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

Elizabeth Salesky¹, Matthias Sperber², and Alex Waibel¹,²
¹Carnegie Mellon University, Pittsburgh PA, U.S.A.
²Karlsruhe 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 models linked with phrase-based MT using lattices.

- 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 4x downsampling in time
  - Benefit of added depth with fewer parameters
- We use 40-dimensional mel filterbank features with pre-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: lowcasing 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. Next fluent (FLT) references (Salesky et al., 2018) rewrite utterances without disfluencies.

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 c=1.0. 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 disfluency references, test

<table>
<thead>
<tr>
<th>Model</th>
<th>1Ref</th>
<th>2Ref</th>
<th>1Ref</th>
<th>2Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postproc. Filter</td>
<td>13.6</td>
<td>16.5</td>
<td>13.5</td>
<td>16.8</td>
</tr>
<tr>
<td>Postproc. MonoMT</td>
<td>14.4</td>
<td>17.8</td>
<td>14.4</td>
<td>18.9</td>
</tr>
</tbody>
</table>

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

Notes on Output:

- Training with fluent target data constrains output vocabulary: filler words such as ‘um’, ‘uh’, ‘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)

Evaluation using existing metrics requires care

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

End-to-end or Post-processing Step?

We evaluate 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 c=1.0. 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.

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 c=1.0. 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.

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

<table>
<thead>
<tr>
<th>SRC</th>
<th>ORG</th>
<th>FLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>eh, eh, um, yo pienso que as</td>
<td>oh, uh, uh, i think it’s like that</td>
<td>i think it’s like that</td>
</tr>
<tr>
<td>SRC</td>
<td>tambien tengo um eh estoy tomando una clase</td>
<td>i also have um eh i’m taking a marketing class</td>
</tr>
<tr>
<td>SRC</td>
<td>porque que se, may be ya se acuerda que</td>
<td>because what is, mhm do you recall now that</td>
</tr>
<tr>
<td>SRC</td>
<td>y entonces am es entonces la universidad donde</td>
<td>and so am and so the university where</td>
</tr>
<tr>
<td>SRC</td>
<td>y estoy es university of pennsylvania</td>
<td></td>
</tr>
</tbody>
</table>

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

- Disfluent model outputs are 13% shorter with 1.5 fewer tokens per sentence
- Fluent model outputs are 13% shorter with 1.5 fewer tokens per sentence
- Most common utterances in dataset are 1-2 token backchanneling
- 10.5% of all utterances marked only disfluences

Table 3: Evaluating isolated disfluences

<table>
<thead>
<tr>
<th>DEV</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluent</td>
<td>BLEU 16.6</td>
</tr>
<tr>
<td>Disfluent</td>
<td>BLEU 19.0</td>
</tr>
<tr>
<td>Fluent</td>
<td>METEOR 21.8</td>
</tr>
<tr>
<td>Disfluent</td>
<td>METEOR 25.1</td>
</tr>
</tbody>
</table>

Filtering boosting fluency

Saving model scores but requires the same resources.