Proposal for a Challenge Task June 14, 2024

Transfer of Pragmatic Intent in Speech-to-Speech Translation

Motivation Current speech-to-speech translation systems work well for many purposes, but are less well suited for informal dialog [LL20]. In particular, their outputs are usually insensitive to the discourse context of the utterance and the interpersonal and pragmatic goals of the source. One reason for these limitations is the fact that evaluation methods currently focus on semantic fidelity and output naturalness.

We accordingly propose a new evaluation, in terms of pragmatic fidelity. For example, given the input *did you see her?* with its specific prosody --- perhaps showing breathless interest, encouraging the interlocutor to continue his story, and displaying sensitivity to the complex emotions he's feeling about a recent break-up and knowing that his ex-girlfriend is still around --- the output should be judged not only on its ability to faithfully convey the lexical content, but also those elements of intent and stance.

Setting up such a challenge task is a workitem of our 3-year NSF-funded project "Modeling Prosody for Speech-to-Speech Translation," and as such the necessary funding is in place.

Goals

Through this challenge task, we plan to:

1. Obtain a baseline of how well current systems succeed at this aspect of translation.
2. Discover and highlight methods that enable systems to do better at this aspect of translation.
3. Discover what types of pragmatic function can be handled by existing methods, and which remain topics for future research, aiming to inform linguistic inquiry, model design, and new data resource collection.
4. Establish a new protocol for systems evaluation, which we hope will be a long-term standard, complementing existing semantic- and fluency-based evaluation methods.

The results for 1, 3, and 4 will be reported in the shared-task overview paper. The results for 2 will be reported in the individual papers describing systems submitted to the challenge. We anticipate that these will include ablation studies and studies of the effectiveness of using various training data. We will also release all the data and software that we use for evaluation.

Overall, we aim to jumpstart a new direction in speech-to-speech translation research. Over the long term, this will enable the evaluation and development of systems able to translate pragmatic intent, and thereby better support speakers needing to convey feelings, stances and intentions, needing to organize the flow and topic structure of dialog, and seeking establish shared understandings and rapport. These systems will support many use cases that are currently ill-served.

Task Systems will take as input a folder of about 1000 English utterances extracted from dialog and output a folder of Spanish utterances. The inputs will be audio-only: no transcriptions will be provided. As a variant condition, systems inputs will include not only the source language utterances

but also their full left contexts in the dialog. Another variant condition will be in the Spanish-English direction, using a disjoint set of utterance pairs.

Evaluation Data. We will exploit the "Dialogs Reenacted across Languages" (DRAL) corpus WA23]. This consists of paired Spanish-English utterances, closely matched for communicative intent. These were created by having bilinguals engage in real, spontaneous conversations, and subsequently re-create selected utterances in their other language. Utterances are mostly 1-4 seconds in length. This data set is provisionally released at https://www.cs.utep.edu/nigel/dral/ , with a final version to be released by the Linguistic Data Consortium this summer.

Training Data We will provide only about 500 English-to-Spanish audio pairs. While teams may use these for tuning meta-parameters or any other purpose, we expect that systems will be trained on other resources, perhaps monolingual dialog data and bilingual monologue data.

Evaluation

1. **Human Quantitative Evaluation**. A panel of 5-6 judges will independently score the pragmatic fidelity of the translation on a scale from 1-5, using an adaptation of the protocol of [WM24]. As this is time-consuming, we will probably do this for only 80 utterances. Assuming that we have submissions from only 5 teams, we can probably do this all in one morning.
2. **Segura Metric.** Outputs will be scored by their similiarity to the human-generated, gold translation. This similarity metric [S24] uses the cosine similarity between the representations of two utterances in terms of 103 HuBert features, that were selected to maximize match to a collection of pragmatic similarity judgments [WM24].
3. **Explainable Metric.** The same, but the similiarity metric will be based on the Euclidean distance based on a few dozen features, with appropriate weights, that describe the pragmatically-important prosodic features of each utterance. The weights also will be learned to maximize match to the [WM24] judgments.
4. **Human Qualitative Evaluation.** A focus group of bilinguals will listen to about 10 outputs for each system, having access to the reference translation and to the source utterance. We will apply qualitative-inductive methods to identify common strengths and weaknesses, and also to identify illustrations of ways in which systems differ.

For the first three metrics we will report a topline, the ratings of human translations produced by an independent bilingual speaker. We may also generate baseline scores using the outputs of some widely-used, but prosodically/pragmatically insensitive system, like Google Translate. For the explainable metric we will also report the baseline of directly transferring the prosody of the source.

Expected Types of Participation

1. Developers of large research systems, such as Meta's Seamless and Google's AudioPaLM, might use this task to explore the utility of different configurations of their systems for this purpose. While pragmatic-intent transfer is not a specific priority for any research group that we know of, there is much interest in related topics, including speaker-property transfer and transfer of emotion and expressiveness through prosody.
2. Academic research teams may have new ideas, new systems or novel modules (such as post-hoc resynthesis for prosodic adjustments) that could be tested using this task.
3. The operators of production systems, such as Google Translate and Skype Translator, may want to use this task to investigate how well their systems currently work for pragmatics-intensive use cases.

Status and Support

Funding: We will use NSF funds for a research assistant to set up the workflow and answer any questions from participants, for human subjects for the quantitative and qualitative evaluations, for further data collection if needed, etc.

Data: We have adequate data. We may collect additional data to have multiple references per input. (In future years, we are planning to collect more, initially Spanish-English and English-Spanish, and then another pair. We can accelerate the schedule if needed.)

The Segura Metric. This is functional (Segura 2024), fast enough, and released, so teams can use it themselves.

The Explainable Metric. This will be prosody-based, as an improved version of one we developed last year (Avila, 2023).

Quantitative Evaluation Protocol. This will be very minor variant of an IRB-approved protocol that we used in May.

Metric Relations: A fundamental assumption of the automatic metrics is that similarity to the reference translation is a good proxy for fidelity of the translation. This is a common assumption in MT evaluation, but has never been tested for *pragmatic* similarity. We are planning to empirically study the relationship.

Qualitative Evaluation Protocol. This will use methods that we have applied in the past.

Timeline Possibilities

If Interspeech:

- Proposal: due late December, accepted mid-January
- Data Release: Jan 20: final release of data and metrics; Jan 25: dry-run evaluation to test workflow, ideally done by all teams;
- Evaluation: Feb 15-25 evaluation (teams upload final submissions, organizers report quantitative results to teams, qualitative results slightly later
- Paper submission: early March: teams submit titles with abstracts, then the updated papers
- Conference: August, Rotterdam

If IWSLT:

- Proposal: due in September? Accepted in October?
- Data Relaease: January
- Evaluation: April 1-15
- Paper submission: April 30
- Conference: July?

Notes: An Interspeech Challenge would be seen more widely. A IWSLT Challenge would have a better timeline, and be more likely to allow continuation for a second year.

Organizing Committee

Nigel Ward, University of Texas at El Paso, nigel@utep.edu

Olac Fuentes, University of Texas at El Paso, ofuentes@utep.edu

others TBD

References

[AW23] Towards Cross-Language Prosody Transfer for Dialog. Jonathan E. Avila and Nigel G. Ward. Interspeech 2023.

[LL20] Liebling, Daniel J., et al. "Unmet needs and opportunities for mobile translation AI." Proceedings of the CHI conference on human factors in computing systems. 2020.

[WM24] A Collection of Pragmatic-Similarity Judgments over Spoken Dialog Utterances. Nigel G. Ward and Divette Marco. LREC-COLING 2024. Data at https://www.openslr.org/152/

[WS24] Towards a General-Purpose Model of Perceived Pragmatic Similarity. Nigel G. Ward, Andres Segura, Alejandro Ceballos, Divette Marco. Interspeech 2024. Code at https://github.com/andysegura89/Pragmatic_Similarity_ISG

[WA23] Dialogs Re-enacted Across Languages, Version 2. Nigel G. Ward, Jonathan E. Avila, Emilia Rivas, Divette Marco. Technical Report UTEP-CS-23-27, 2023. Data samples and full download at https://www.cs.utep.edu/nigel/dral/