Annotating Coherence Relations for Studying Topic Transitions in Social Talk

Alex Lưu
Brandeis University
alexluu@brandeis.edu

Sophia A. Malamud
Brandeis University
smalamud@brandeis.edu

Abstract

This study develops the strand of research on topic transitions in social talk which aims to gain a better understanding of interlocutors' conversational goals. Lưu and Malamud (2020) proposed that one way to identify such transitions is to annotate coherence relations, and then to identify utterances potentially expressing new topics as those that fail to participate in these relations. This work validates and refines their suggested annotation methodology, focusing on annotating most prominent coherence relations in face-to-face social dialogue. The result is a publicly accessible gold standard corpus with efficient and reliable annotation, whose broad coverage provides a foundation for future steps of identifying and classifying new topic utterances.[1]

1 Introduction

In natural language interactions, speakers' conversational goals determine the course of conversation. Conversational goals are therefore an essential component of any dialogue model that possesses a genuine capability of natural language understanding and reasoning. While this component is predefined or able to be inferred from visible linguistic content in the dialogue systems for task-oriented conversation, it is inadequately represented or missing in the current dialogue systems implemented for social talk. For instance, open-domain dialog systems that evolve from task-oriented dialog systems usually extend the space of information to be exchanged, e.g. in terms of intents and topics, and treat social talk as multi-domain task-oriented talk; while end-to-end trained neural chatbots focus more on the utterance generation task, and do not have any explicit representation of conversational goals.

To gain better insight into conversational goals of the interlocutors in social talk, Lưu and Malamud (2020) conducted a pilot annotation study of new-topic utterances (NTUs), which begin a new topic not related to the content of prior discourse. While such utterances are legitimate in social talk, current models of dialogue would treat them as incoherent conversational moves because these models only focus on utterances within a topically coherent discourse segment. By identifying NTUs and studying patterns of the disjunctive topic changes (DTCs) which are signaled by these utterances, Lưu and Malamud (2020) was able to introduce new sequence-based social intents that are absent in traditional taxonomies of speech acts but clearly reflect the goal-directed aspect of social talk, and therefore advocate the enrichment of dialogue models for social talk.

We build on (Lưu and Malamud, 2020) to evaluate and refine their annotation methodology for identifying NTU candidates. Here, we report on a full annotation project which features a publicly accessible dataset of face-to-face casual dialogues in American English, refined annotation guidelines, detailed analyses of the annotation process, and fully adjudicated annotation results which can serve as a test set for computational evaluation.

This paper is organized as follows. Section 2 provides an overview of related prior research. Section 3 presents our work constructing a gold standard corpus of discourse relation annotation for the purpose of identifying NTUs. Section 4 reports quantitative and qualitative analyses of our annotation process and methodology, while Section 5 concludes and presents a plan for future work.

[1]The live version of this publication is located at https://osf.io/7t4rf/.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

The 14th Linguistic Annotation Workshop, pages 174–179
Barcelona, Spain (Online), December 12, 2020.

2 Related Work

Riou (2015b) uses a mixed-methods approach to identify topic transitions in conversational interaction, including (1) the preliminary segmentation of conversation into turn-constructional units (TCUs), widely used in the Conversation Analysis framework as minimal interactional moves, and (2) the actual classification of each TCU into different categories such as topic continuity, stepwise topic transition, and disjunctive topic transition (i.e. DTC) (Riou, 2015a). The latter relies on the annotators' internal analysis rather than any explicit guidelines. Focusing on identifying DTCs, Lưu and Malamud (2020) propose a more formal approach which relies on well-tested coherence relation annotation to single out the NTU candidates as those utterances which do not bear any coherence relation to the content of prior discourse.

The annotation of coherence relations in (Lưu and Malamud, 2020) is different from previous attempts in two main aspects. First, as its ultimate goal is to identify NTUs, which do not participate in any coherence relations with prior discourse, the annotation aims at annotating coherence relations that cover as many utterances in the discourse as possible, rather than exhaustively annotating all available coherence relations of interest as in (Tonelli et al., 2010; Rehbein et al., 2016). Second, it relies on the superset of all well-known and tested coherence relations, rather than a specific taxonomy which may require ad hoc additions along the annotation process, as in (Xue et al., 2016; Yung et al., 2019).

It is noted that the pilot annotation in (Lưu and Malamud, 2020) does not fully comply with their proposed methodology. Instead of annotating targeted coherence relations from the ground up, they start with the Disco-SPICE corpus's *1,273* annotated coherence relations defined in the early version of the Penn Discourse Treebank (PDTB) 3.0 scheme (Webber et al., 2016). That optimizes the workload but does not provide an accurate view of the proposed annotation process because the goals of two annotation projects are different (as mentioned above).

This work seeks to elaborate and enrich Lưu and Malamud (2020)'s methodology for marking coherence relations as a way to identify NTU candidates in social talk, making the following contributions:

- a full development cycle of a high-quality gold standard corpus, which is publicly accessible,
- quantitative and qualitative assessments of the reliability and validity of the annotation process,
- carefully refined annotation guidelines, which are appropriate for large-scale annotation projects.

3 Gold Standard Corpus Creation

3.1 Data, Data Format, and Annotation Platform

The annotation data includes the initial *15*-minute extracts of *7* casual dialogues from the Santa Barbara Corpus of Spoken American English (SBCSAE) (Du Bois et al., 2000), segmented into TCUs by Riou (2015a) (see detail in Table 1) and accompanied by audio files that can be conveniently browsed at TalkBank.org. These face-to-face conversations are a perfect complement to the telephone dialogues annotated by Lưu and Malamud (2020). Working on the version segmented by Riou (2015a), we adopt her assumption that TCU-based utterances are the smallest units for topics in the conversational discourse, and will be able to use our annotation results to assess the reliability of her methodology.

The data is converted into the FoLiA format and annotated in the FoLiA Annotation Tool (FLAT), as in (Lưu and Malamud, 2020). Being a rich XML-based format for multiple linguistic annotation types, FoLiA stands out for its human-readability and portability by allowing both in-line and stand-off annotation layers and including all of these layers in a single file (van Gompel and Reynaert, 2013). It is especially suitable for discourse-level annotations which are likely to be enriched with the annotation of more form-constrained linguistic layers to facilitate the research on the form-function interface. The well developed and maintained FoLiA libraries such as FoLiApy for Python also allow us to easily manipulate the data at any point of the corpus development cycle, making the workflow flexible and interactive. An example of this is our use of various Python scripts described at the end of Subsection 3.2.

3.2 Corpus Development Cycle

We use the Model-Annotate-Model-Annotate (MAMA) cycle, introduced in (Pustejovsky and Stubbs, 2012). Specifically, one pilot annotation round and two MAMA iterations (Table 1) were performed by a team of three annotators,who are rising junior Linguistics majors and native American English speakers.

- Pilot round: Each team member annotates 'SBC034Time' to acclimate to the task and estimate annotation rate (TCUs per hour). Each person then adjudicates the others' annotations.
- MAMA iterations: each of the remaining six dialogues is annotated by two team members and adjudicated by the third.

Dialogue name	**Number of TCUs**	**Average hours to annotate**	**Hours to adjudicate**
Pilot:	***379***	***6.00 ± 0.67***	***4.08 ± 1.22 (average)***
– SBC034Time	379	6.00 ± 0.67	4.08 ± 1.22 (average)
MAMA 1:	***391 (average)***	***5.33 ± 0.67***	***3.00***
– SBC007Tree	344	4.38 ± 0.63	4.00
– SBC005Book	393	5.38 ± 1.13	2.50
– SBC017Notions	437	6.25 ± 0.25	2.50
MAMA 2:	***567 (average)***	***6.67 ± 0.33***	***4.75***
– SBC047Lot	491	6.00 ± 0.00	4.75
– SBC043Spoonfuls	561	6.50 ± 0.50	3.00
– SBC006Cuz	648	7.50 ± 0.50	6.50
Total	**3253**	**42.00 ± 3.67**	**27.33 ± 1.22 (average)**

Table 1: Annotation and adjudication time

The accelerated annotation task yielded an average annotation rate of *77.45* TCU/hr, with natural improvement from the pilot (*63.17* TCUs/hr) to MAMA 1 (*73.38* TCU/hr) to MAMA 2 stage (*85* TCU/hr).

The annotators are asked to put themselves in the interlocutors' shoes and label the most easily recognizable coherence relation between each TCU, or a discourse segment containing this TCU, and its prior discourse, which can also consist of one or more TCUs (example in Table 2), using labels based on:

- the latest PDTB 3.0 taxonomy of discourse relations (Webber et al., 2019),
- semantic relations from (ISO, 2016) and (ISO, 2012), which model the interactive nature of dialogue in a finer-grained manner than those of PDTB 3.0, e.g. *feedback* or *prop Q-A* in Table 2b.

Utterance	**Simplified transcript**
1728-JIM	*And, so much of today's technology is soulless.*
1729-JIM	*And has nothing to do with peace, it has to do with, just generally, chewing up, you know, consumerism basically and,*
1730-MIC	*Mhm.*
1731-JIM	*chewing up chewing up new uh, chewing up the human experience, and turning it into, some kind of consumer need.*
1732-MIC	*Did you ever get into Tesla?*
1733-JIM	*Uh, just, ever so peripherally.*
1734-MIC	*He had a lot of real wacky ideas on big levels.*
1735-MIC	*He wanted a world power system, that you could um, tap into the air basically, and get power anywhere on earth.*

(a) Excerpt from 'SBC017Notions'

Relation	**Arg 1**	**Arg 2**
conjunction	1728	1729
feedback	1729	1730
entity-based	1729	1731
prop Q-A	1732	1733
entity-based	1732	1734
instantiation	1734	1735

(b) Annotated relations in (a)

Table 2: Annotation example

The annotators will skip less salient coherence relations which:

- are available between the TCU under consideration and other parts of prior discourse, or
- concurrently exist between the TCU under consideration and its counterpart in the annotated relations (as discussed in (Moore and Pollack, 1992; Rohde et al., 2018; Webber et al., 2019)).

This is because our goal is to identify NTUs - utterances that do not participate in any identifiable relation to their prior linguistic context, and therefore once we annotate even a single relation between two given discourse segments, we know that the utterances in them are not NTUs. Focusing on existence of relations rather than identity of these relations allows us to pursue this goal, and has additional methodological benefits. First, this reduces the annotators' cognitive load, enhances their concentration, and fosters their natural interpretation of the dialogues. Second, it provides valuable insight into the annotators' ranking of the coherence relations based on their salience. A possible trade-off are inter-annotator inconsistencies with respect to the identity of annotated relations, but this does not affect annotation consistency with respect to existence of relations between given utterances.

We never encountered a case when an annotator could not find a label for a coherence relation, which is evidence for the breadth of coverage of our predefined label set. We accelerated the annotation process with the following decisions:

- The annotators do not mark the actual discourse connectives or equivalent linguistic materials which lexicalize the annotated coherence relations.
- The label set is selected so that the two arguments of each coherence relation can be indexed chronologically, i.e. the second argument always appears after the first argument in the conversational flow, in an automatic manner, specifically:
  - we only use Level-2 discourse relations of PDTB 3.0 sense hierarchy, which are indirectional (symmetric)
  - we rely on the the chronological nature of the arguments of ISO-based semantic relations (e.g. answers always follow questions)

The refined annotation guidelines are publicly accessible at `https://alexluu.flowlu.com/hc/6/223--annotation-guidelines`.

Each input file for pair-wise adjudication is created as a spreadsheet using a Python script. Identical relations annotated by two different annotators are automatically accepted. For discrepancies, the adjudicator can agree or disagree on the relation labels or the argument spans or both. Finally, the adjudication results are reconstructed into the gold standard version of a FoLiA file by another Python script. The final gold standard corpus is publicly accessible at `https://alexluu.flowlu.com/hc/6/250--annotated-corpus`.

4 Annotation Analyses

4.1 Quantitative Analysis

Based on (Artstein and Poesio, 2008), we identify each utterance as an item (markable) for inter-annotator agreement (IAA) calculation wherein the categories which can be assigned to each item are:

- there is no coherence relation between the item and prior discourse.
- there is a coherence relation between the item and prior discourse which consists of two attributes: the first argument of the relation, and the label of the relation.

The pair-wise IAA statistics are presented in Table 3. The column '**Agree on arg 1**' shows the ratio of the number of relations agreeing on their first arguments to the total number of utterances, while the column '**Agree on arg 1** & **label**' shows the ratio of the number of identical relations to the total number of utterances. The final column displays the number of coherence relations accepted to the final gold standard corpus. It is worth noting that the IAA significantly improves between two MAMA cycles, which will become clearer in light of the qualitative analysis.

4.2 Qualitative Analysis

We conducted interviews and a post-project survey with the annotators. Based on these qualitative measures, the most confused pair of relations are *agreement* and *feedback* (audio helps, but not always). In general, *agreement* relates to a directive or the truth of a statement in the previous utterance, while *feedback* signals the success/failure of processing the message in the previous utterance, without signaling agreement. TCUs are generally good units for our annotation, but there are 'too short' cases which require the creation of multi-TCU arguments and therefore slow down the annotation process.

Dialogue name	Agree on arg 1	Agree on arg 1 & label	# of accepted relations
Pilot:	***0.527 ± 0.041 (avg)***	***0.365 ± 0.021 (avg)***	***279 ± 10 (avg)***
– SBC034Time	0.527 ± 0.041 (avg)	0.365 ± 0.021 (avg)	279 ± 10 (avg)
MAMA 1:	***0.527***	***0.334***	***911 (total)***
– SBC007Tree	0.753	0.491	303
– SBC005Book	0.354	0.204	266
– SBC017Notions	0.506	0.327	342
MAMA 2:	***0.624***	***0.400***	***1330 (total)***
– SBC047Lot	0.617	0.436	411
– SBC043Spoonfuls	0.652	0.387	461
– SBC006Cuz	0.603	0.384	458
		Total	**2520**

Table 3: Pair-wise IAA statistics.

The annotators rate their independence as very high, slightly decreasing throughout the corpus development cycle due to our weekly meetings and the fact that annotators of some files are adjudicators of others. The trade-off is that the annotators become more confident, performing less pure but higher quality annotation. They also feel more competent at adjudication, and notice more convergence in their teammates' annotations. The annotators report having a very positive experience and are willing to participate in future projects on dialogic discourse.

5 Conclusion and Future Work

In summary, we have constructed a gold-standard corpus of most salient discourse relations as a way to identify NTU candidates in conversation. Our accelerated annotation methodology and the use of established relation taxonomies allowed for efficient and reliable annotation process with broad coverage.

Our next steps are, first, to compare the topic transitions inferred from our annotation and Riou (2015a)'s work to evaluate her methodology; and second, to identify NTUs and patterns of DTCs to enrich the classification of NTUs in (Lưu and Malamud, 2020). We plan to use this new classification as a step towards developing a dialogue agent capable of true understanding of the flow of social talk.

Acknowledgements

We are extremely grateful to Marine Riou and John W Du Bois who gave us their full support for making our annotation publicly accessible. Our deepest gratitude goes to our well-rounded annotation team: Julia Kenneally, Tali Tukachinsky and Cole Peterson. Finally, we would like to thank our anonymous reviewers for their crystal clear and thoughtful feedback that made our revision process very efficient.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. 2000. Santa Barbara corpus of spoken American English. *CD-ROM. Philadelphia: Linguistic Data Consortium.*

2012. Language resource management – semantic annotation framework (SemAF) – part 2: Dialogue acts. Technical report, International Organization for Standardization.

2016. Language resource management – semantic annotation framework (SemAF) – part 8: Semantic relations in discourse, core annotation schema (DR-core). Technical report, International Organization for Standardization.

Alex Lưu and Sophia A. Malamud. 2020. Non-topical coherence in social talk: A call for dialogue model enrichment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 118–133, Online, July. Association for Computational Linguistics.

Johanna D. Moore and Martha E. Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. O'Reilly Media, Inc.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1039–1046, Portorož, Slovenia. European Language Resources Association (ELRA).

Marine Riou. 2015a. *The Grammar of Topic Transition in American English Conversation. Topic Transition Design and Management in Typical and Atypical Conversations (Schizophrenia)*. Ph.D. thesis, Université Sorbonne Paris Cité.

Marine Riou. 2015b. A methodology for the identification of topic transitions in interaction. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (16).

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267, Melbourne, Australia, July. Association for Computational Linguistics.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Maarten van Gompel and Martin Reynaert. 2013. FoLiA: A practical XML format for linguistic annotation – a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, Dec.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31, Berlin, Germany, August. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 annotation manual. Technical report, University of Edinburgh.

Nianwen Xue, Qishen Su, and Sooyoung Jeong. 2016. Annotating the discourse and dialogue structure of SMS message conversations. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 180–187, Berlin, Germany, August. Association for Computational Linguistics.

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25, Florence, Italy, August. Association for Computational Linguistics.