RASwDA: Re-Aligned Switchboard Dialog Act Corpus for Dialog Act Prediction in Conversations

Run Chen[1] Eleanor Lin[1] Shayan Hooshmand[1] Mariam Mustafa[1] Rose Sloan[1,2]
Ritika Nandi[1] Alicia Yang[1] Andrea Lopez[1] Ansh Nikhil Kothary[1] Isaac Suh[1]
Catherine Lyu[1] Eric Chen[1] Sophia Horng[1] Julia Hirschberg[1]

[1]Columbia University [2]Bard College

Motivation

The **Switchboard Dialog Act Corpus (SwDA)** [1] is important for modeling dialog act prediction and production. Although several corpora offer sizable annotated speech data in multi-participant meetings, only SwDA exclusively comprises dialogs between two individuals, making it particularly relevant for modeling the types of two-party interactions prevalent in conversational systems today. However, it suffers from a critical limitation: **inaccurate alignment**.

Related Work

Switchboard Dialog Act Corpus (SwDA)

- based on the *Switchboard Corpus*, 2,400 two-sided telephone conversations [2]
- 1,155 conversations, 42 dialog act labels
- force-aligned using GMM-HMM speech recognition system [3]
- inaccurate due to ASR errors and background noise

There is very little evidence that the use of speech features from the currently aligned corpus significantly improves their results in any way and sometimes even leads to worse performance [4, 5].

While the NXT-format Switchboard Corpus links the transcriptions in SwDA with accurate manual alignments, it does so for only 642 of the 1,155 conversations in SwDA [6].

To date, no one has produced a full re-alignment of all 1,155 SwDA conversations.

SwDA Alignment Diagnosis

- Misaligned transcript w.r.t. speech
- Incorrect transcript
- Missed short/overlapping speech (e.g. backchannel)
- Recorded on the wrong channel (27 files found so far)
- Propagated from early errors

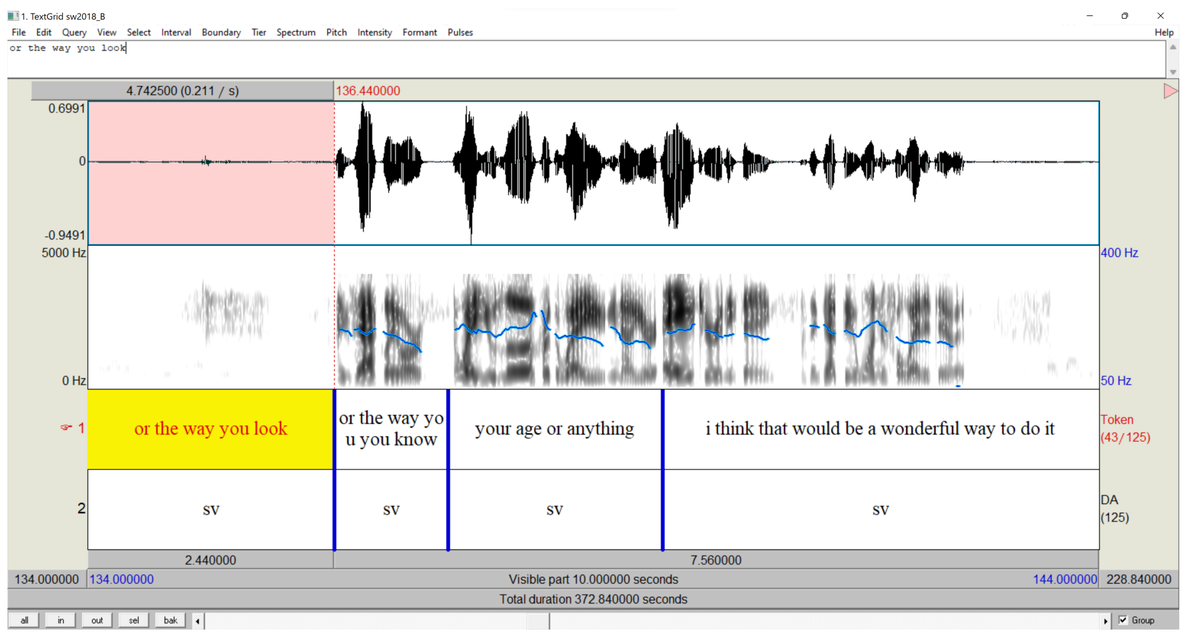
Re-alignment Methods

Step 1: Dialog Parsing

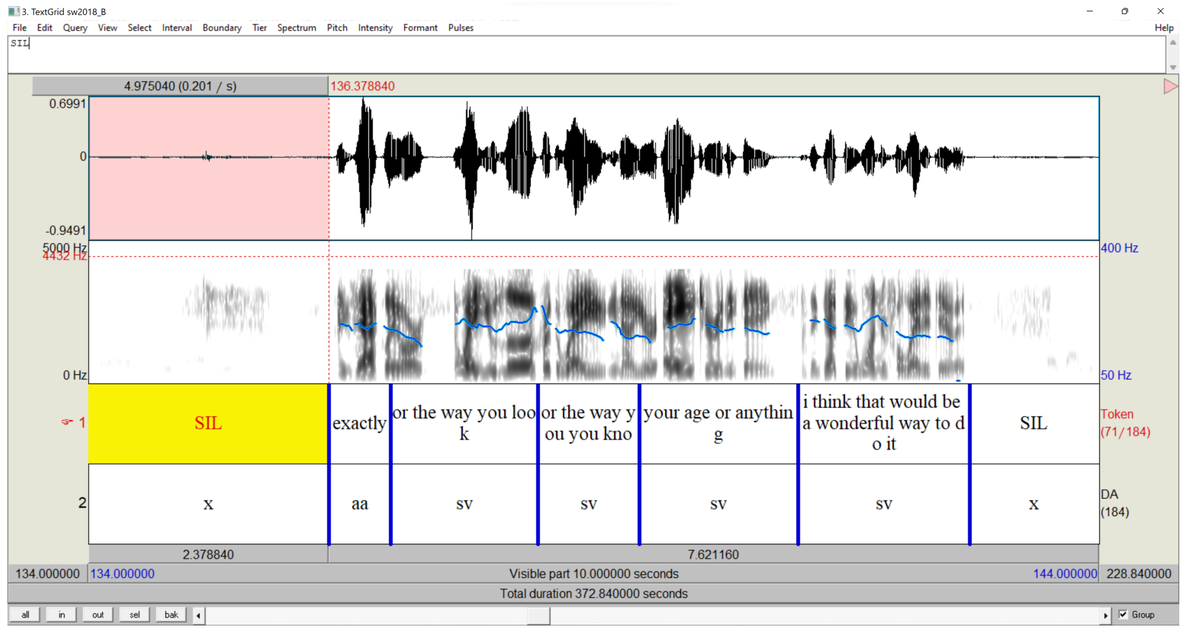
- parse 1155 dialogs into TextGrid files
- 642 NXT-format XML files [6]
- 513 forced alignment with *aeneas* library

Step 2: Manual Correction

- adjust timestamps
- correct transcripts
- speaker overlap: "`SIL`"
- laughters: "`<laughter>`" tokens



(a) Automatic alignment.



(b) Automatic alignment + manual correction.

Figure 1. A section of a SwDA transcript in the Praat interface (a) before and (b) after manual correction of the automatic alignment generated by *aeneas*. Praat allows aligners to view the waveform and spectrogram of the speech signal (top two sections of display) and a TextGrid transcript (bottom section of display) simultaneously.

Results

Re-alignment Progress

Our Re-Aligned Switchboard Dialog Act (RASwDA) corpus currently consists of **537.5** manually re-aligned and validated conversations (1075 single speaker transcripts) from the 1155 SwDA conversations (Table 1).

DA	Description	Count (Full)	% (Full)	Count (RASwDA)	% (RASwDA)
sd	Statement-non-opinion	75145	34.26	32406	24.53
b	Acknowledge (Backchannel)	38298	17.46	16297	12.34
sv	Statement-opinion	26428	12.05	11762	8.90
%	Abandoned, Turn-Exit, or Uninterpretable	15550	7.09	6729	5.09
aa	Agree/Accept	11133	5.08	4973	3.76
x	Non-verbal	3630	1.65	3591	2.6
qy	Yes-No-Question	4727	2.15	2053	1.55
ba	Appreciation	4765	2.17	1799	1.36
ny	Yes answers	3034	1.38	1252	0.95
fc	Conventional-closing	2582	1.18	1056	0.80
qw	Wh-Question	1979	0.90	874	0.66
nn	No answers	1377	0.63	595	0.45
bk	Response Acknowledge-ment	1306	0.60	555	0.42
h	Hedge	1226	0.56	507	0.38
qyd	Declarative Yes-No-Question	1219	0.56	472	0.36
bh	Backchannel in question form	1053	0.48	445	0.34
bf	Summarize/re-formulate	952	0.43	444	0.34
q	Quotation	983	0.45	427	0.32
fo_o_fw_"_by_bc	Other	883	0.40	408	0.31
na	Affirmative non-yes answers	847	0.39	351	0.27

Table 1. Comparison of the top 20 original SwDA DA counts ("Count (Full)") and our re-aligned corpus RASwDA DA counts ("Count (RASwDA)"). Full table showing all 42 DAs in paper.

Improvement on Dialog Act Classification (DAC) Task

Our model uses a convolutional neural network (CNN) and treats DAC as an image classification task on spectrograms of the speech signal (Figure 2).

We reached **59.53** accuracy on the validation set, an improvement of the state of the art model results [5], even with a much smaller training set. We believe that as we continue to build RASwDA by re-aligning the rest of the SwDA conversations, the model performance will further improve with a larger, more accurate dataset.

Model	[5]	Ours
Dataset	SwDA	RASwDA
Accuracy	56.97	**59.53**
Train	192,768	55,049
Validation	3,196	13,762
Test	4,088	–

Table 2. Dialog act classification (DAC) accuracy on speech from SwDA and RASwDA corpora, along with sizes of training, validation, and test splits in numbers of utterances.

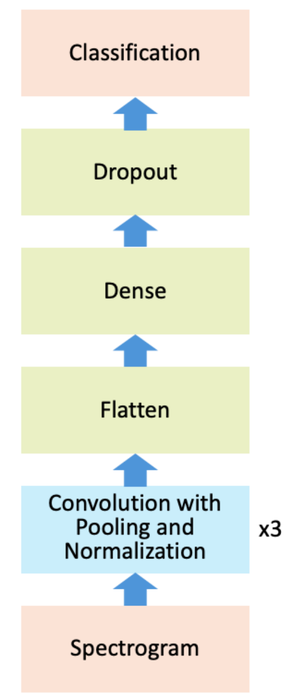


Figure 2. Model architecture

Conclusions

We have identified inaccuracies in the current automatic alignments of the Switchboard Dialog Act (SwDA) corpus and have undertaken a manual re-alignment process for a subset of **537.5** out of 1155 conversations. Our Re-Aligned Switchboard Dialog Act (RASwDA) subset has already demonstrated **improved performance** of state-of-the-art models on the dialog act classification task. We plan to continue the re-alignment process for the remainder of the SwDA corpus and make it publicly available for the wider speech community.

Acknowledgement

We express our gratitude to the students who contributed to this project by assisting with annotation and validation.

References

[1] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13," Tech. Rep. 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO, 1997.

[2] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 517–520 vol.1, 1992.

[3] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. Van Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?," *Lang. and speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[4] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Comput. linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[5] K. Wei, D. Knox, M. Radfar, T. Tran, M. Müller, G. P. Strimel, N. Susanj, A. Mouchtaris, and M. Omologo, "A neural prosody encoder for end-to-end dialogue act classification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7047–7051, IEEE, 2022.

[6] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, pp. 387–419, 2010.