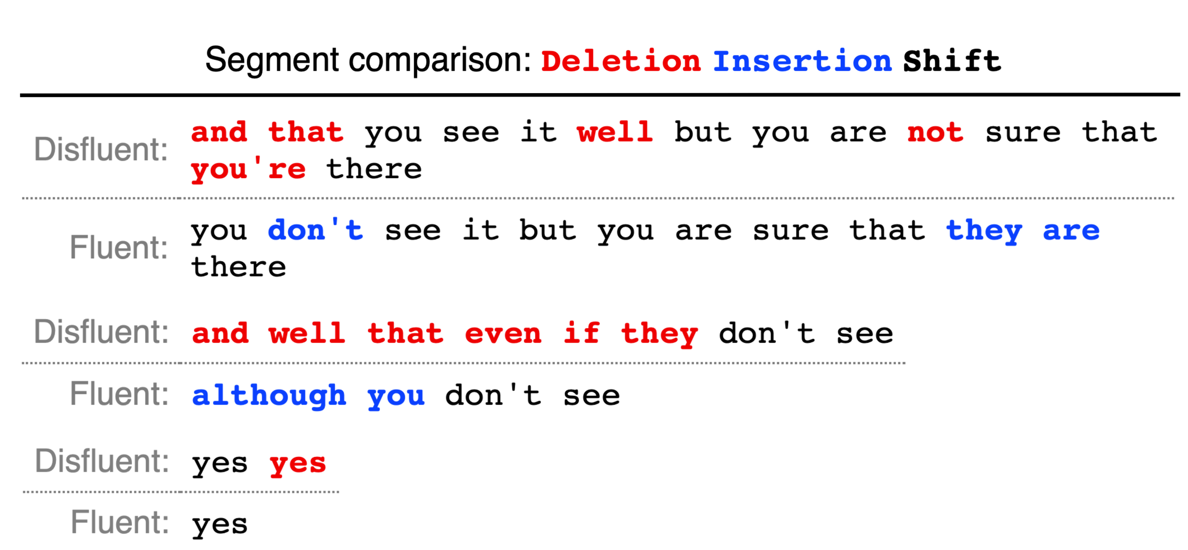


Figure 1: Comparison of example outputs from disfluent and fluent models created with CharCut (Lardilleux and Lepage, 2017).

Notes on Output:

- Training with fluent target data constrains output vocabulary: filler words such as *'um', 'ah', 'mhm'* are not generated.
- Significant reductions in repetitions of both words and phrases
- Instances where the fluent model generates a shorter paraphrase of a disfluent phrase (2nd example above)

Treating disfluency removal as a filtering task can reduce fluency. Removal via MT allows reordering and insertions, **boosting fluency**:

Disfluent *mm well and from and the email is a scandal the spam.*
Fluent *the email is a scandal it's spam.*

Stats Impacting Evaluation:

- Fluent model outputs are 13% shorter with 1.5 fewer tokens per utterance than the disfluent model: avg. utt lengths of 10-11 tokens.
- Scoring against original disfluent refs, shorter length significantly lowers scores: BLEU brevity penalty is 0.86 compared to 0.96-1.0.
  - Removing BP, 1Ref scores are boosted to 19.3 and 19.8 from 16.6 and 17.0 for dev and test – **as good as disfluent model on original data** (Table 3).
  - Fairer comparison: we don't want fluent outputs to match disfluent sequence lengths, and the disfluent models are not penalized due to length.

☞ ***Evaluation using existing metrics requires care***

Evaluation

We evaluate using both **BLEU** and **METEOR**.

- **METEOR** is more 'semantic': we want METEOR scores to be the same with both fluent and disfluent references
- **BLEU** uses modified n-gram precision with a brevity penalty $e^{(1-r/c)}$. We expect scores against fluent references to be lower

METEOR will indicate if meaning is maintained, but not assess disfluency removal, while **BLEU** changes will indicate whether disfluencies have been removed.

Results

Baseline results on *original disfluent* references, test

- 33.7 BLEU, 30.9 METEOR (*4Ref*)
- 19.6 BLEU, 26.1 METEOR (*1Ref*)
- Improvement of 4 BLEU and 2 METEOR over Bansal et al. (2018)
- Do not match Weiss et al. (2017); significantly smaller network

Target Task: disfluent speech → fluent translations

- METEOR scores are almost the same while BLEU scores are lower with the disfluent model 👍
- Fluent outputs should be semantically the same as disfluent outputs but with disfluencies removed
- Scores are lower than disfluent models: fluent references are shorter, so single token changes carry greater weight for BLEU

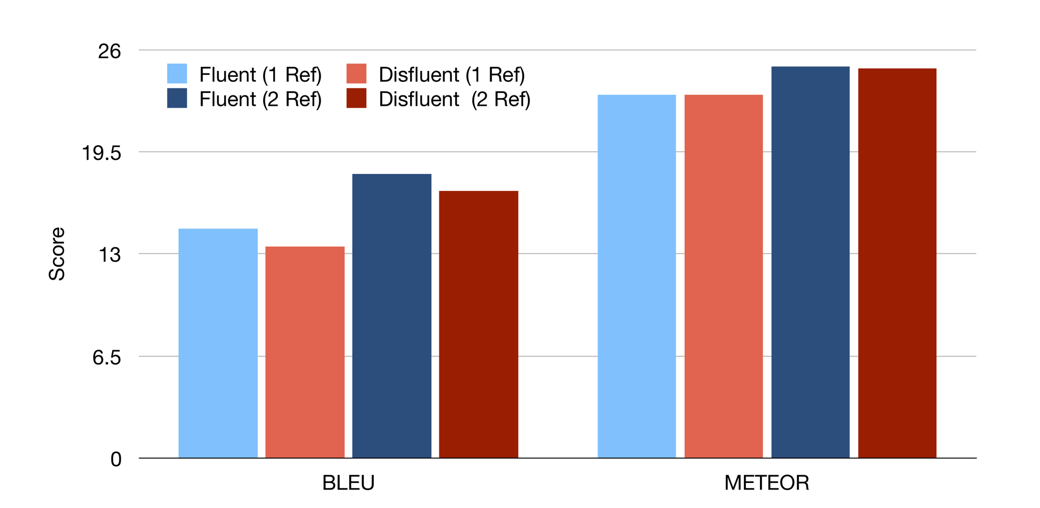


Figure 2: End-to-end model performance evaluated with **new fluent references**. Comparing avg. single reference scores (1Ref) vs multi-reference scores using both generated references (2Ref).

End-to-end or Post-processing Step?

We compare disfluency removal as a post-processing step, using filtering (**Filter**) and monolingual translation (**MonoMT**).

- **Filter** requires labeled spans and may not capture all false starts or repetitions
- **MonoMT** allows for reordering and insertions, boosting fluency
- *Performance:* **Filter** shows slight improvement over disfluent models on dev but not test. **MonoMT** approaches end-to-end model scores but requires the same resources.

	dev		test	
Model	**1Ref**	**2Ref**	**1Ref**	**2Ref**
Postproc. Filter	13.6	16.5	13.5	16.8
Postproc. MonoMT	14.4	17.8	14.4	18.0

Table 2: End-to-end disfluent model with different post-processing steps. Performance evaluated with **new fluent references**.

Comparing to Original References:

- Fewer long n-gram matches with disfluencies removed, BLEU ↓
- Low disfluency recall (filler words, backchanneling), METEOR ↓
- Recall is reduced by ∼14% with the fluent model
  = *approx. % disfluencies in the original data* 👍

		dev		test	
Model	**Metric**	**1Ref**	**4Ref**	**1Ref**	**4Ref**
Fluent	BLEU	16.6	29.8	17.0	30.4
Disfluent	BLEU	19.0	32.4	19.6	33.7
Fluent	METEOR	21.8	25.9	22.7	27.0
Disfluent	METEOR	25.1	30.0	26.1	30.9

Table 3: Evaluating with **original disfluent references**.