Interspeech 2018
2-6 September 2018, Hyderabad

# Coherence Models for Dialogue

*Alessandra Cervone, Evgeny A. Stepanov, Giuseppe Riccardi*

Signals and Interactive Systems Lab, DISI, University of Trento, Italy
{alessandra.cervone, evgeny.stepanov, giuseppe.riccardi}@unitn.it

## Abstract

Coherence across multiple turns is a major challenge for state-of-the-art dialogue models. Arguably the most successful approach to automatically learning text coherence is the entity grid, which relies on modelling patterns of distribution of entities across multiple sentences of a text. Originally applied to the evaluation of automatic summaries and the news genre, among its many extensions, this model has also been successfully used to assess dialogue coherence. Nevertheless, both the original grid and its extensions do not model intents, a crucial aspect that has been studied widely in the literature in connection to dialogue structure. We propose to augment the original grid document representation for dialogue with the intentional structure of the conversation. Our models outperform the original grid representation on both text discrimination and insertion, the two main standard tasks for coherence assessment across three different dialogue datasets, confirming that intents play a key role in modelling dialogue coherence.
**Index Terms**: dialogue systems, coherence models

## 1. Introduction

This work addresses the problem of automatic coherence assessment of dialogue. Coherence – what makes a text unified rather than a random group of sentences – is an essential property to pursue for a system aimed at conversing with humans. Nonetheless, producing coherent responses across conversation turns remains an open research problem for state-of-the-art (SoA) open-domain dialogue models [1, 2].

Furthermore, progresses in open-domain dialogue modelling are currently curbed by a lack of standardized automatic metrics to evaluate and compare conversational systems [3]. Most available automatic metrics for dialogue evaluation either rely on surface features such as the words used (e.g. BLEU [4]), try to replicate generic human judgments [5], or work only for task-based dialogue systems [6]. For evaluation, the field still relies heavily on user satisfaction, an expensive and time-consuming process which poses its own challenges given the subjectivity of human judgment. While coherence has been proposed multiple times as an important metric to evaluate open-domain dialogue, there have been only few studies on open-domain dialogue coherence assessment [7, 8, 9].

On the other hand, the Natural Language Processing (NLP) literature has made several attempts [10, 11] to formalize the notion of text coherence into *coherence models*. The entity grid, the most popular approach to coherence modelling in this community, proposes to represent documents according to the patterns of distribution of entities mentioned in the text across adjacent sentences [11]. Besides its correlations with human judgment, among the reasons behind the success of this approach is the fact that it is linguistically motivated, capturing important aspects of discourse coherence related to entities distribution [12, 13]. Since its original proposal, the entity grid has undergone multiple extensions and has been widely applied to different tasks such as text coherence rating, automatic summaries assessment and sentence ordering, among others [11, 14]. It has also been successfully applied to dialogue [15, 16], for example for chat disentanglement.

Being a local coherence model, i.e. modelling paragraphs internal coherence rather than the global coherence of the entire text, the extensions of the grid proposed for dialogue do not take into account one essential characteristic of dialogue coherence that has been studied for several years: its intentional structure.

Several theories studying dialogue coherence are indeed rooted on the idea of an internal structure given by participants' intents in a conversation [17, 18, 19, 20]. In many approaches, the basic units of these sequences are a variation of Dialogue Acts (DAs), a concept based on Speech Acts theory [21], that conveys the illocutionary function of an utterance in a conversation; and represents a formalized and generalized lexicon of speaker intents. Attempts to formalize computationally similar theoretical intuitions about dialogue coherence [22, 23] did not find wide-spread application, since they require extensive expertise and significant manual annotation effort.

We propose entity-grid inspired coherence models for dialogue augmented with intentional information, represented by DA transitions across turns. To the best of our knowledge, this work is the first to combine entity grid coherence models with DAs. We compare our models to the original entity grid on the two de-facto standard tasks for coherence, i.e. sentence (in our case turn) ordering discrimination and insertion. We perform our experiments on three publicly available datasets conveying different types of dialogue (task-based and open-domain) and DAs annotation schemes, namely Switchboard [24], AMI [25] and Oasis [26]. Our results show the crucial importance of the DA information for assessing dialogue coherence.

## 2. State of the art

The most fertile framework for local coherence modelling in text is arguably the *entity grid* [11]. As shown in Figure 1, this approach proposes to represent the structure of a document (in our case a dialogue) through a grid displaying transitions in the syntactic roles of entities (the heads of Noun Phrases (NP)) across neighbouring sentences in the text. In the grid, the rows represent subsequent sentences (turns in our case, as in [16]) while each entity is represented by a column. A grammatical role can be: subject ($S$), direct object ($O$) or neither ($X$), plus a symbol ($-$) to signal that an entity does not appear in that turn $t$. The assumption is that the grid topology of coherent texts exhibits certain regularities associated to the way entities are introduced and become the focus of the discourse. For example, in the case of the grid represented in Figure 1, Table A we can notice how the sentences are connected by the continuity of the entity "drugs" across different turns. If an entity appears more than once in the same turn the most prominent syntactic role is chosen ($S > O > X$).


10.21437/Interspeech.2018-2446

| | company | drugs | policy | convictions | clients |
|---|---|---|---|---|---|
| **t1** | S | X | - | - | - |
| **t2** | X | S | O | X | S |
| **t3** | - | - | - | - | - |
| **t4** | - | - | - | - | - |
| **t5** | - | X | - | - | - |

A: *Entity Grid*

| | company | drugs | policy | convictions | clients | *no_entities* |
|---|---|---|---|---|---|---|
| **da1** | qy | qy | - | - | - | - |
| **da2** | - | na | na | - | - | - |
| **da3** | - | - | - | - | - | sdê |
| **da4** | sd | sd | sd | sd | sd | - |
| **da5** | - | - | - | - | - | % |
| **da6** | - | - | - | - | - | qo |
| **da7** | - | - | - | - | - | nn |
| **da8** | - | sdê | - | - | - | - |

B: *Modified Grid*

***A.* t.1 da.qy** Well does [the company ]$_S$ you work for test for[drugs]$_X$?
***B.* t.2 da.na** Actually, they just recently started [a policy]$_O$ of testing [drugs]$_X$, which was kind of interesting,
***B.* t.2 da.sdê** because when I went to work for them, uh, they didn't do that
***B.* t.2 da.sd** but, uh, since then they've started a [drug]$_O$ testing [policy]$_O$, not because of their own, uh [convictions]$_X$, but because [the clients]$_S$ of [our company]$_X$ are requests that we do that.
***A.* t.3 da.%** Huh.
***B.* t.4 da.qo** How about you?
***A.* t.5 da.nn** Uh, no
***A.* t.5 da.sdê** we're not being tested for [drugs]$_X$ at all, uh

Figure 1: *Entity grid example (A) vs. our modified grid (B) for the extract of the Switchboard dialogue 002_4330 (on the right). Entities in the sentences are annotated with their syntactic role: subject (S), object (O) or neither (X). The Dialogue Act tags are directly taken from the SWBD DAMSL annotation: qy (yes-no-question), na (affirmative-non-yes-answers), sde (Statement expanding y/n answer), sd (statement-non-opinion), % (uninterpretable), qo (open-question), nn (no-answers).*

By computing the probabilities over all possible transitions of length *n* from one category to all others (thus $\{S, O, X, -\}^n$) we can turn this representation into a feature vector, similar to a language model over the entity tags, representing the syntactic role transitions of entities in the whole document. It is important to notice that the entity grid is not lexicalised, since this information is lost when creating the feature vectors.

In [11], the authors use these feature vectors to train a Support Vector Machine (using SVM$^{light}$ [27]) modelling coherence as a ranking problem and using as a training dataset a set of original documents as positive (coherent) examples, paired with a set of the same documents with the sentences randomly permuted as negative (incoherent) examples (a procedure called pairwise training). The authors also experiment with models using different degrees of saliency (entity frequency) and transitions lengths (between 2 and 4), and by employing coreference resolution systems to detect entity chains (however given the performances of SoA systems this addition does not provide improvements).

The algorithm proposed in [11] derives thus automatically an abstract representation for a text, with as the only requirement a syntactic parser and a dataset. Among the weak points of this framework, however, is the fact that it models only local coherence (patterns of distribution across adjacent sentences) and a data sparsity problem.

Over the years, the entity grid model inspired numerous extensions [28, 29, 14] and similar implementations. Some approaches [29], for example, augmented the model using the semantic relatedness of the entities but without much improvement. Others [14] showed the usefulness of incorporating entity–specific features such as named entity information and considering also nouns which do not head NPs (as in Figure 1, turn 2, where in the NP "a drug testing policy" we consider both "drug" and "policy" as entities).

The typical tasks on which local coherence models are currently evaluated are: sentence ordering *discrimination*, where the system needs to rank original documents higher than randomly permuted ones, and *insertion*, introduced by [14], where the system has to rank the position of a sentence removed from a document. The state of the art for these tasks was recently achieved by [30], which uses the entity grid as input to a Convolutional Neural Network. The authors report an accuracy of 88.69 (compared to 81.58 of the original grid with both head and non-head nouns) and an insertion score of 25.95 (compared to 22.13 of the same model). One of the advantages of the neural model compared to the original one is its ability to model long range entity transitions. Other recent works inspired by the entity grid include coherent paragraph generation [31], and applications to automated essay scoring [32] and neural stories text generation [33].

Entity-based local coherence models apply well to dialogue as is or with some extra features, but not DAs [15, 16]. Dialogue coherence has been explored outside of the entity grid approach as well [8, 7, 9]. In [7], the authors propose a semi-automatic approach to evaluate dialogue coherence using only DA and relying on turn level coherence ratings from multiple sources. To the best of our knowledge, the only approach that combines entity and DA information for dialogue coherence evaluation is [8], which did not utilize the entity grid and models coherence as a binary classification task on utterance pairs rather than the whole conversation.

| | **SWBD** | **AMI** | **Oasis** |
|---|---|---|---|
| # DA tags | 42 | 16 | 41 |
| Av. # tokens/turn | 13.1 | 15.1 | 10.6 |
| Av. # turns/dialogue | 109 | 86.4 | 10.6 |
| # Train dialogues | 740 | 356 | 191 |
| # Test dialogues | 231 | 111 | 59 |
| # Dev dialogues | 184 | 89 | 45 |

Table 1: *We report the count of Dialogue Act tags, the average number of tokens per turn, the average count of turns per dialogue and counts of our Training, test and Developments splits for our three datasets: SWBD, AMI and Oasis. The document counts are given for the original documents, therefore need to be multiplied times 20 (pairs) for the discrimination task and times 100 (10x10) for the insertion task.*

## 3. Methodology

Both the original and its SoA extensions for coherence assessment focus on modelling local (entity-based) coherence, which is a form of surface coherence of the text (cohesion in Pragmatics theory [34]). However we can easily imagine how the entity grid or its extensions would not capture the lack of

coherence in the following example:

A. Do you have dogs?
B. What is the average height of dogs?

In this case the text would be judged coherent given the continuation of the entity "dogs" across both turns. Nonetheless this example is incoherent because B does not answer A's question, but rather introduces an unrelated question.

In this work we augment the original entity grid document representation with a notion of global coherence, as provided by the intentional structure of the conversation in the form of Dialogue Acts. Our hypothesis is that DA information could improve coherence models performance on dialogue. This hypothesis is also motivated by the fact that syntactic roles might no be so prominent or reliable when transferred to the spoken dialogue domain, since for some dialogue types turns tend to be quite short and syntactic parsers are not very robust when there is no punctuation.

In order to test our hypothesis, we experiment with various grid constructions in order to find the best way to combine the DAs information with the original representation. For clarity, we follow a template <row>-Grid:<cell> for naming our different document representations. In particular the <row> refers to text span (row in the grid) chosen, either the Turn (T) as in [16] or the text span of the DA (D); the <cell> refers to the category in the grid cells, either the syntactic role (*role*), the presence of the entity (*presence*, reducing the vocabulary to entities presence ($X$) or not ($-$) already proposed in [11]) or the DA tag (*DA*, which varies according the DA schema of each dataset). In the rest of the section we detail the document representations in our experiments.

**Baselines**: The baselines *T-Grid:roles* and *T-Grid:presence* replicate respectively the original entity grid in its all nouns variant (proposed by [14]) and a simplified version of the grid where the vocabulary is restricted to two items.
**D-Grid:role**: This variation differs from the *T-Grid:roles* only for the fact that the text span units are DAs, rather than turns, while the vocabulary is still composed by syntactic roles. The disadvantage of this representation is that it is more sparse than its preceding one, but it is able to capture in-turn entities transitions.
**D-Grid:DA**: In this variant the syntactic roles tags are substituted by the DA categories (according to the dataset's DA scheme). This is the modified grid shown in Figure 1, Table B. In this document representation an extra "no_entities" column is added to capture the DA tags where no entity is mentioned.
**Only DAs**: This text representation is the same as the previous one, with the difference that here all entities are dropped and we keep only one column with all the DAs.
**Combinations**: *T-Grid:presence + Only DAs* and *T-Grid:role + Only DAs* represent the combination of *Only DAs* with the two baselines by simply concatenating their feature vectors. These variations combine the entities and DAs feature vectors as two separate sources of information.

## 4. Experimental setup

**Tasks**: We evaluate our models on the sentence ordering *discrimination* task proposed in the original [11] and on the *insertion* task proposed in [14], which represent the standard evaluation tasks for coherence models. In order to ensure comparability across our experiments, when permuting the order in the documents, we always permute the entire turn (therefore multiple rows in case we have several DAs in the same turn) and the same permutations are kept across all settings.

The first task, discrimination, is usually evaluated as accuracy of the model in ranking the original text higher than a permuted one (we use 20 permutations per document following previous work [11, 14, 30]). In order to better analyse our results, we add to this metric two widely used ranking metrics, i.d. Mean Reciprocal Rank (MRR, the average of reciprocal ranks in a set of queries) and Precision at One (P@1, the ability of the model to rank the original higher than all the permutations). In both these metrics, instead of comparing the original document with each of its permutation we compare the rank of the original document to all its permutations at the same time.

On the other hand, the insertion task is evaluated as the average number of sentences per document inserted in the correct position (therefore the average of the P@1). For the insertion task, we randomly pick 10 turns per dialogue and insert each one in 10 random positions (for each dataset we used the same turns and positions to ensure intra-dataset comparability).
**Datasets**: In order to verify the robustness of our models across different DAs schemes and dialogue types, we perform all our experiments on three different publicly available datasets with DA annotation, namely BT Oasis[26], AMI[25] and the Switchboard Dialogue Act corpus [24] (SWBD). Table 1 shows some differences across the datasets.[1]

The dialogues in SWBD are open-domain telephone conversations. The individual turns tend to be quite long while the dialogues are the longest across the three datasets. For the DA categories we employ the 42 DAMSL ones. Oasis, on the other hand, is quite the opposite. A dataset of task-based conversations between clients and British Telecom help desk, here the turns tend to be quite short and the dialogues very short. AMI presents yet another type of dialogue data. Compared to the other datasets here the dialogues are between multiple speakers. In these dialogues participants were asked to discuss a project, so turns tend to be very long. This is also the dataset with the less rich annotation scheme compared to the previous two (only 16 DA categories).
**Parameters**: As in the original entity grid paper we test all our models using the preference kernel implemented in SVM$^{light}$ [27] with default parameters. We follow the default original grid parameters (saliency:1, transitions length:2) for all our experiments. This was done to ensure a fair comparison between the datasets with few entities and short dialogues (Oasis) and those with many turns and several entities (Switchboard, AMI). For preprocessing the text to extract noun phrases and their syntactic roles we use spacy [35].

## 5. Results

We report the results of our experiments in Table 2. To the model described in Section 3 we add a Random baseline, to give a measure of how the difficulty of both tasks vary across the datasets. To assess the respective significance of the coherence models, for discrimination accuracy and P@1 we use the McNemar test, while for discrimination MRR and the insertion Average P@1 we use Fisher's randomization test.

Regarding the *discrimination* task, the first thing to notice is how Only DAs, the model capturing DAs transitions without taking into account entities information, is a very competi-

[1]The code is available at: `https://github.com/alecervi/Coherence-models-for-dialogue`

| | SWBD | | | | AMI | | | | Oasis | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Discr. | | | Ins. | Discr. | | | Ins. | Discr. | | | Ins. |
| | **Acc.** | **MRR** | **P@1** | **Av. P@1** | **Acc.** | **MRR** | **P@1** | **Av. P@1** | **Acc.** | **MRR** | **P@1** | **Av. P@1** |
| Random | 50.00 | 16.98 | 4.76 | 8.70 | 50.00 | 18.93 | 6.31 | 9.44 | 50.00 | 17.39 | 5.08 | 9.16 |
| Only DAs | **99.76** | 98.76 | 97.80 | 45.45 | **98.78** | **95.27** | **92.79** | 30.75 | 91.53 | 68.47 | 54.24 | 41.44 |
| T-Grid:presence | 70.65 | 38.60 | 24.24 | 10.74 | 76.71 | 40.88 | 25.23 | 7.21 | 72.03 | 33.94 | 18.64 | 23.49 |
| T-Grid:role | 64.78 | 29.39 | 13.85 | 12.08 | 79.59 | 46.73 | 28.83 | 11.71 | 65.25 | 26.34 | 10.17 | 18.80 |
| D-Grid:role | 63.25 | 28.50 | 13.85 | 10.00 | 59.41 | 25.40 | 11.71 | 11.71 | 49.58 | 17.08 | 3.39 | 15.52 |
| D-Grid:DA | 99.57 | 97.36 | 95.67 | 38.79 | 95.41 | 83.02 | 75.68 | 19.25 | 87.80 | 57.64 | 40.68 | 28.96 |
| T-Grid:presence + Only DAs | **99.76** | 98.76 | 97.84 | **45.58** | 98.47 | 93.74 | 90.09 | 31.41 | **92.46** | 69.75 | **57.63** | **42.74** |
| T-Grid:role + Only DAs | 99.68 | **99.17** | **98.70** | 44.98 | 98.51 | 94.56 | 91.89 | **32.43** | 91.78 | **70.39** | **57.63** | 42.49 |

Table 2: *For each of the three datasets considered (SWBD, AMI and Oasis) we report results on the two tasks of Discrimination and Insertion. For Discrimination, we report the standard Accuracy (Acc.), plus Mean Reciprocal Rank (MRR) and Precision at one (P@1). For Insertion, we report the standard metric for this task, i.e. Precision at one (P@1) averaged for the dialogue.*

tive model across all the three datasets. Indeed the intentional structure information alone is so strong that on SWBD and AMI, the task of discriminating an original document from randomly shuffled re-orderings of the same document seems even too easy. With similar setup and data (also Switchboard but a different subset of dialogues with 505 original dialogues for training and 153 for testing) [16] reports an accuracy of 86.0 for its extended version of the grid. The strength of the intentional structure information is still prominent, but less visible in Oasis, where the dialogues are much shorter compared to the previous two datasets and it might be possible that random shuffling of turns might not disrupt the dialogue coherence so effectively.

In general, we notice the importance of DA information across the three datasets also for the rest of the proposed models for the discrimination task. As expected, the lowest results are achieved by the D-Grid:role model, which are still much better than the Random baseline. This model is similar to the original grid with the disadvantage of increasing the sparsity of entities.

The next lowest scores are then achieved by the T-Grid:presence and T-Grid:role. While the second performs better on AMI, where turns are the longest and we can expect sentence structure to be more complicated, the T-Grid:presence outperforms T-Grid:role both on Oasis and SWBD, confirming our hypothesis regarding the diminished importance of syntactic roles in dialogue. The next best model across all datasets for discrimination is D-Grid:DA with a large distance compared to T-Grid:presence and T-Grid:role.

The best performing models are the combinations, where the entity and DA information are encoded separately. These models achieve the best results on SWBD and Oasis, while their distance to Only DAs is not statistically significant on AMI.

The observations made on the discrimination task are reinforced on the *insertion* task. Only by looking at the performances (between 8.70 and 9.44) for the Random baseline, we notice how much the task is harder than the previous one (as mentioned in 2 the SOA in the Wall Street Journal is 25.95). The noticeable difference in the results for the T-Grid:presence, T-Grid:role compared to D-Grid:DA for insertion confirms once again how crucial is the intentional information. While also for insertion the intentional structure alone gives a very strong signal across all the datasets, the best results are achieved by combining the DAs with the entity information. This result is consistent with the nature of the task, where entity information could provide an important contribution to locating the exact place of a turn in the conversation. Also for this task, the syntactic role information yields the highest scores only for AMI, the dataset with the longest turns, while on SWBD and Oasis the best results are achieved by the simpler model – T-Grid:presence + Only DAs.

The Only-DAa model significantly outperforms the entity grid coherence models without DAs. However, while the models using the combination of entity grid and DAs (T-Grid:presence + Only DAs, T-Grid:role + Only DAs) yield better performance on SWDA and Oasis, overall their differences are not statistically significant.

## 6. Conclusions

In this paper, we applied the entity grid local coherence approach to dialogue. We experimented with different variations of its document representation in order to find the best way to augment it with participants' intents, an expression of global coherence and a signal which has been widely studied in dialogue to describe the structure of conversations. Our experiments confirm the crucial importance of the intentional structure for dialogue coherence, but also show how its combination with entity information could be useful for harder tasks connected to dialogue coherence, such as insertion.

Furthermore, our experiments show how the task of sentence ordering discrimination might be too easy on dialogue, where the DAs already give a very strong signal. On the other hand, the task of insertion is by far more difficult. For future work, we plan to explore other tasks for coherence modelling that might be more useful for dialogue, such as automatic prediction of the next dialogue turn.

It is also important to notice that our proposals for document representation are independent of the Machine Learning models employed. They could therefore be used, for example, in combination with a CNN as implemented in [30]. Another application we foresee for these models is to be used in the reward function for training dialogue systems in a Reinforcement Learning setting. Moreover, it is worth noticing that our experiments were performed using gold DAs. One of the first future experiments to perform would be to replicate the experiments with predicted DA labels, rather than gold ones to verify the robustness of the approach when using a DA tagger (the current approaches to DA tagging on Switchboard report accuracies above 75% [36, 37]). In such a setting, we imagine that the entities information might play even more important role in assessing dialogue coherence. Other possible directions include applying our coherence models to chat disentanglement, as well as the automatic evaluation of conversational agents' coherence.

# 7. References


[1] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of NAACL-HLT*, 2016, pp. 110–119.

[2] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1192–1202.

[3] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2122–2132.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[5] R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "On the evaluation of dialogue systems with next utterance classification," in *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, p. 264.

[6] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "Paradise: A framework for evaluating spoken dialogue agents," in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997, pp. 271–280.

[7] S. Gandhe and D. Traum, "A semi-automated evaluation metric for dialogue model coherence," *Situated Dialog in Speech-Based Human-Computer Interaction*, p. 217, 2016.

[8] R. Higashinaka, T. Meguro, K. Imamura, H. Sugiyama, T. Makino, and Y. Matsuo, "Evaluating coherence in open domain conversational systems," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[9] A. Venkatesh, C. Khatri, A. Ram, F. Guo, R. Gabriel, A. Nagar, R. Prasad, M. Cheng, B. Hedayatnia, A. Metallinou, R. Goel, S. Yang, and A. Raju, "On evaluating and comparing conversational agents," in *NIPS 2017 Conversational AI workshop*, 2017.

[10] B. J. Grosz, S. Weinstein, and A. K. Joshi, "Centering: A framework for modeling the local coherence of discourse," *Computational linguistics*, vol. 21, no. 2, pp. 203–225, 1995.

[11] R. Barzilay and M. Lapata, "Modeling local coherence: An entity-based approach," *Computational Linguistics*, vol. 34, no. 1, pp. 1–34, 2008.

[12] A. K. Joshi and S. Kuhn, "Centered logic: The role of entity centered sentence representation in natural language inferencing." in *IJCAI*, 1979, pp. 435–439.

[13] T. Givón, "Beyond foreground and background," *Coherence and grounding in discourse*, vol. 11, pp. 175–188, 1987.

[14] M. Elsner and E. Charniak, "Extending the entity grid with entity-specific features," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 125–129.

[15] A. Purandare and D. J. Litman, "Analyzing dialog coherence using transition patterns in lexical and semantic features." in *FLAIRS Conference*, 2008, pp. 195–200.

[16] M. Elsner and E. Charniak, "Disentangling chat with local coherence models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 1179–1189.

[17] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *language*, pp. 696–735, 1974.

[18] H. Sacks and G. Jefferson, "Lectures on conversation," 1995.

[19] E. A. Schegloff, "Sequencing in conversational openings," *American anthropologist*, vol. 70, no. 6, pp. 1075–1095, 1968.

[20] E. A. Schegloff and H. Sacks, "Opening up closings," *Semiotica*, vol. 8, no. 4, pp. 289–327, 1973.

[21] J. L. Austin, *How to do things with words*. Oxford university press, 1975.

[22] B. J. Grosz and C. L. Sidner, "Attention, intentions, and the structure of discourse," *Computational linguistics*, vol. 12, no. 3, pp. 175–204, 1986.

[23] J. F. Allen and C. R. Perrault, "Analyzing intention in utterances," *Artificial intelligence*, vol. 15, no. 3, pp. 143–178, 1980.

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

[25] J. Carletta, "Announcing the ami meeting corpus," *The ELRA Newsletter 11(1), January-March, p. 3-5.*, 2006.

[26] G. Leech and M. Weisser, "Generic speech act annotation for task-oriented dialogues," in *Procs. of the 2003 Corpus Linguistics Conference, pp. 441Y446. Centre for Computer Corpus Research on Language Technical Papers, Lancaster University*, 2003.

[27] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[28] C. Guinaudeau and M. Strube, "Graph-based local coherence modeling." in *ACL (1)*, 2013, pp. 93–103.

[29] K. Filippova and M. Strube, "Extending the entity-grid coherence model to semantically related entities," in *Proceedings of the Eleventh European Workshop on Natural Language Generation*. Association for Computational Linguistics, 2007, pp. 139–142.

[30] D. T. Nguyen and S. Joty, "A neural local coherence model," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 1320–1330.

[31] J. Li and D. Jurafsky, "Neural net models of open-domain discourse coherence," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 198–209.

[32] Y. Farag, H. Yannakoudakis, and T. Briscoe, "Neural automated essay scoring and coherence modeling for adversarially crafted input," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, vol. 1, 2018, pp. 263–271.

[33] E. Clark, Y. Ji, and N. A. Smith, "Neural text generation in stories using entity representations as context," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, vol. 1, 2018, pp. 2250–2260.

[34] "Cohesion in english, author=Halliday, Michael Alexander Kirkwood and Hasan, Ruqaiya, year=1976, publisher=Routledge."

[35] M. Honnibal and M. Johnson, "An improved non-monotonic transition system for dependency parsing," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1373–1378. [Online]. Available: https://aclweb.org/anthology/D/D15/D15-1162

[36] Y. Ji, G. Haffari, and J. Eisenstein, "A latent variable recurrent neural network for discourse relation language models," in *Proceedings of NAACL-HLT*, 2016, pp. 332–342.

[37] S. Mezza, A. Cervone, G. Tortoreto, E. A. Stepanov, and G. Riccardi, "ISO-standard domain-independent dialogue act tagging for conversational agents," in *COLING*, 2018.