Completing and Validating the Re-Aligned Switchboard Dialog Act Corpus

Run Chen[1,2], **Zihao Tao**[2], **John Prado**[2,3], **Ignazio LaManna**[2]
Ryan Puterbaugh[2], **Mim Datta**[2], **and Julia Hirschberg**[2]
[1]Google, USA [2]Columbia University, USA [3]University of Alberta, Canada
julia@cs.columbia.edu

Abstract

Although widely used in dialog act prediction and generation, the Switchboard Dialog Act (SwDA) corpus has performed poorly in models incorporating prosodic information because of misalignment between speech and text data. In this paper, we report our completion of the work begun in Chen et al. (2024) in addressing these misalignment issues with an improved SwDA corpus called RASwDA (Re-Aligned Switchboard Dialog Act Corpus). Now fully realigned and validated, RASwDA finally meets standards of accuracy allowing for classification models trained on it to exceed classification benchmarks set by models trained on other Switchboard subcorpora.

1 Introduction

Since its creation, the Switchboard Dialog Act (SwDA) corpus has been widely used for dialog act prediction and generation tasks. However, due to the misalignment between speech and text data in SwDA, models that incorporate prosodic information have shown poor performance. This paper presents the completed Re-Aligned Switchboard Dialog Act (RASwDA) corpus, as introduced in Chen et al. (2024), and reports novel state-of-the-art dialog act classification (DAC) results obtained directly from the efforts to manually correct and validate the previously force-aligned speech and text data. Over the DEVtest of Chen et al. (2024), this paper makes four contributions:

- We re-align the remaining 617.5 SwDA conversations and validate all 1,155, with a Cohen's κ of 0.90 between the post-alignment and post-validation stages.
- We analyze inter-stage disagreement and find that Statement-Non-Opinion $\leftrightarrow$ Statement-Opinion accounts for the majority of label changes.
- We train an acoustic-feature only BiLSTM model that saw a 41% improvement from training on the files from the original alignment to the post-alignment, showing that the incorrect alignments left prosodic features too noisy to learn from.
- We also train a late-fusion multi-modal RoBERTa model that sets a new DAC benchmark on SwDA (87.8% accuracy).

In §2, we outline some history of the Switchboard Telephone Speech Corpus. In §3, we summarize the procedure of Chen et al. (2024), present additional steps taken to validate the data set, and explain the models chosen for the DAC task. In §4 we present the results of an inter-annotator reliability test and classification models, and in §5 we discuss our conclusions.

2 Background

The **Switchboard Telephone Speech Corpus**, consisting of roughly 2,400 two-way telephone conversations between speakers of various US dialects, was created under DARPA sponsorship by Texas Instruments in 1990 but first released by the National Institute for Standards and Technologies in 1992 (Godfrey et al., 1992). The automated collection process prioritized the pairing of speakers who had not yet conversed either with each other or about the predetermined prompt. Participants initiated data-collection themselves, by giving a signal to begin the recording once prepared. The original corpus included time-aligned transcriptions of the speakers' utterances, as well as the first attempt at automatic word-by-word time-stamping based on those transcriptions (Wheatley et al., 1992).

The original approach to aligning the transcriptions and DA tags with the appropriate utterances was a GMM-HMM Switchboard recognition system to generate the corresponding time intervals

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 173–177

(Shriberg et al., 1998, p. 454). However, consumer recording devices and automatic speech recognition systems have come a long way since then, and in fact the original outputs were so poorly aligned that they severely impaired research efforts in automatic tagging (Stolcke et al., 2000). Most force-alignment errors appear during quiet or low-energy utterances, but by manual inspection and verification of hundreds of audio files, we also find errors caused by background noise and line static. Similarly, the transcriptions contain incorrect or missing words, especially during speakers' dialectical or nonstandard phrasings.

Since its creation, subcorpora of **Switchboard** have served as the basis for many re-annotations and re-segmentations. We discuss a selection of the most relevant to the present paper.

In 1997, researchers from UC Boulder and SRI International completed an annotation of roughly half of this corpus, LDC's Switchboard-1 Release 2 (LDC97S62) (Jurafsky and Shriberg, 1997), with an augmented version of the Discourse Annotation and Markup System of Labeling (DAMSL) set of tags (Allen and Core, 1997). This **SwDA corpus** consists of 1,155 of the original conversations, and each of the roughly 205,000 utterances is tagged as one of 42 Dialog Acts, including Statement-opinion (sv), Acknowledge / Backchannel (b), Yes-No-Question (qy), etc.

By 1999, the third iteration of the Penn Treebank Project, **Treebank3**, included tagged, parsed, and disfluency-annotated text from 650 of the 2,400 Switchboard conversations (Marcus et al., 2000). This subcorpus formed the basis of the 2010 conversion of the corpus using the NITE XML Toolkit (hereafter **NXT**) (Carletta et al., 2005), which integrated different annotation protocols and allowed for easier traversal and querying of the data and its features (Calhoun et al., 2010). The benefits of this more complex representation included the ability to track long-distance dependencies in speech, to search the annotations as a set of attribute n-tuples, and to preserve the internal consistency of the constituent annotations.

In 2024, a team at Columbia University's Department of Computer Science conducted a DEVtest to assess the efficacy of manual correction of previous forced alignments of DAMSL DA tags and corresponding Switchboard conversations (Chen et al., 2024). They used TextGrids, Praat's tiered annotation format in which each tier holds a series of labeled time-stamped intervals. This DEVtest included 537.5 of the 1,155 SwDA conversations, relying on **NXT-format Switchboard**'s XML files to create time-aligned TextGrids for the conversations included, and the *aneas* library for the rest (Chen et al., 2024).

3 Procedure

Our continuation of the work of Chen et al. (2024), to complete the **RASwDA** corpus, is the result of a year-and-a-half long process consisting of three stages: manual re-alignment of the remaining **SwDA** corpus, a manual validation of this re-alignment, and an evaluation stage. We follow the earlier data-preparation procedure (Chen et al., 2024), whose DEVtest re-aligned and validated 537.5 conversations (1075 individual speaker transcripts) of the 1,115 SwDA conversations (2,310 individual speaker transcripts): the remaining 617.5 conversations were drawn from both the NXT-format Switchboard Corpus (Calhoun et al., 2010) and LDC Switchboard-1 Release 2 (Godfrey et al., 1992). The NXT-format Corpus consists of 642 conversations in XML file-format, which we parsed into TextGrid format. For conversations not already in the NXT Corpus, we parsed each conversation's transcript into separate transcripts, one for each speaker, and likewise separated each conversation's audio into two WAV files using SoX. We computed the forced alignment for each utterance in each speaker transcript with the *aeneas* library (Pettarin, 2017), which also resulted in a set of TextGrid files. The issues affecting forced alignment by *aeneas* (e.g. background noises) observed in the DEVtest conducted by Chen et al. (2024) persisted into the remaining 617.5 conversations.

3.1 Manual Re-alignment Stage

Like in Chen et al. (2024), we used the Praat speech analysis software (Boersma and Weenink, 2026) to manipulate these TextGrid-format SwDA transcripts. In addition to correcting the alignment of transcript time-intervals to the associated audio, the team of aligners was instructed to mark speaker overlap and laughter with special "SIL" and "<laughter>" tokens and to correct mis-transcriptions, other segmentation errors, and omissions in the transcript. To the best of our native-English-speaking judgment, we also resolved the mis-transcriptions and segmentation errors which the original SwDA annotators themselves marked

for correction at a later date (Jurafsky and Shriberg, 1997). Our aligners included six undergraduates, one post-baccalaureate, and one graduate student in computer science and linguistics programs, some compensated for their time in course credit.

3.2 Validation & Evaluation Stages

After completing manual re-alignment of the remaining 617.5 SwDA conversations, a validation stage began. A small team of validators (a subset of the aligners) reviewed all 1,155 conversations of the completed RASwDA corpus to verify that re-alignment had been performed for each conversation and to ensure any remaining marks were resolved. No student worked with the same file more than once from the re-alignment stage to the validation stage. General disagreements which validators brought against decisions made by aligners during the previous stage were discussed and addressed during team meetings. To formally assess annotator (dis)agreement among the aligners and validators, we computed the inter-stage reliability between post-alignment and post-validation DA tags. First, we extracted every DA tag from each TextGrid file in both versions of the corpus, recording all start times, end times, and labels. To compare the same utterance between the alignment and the validation stages, we matched tags by interval midpoint—rather than boundary—in order to tolerate segment boundary shifts of up to half the interval's duration. The corresponding DA tags were then compared using Cohen's Kappa, which measures agreement between the two annotation passes beyond what would be expected by chance.

Figure 1: Most-agreed-upon SwDA dialog acts between the alignment and validation stages.

To identify which DAMSL categories carry inherently opaque identification boundaries- information that downstream users of RASwDA need when interpreting model confusion on this corpus, we visualized inter-stage (dis)agreement and generated word bubbles for the DAMSL DA tags that were

Figure 2: Most-disagreed-upon SwDA dialog acts between the alignment and validation stages.

most- and least-frequently changed. For the agreement word bubble (Figure 1), we pulled the diagonal values from the confusion matrix (i.e., the labels that validators left unchanged), sorted by frequency, and kept the top 10 most-agreed-upon tags. We then mapped each DAMSL code to its full dialog act name using a dictionary built from the DAMSL manual in Jurafsky and Shriberg (1997) and generated a word bubble sized by frequency. For the disagreement word bubble (Figure 2), we pulled all off-diagonal values from the confusion matrix (i.e., where row $\neq$ column: the labels that validators changed from the original to a new DA tag). We then summed each tag's total outgoing changes and kept the top 10 most-disagreed-upon tags. Using the same DAMSL dictionary, we generated another word bubble sized by frequency.

3.3 Dialog Act Classification Models

As an extrinsic measurement of our realignment efforts, we compared the performance of two classification models, an acoustic only Bidirectional Long Short-Term Memory (Bi-LSTM), and a late fusion multi-modal RoBERTa model. We instantiate three instances of these models, trained on text grids from each stage of the re-alignment process. Both models utilize acoustic features extracted using the OpenSMILE's low level descriptor (LLD) feature set, eGeMAPSv02.

3.3.1 Bi-LSTM

The Bi-LSTM encoder processes variable-length acoustic feature sequences timestep-by-timestep, capturing context from both past and future frames simultaneously. We adopt a two-layer architecture with inter-layer dropout: the first layer learns low-level temporal patterns, while the second composes these into higher-level representations, with dropout regularizing the inter-layer activations. The encoder's outputs are collapsed into a fixed-size ut-

terance embedding via learned attention pooling, which weights each frame by its relevance to the classification decision — down-weighting uninformative regions such as silence or background noise. Loss is computed using weighted cross-entropy, which inversely scales each class's loss contribution to account for the severe label imbalance present in naturalistic dialogue data.

3.3.2 RoBERTa

The multi-modal model augments a pretrained RoBERTa-based text encoder with the same acoustic branch used in the Bi-LSTM model. The text encoder produces a 768-dimensional utterance representation from the [CLS] token, which aggregates sequence-level meaning during pretraining. The acoustic branch processes the eGeMAPSv02 feature sequences through an identical two-layer bidirectional LSTM with attention pooling, projecting the result to a 256-dimensional embedding. Rather than concatenating the two modalities directly, a learned scalar gate, conditioned on both the text and acoustic representations, controls how much the acoustic signal contributes for each utterance. This allows the model to down-weight noisy or uninformative acoustic signals when the text alone is sufficient, and to leverage prosodic cues when they are disambiguating. The gated acoustic embedding is concatenated with the text representation and passed through a two-layer MLP classifier trained on 10 dialogue act classes, with loss again computed using weighted cross-entropy.

4 Results

The Cohen's Kappa of 0.90 falls within the 'almost perfect' range (0.81–1.00) on the Landis and Koch (1977) scale, reflecting strong inter-annotator reliability across post-alignment and post-validation annotations. The agreement bubble was dominated by Non-Verbal (12,232), Statement-Non-Opinion (5,456), and Backchannel (3,324). Despite ranking second in agreement, Statement-Non-Opinion also has the most outgoing changes in the disagreement bubble (723), with Statement-Non-Opinion to Statement-Opinion accounting for 578 of those changes, making it the primary source of annotator disagreement. That some DAMSL DA tags appear both frequently agreed-upon and frequently disagreed-upon we ascribe to a fuzziness inherent of the labels themselves.

4.1 Model Performance

Table 1		
Model	F1	Accuracy
Bi-LSTM		
Original	0.06924	15.536
Post-Align	0.30673	56.538
Post-Valid	0.29351	57.223
RoBERTa		
Original	0.72120	86.725
Post-Align	0.67939	87.600
Post-Valid	0.66812	**87.852**

Table 1: Model Results

As seen in Table 1, the results of our efforts produced significantly improved results. The Bi-LSTM is the more telling test of re-alignment because it relies on acoustic features alone. On the original data, it reached 15.5% accuracy, barely above chance for 42 classes. The original forced alignments mapped DA labels to the wrong portions of the audio, so the extracted features were effectively noise. After validation, accuracy rose to 57.2% over the three iterations of manual TextGrid correction, showing that corrected boundaries let the model pick up prosodic patterns it could not access before. The bold figure highlights a new state-of-the-art score for DAC models trained on the SwDA corpus, while the 41.7% increase in the performance from the original data to the post-validation data signifies that our model went from essentially utilizing random chance to a fully-functional classification model which relies on a number of temporal acoustic features from audio signals.

In addition, the performance distribution over the multi-modal DAC models shows that although text accounts for the most informative feature on classification for the accuracy of DAC tasks, corresponding acoustic features clearly provide a performance boost taking the model up from 86.7% to 87.8%.

5 Conclusion

This paper has presented the completed RASwDA corpus: the result of a large-scale manual re-alignment effort, as introduced in (Chen et al., 2024), to address issues present in SwDA (Jurafsky and Shriberg, 1997). With a Cohen's κ of 0.90 and new DAC benchmarks on both acoustic-only and multi-modal models, we have shown the efficacy of manual alignment, annotation, and validation in

producing a well-aligned corpus. The validation stage confirmed the quality of the post-alignment data while also surfacing systematic label ambiguities, particularly between sd and sv, that users of RASwDA should be aware of.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers.

Paul Boersma and David Weenink. 2026. Praat: Doing phonetics by computer.

Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419.

Jean Carletta, Stefan Evert, Ulrich Heid, and Jonathan Kilgour. 2005. The NITE XML Toolkit: Data model and query language. *Language Resources and Evaluation*, 39(4):313–334.

Run Chen, Eleanor Lin, Shayan Hooshmand, Mariam Mustafa, Ritika Nandi Rose Sloan, Alicia Yang, Andrea Lopez, Ansh Kothary, Isaac Suh, Catherine Lyu, Eric Chen, Sophia Horng, and Julia Hirschberg. 2024. RASwDA: Re-aligned switchboard dialog act corpus for dialog act prediction in conversation. In *[Proceedings] International Workshop on Spoken Dialogue Systems Technology*.

John Godfrey, Edward Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520.

Dan Jurafsky and Elizabeth Shriberg. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13.

J. Richard Landis and Gary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mitchell Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 2000. Treebank-3.

Alberto Pettarin. 2017. aeneas.

Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3–4):443–492. PMID: 10746366.

SoX. 2015. SoX: Sound eXchange.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech.

Barbara Wheatley, George Doddington, Charles Hemphill, John Godfrey, Edward Holliman, Jane McDaniel, and Drew Fisher. 1992. Robust automatic time alignment of orthographic transcriptions with unconstrained speech. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 533–536.