Dialogue-Based Relation Extraction

Dian Yu[1]† **Kai Sun**[2]† **Claire Cardie**[2] **Dong Yu**[1]
[1]Tencent AI Lab, Bellevue, WA
[2]Cornell University, Ithaca, NY
{yudian, dyu}@tencent.com, ks985@cornell.edu, cardie@cs.cornell.edu

Abstract

We present the first human-annotated dialogue-based relation extraction (RE) dataset DialogRE, aiming to support the prediction of relation(s) between two arguments that appear in a dialogue. We further offer DialogRE as a platform for studying cross-sentence RE as most facts span multiple sentences. We argue that speaker-related information plays a critical role in the proposed task, based on an analysis of similarities and differences between dialogue-based and traditional RE tasks. Considering the timeliness of communication in a dialogue, we design a new metric to evaluate the performance of RE methods in a conversational setting and investigate the performance of several representative RE methods on DialogRE. Experimental results demonstrate that a speaker-aware extension on the best-performing model leads to gains in both the standard and conversational evaluation settings. DialogRE is available at https://dataset.org/dialogre/.

1 Introduction

Cross-sentence relation extraction, which aims to identify relations between two arguments that are not mentioned in the same sentence or relations that cannot be supported by any single sentence, is an essential step in building knowledge bases from large-scale corpora automatically (Ji et al., 2010; Swampillai and Stevenson, 2010; Surdeanu, 2013). It has yet to receive extensive study in natural language processing, however. In particular, although dialogues readily exhibit cross-sentence relations, most existing relation extraction tasks focus on texts from formal genres such as professionally written and edited news reports or well-edited websites (Elsahar et al., 2018; Yao et al., 2019; Mesquita et al., 2019; Grishman, 2019), while dialogues have been under-studied.

S1:	Hey Pheebs.
S2:	Hey!
S1:	Any sign of your **brother**?
S2:	No, but he's always late.
S1:	I thought you only met him once?
S2:	Yeah, I did. I think it sounds y'know big sistery, y'know, 'Frank's always late.'
S1:	Well relax, he'll be here.

	Argument pair	**Trigger**	**Relation type**
R1	(Frank, S2)	brother	per:siblings
R2	(S2, Frank)	brother	per:siblings
R3	(S2, Pheebs)	none	per:alternate_names
R4	(S1, Pheebs)	none	unanswerable

Table 1: A dialogue and its associated instances in DialogRE. S1, S2: anoymized speaker of each utterance.

In this paper, we take an initial step towards studying relation extraction in dialogues by constructing the first human-annotated dialogue-based relation extraction dataset, **DialogRE**. Specifically, we annotate all occurrences of 36 possible relation types that exist between pairs of arguments in the 1,788 dialogues originating from the complete transcripts of *Friends*, a corpus that has been widely employed in dialogue research in recent years (Catizone et al., 2010; Chen and Choi, 2016; Chen et al., 2017; Zhou and Choi, 2018; Rashid and Blanco, 2018; Yang and Choi, 2019). Altogether, we annotate 10,168 relational triples. For each (*subject, relation type, object*) triple, we also annotate the minimal contiguous text span that most clearly expresses the relation; this may enable researchers to explore relation extraction methods that provide fine-grained explanations along with evidence sentences. For example, the bolded text span *"brother"* in Table 1 indicates the PER:SIBLINGS relation (R1 and R2) between speaker 2 (S2) and *"Frank"*.

Our analysis of DialogRE indicates that the supporting text for most (approximately 96.0%) an-

† Equal contribution.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4927–4940
July 5 - 10, 2020.

notated relational triples includes content from multiple sentences, making the dataset ideal for studying cross-sentence relation extraction. This is perhaps because of the higher person pronoun frequency (Biber, 1991) and lower information density (Wang and Liu, 2011) in conversational texts than those in formal written texts. In addition, 65.9% of relational triples involve arguments that never appear in the same turn, suggesting that multi-turn information may play an important role in dialogue-based relation extraction. For example, to justify that *"Pheebs"* is an alternate name of S2 in Table 1, the response of S2 in the second turn is required as well as the first turn.

We next conduct a thorough investigation of the similarities and differences between dialogue-based and traditional relation extraction tasks by comparing DialogRE and the Slot Filling dataset (McNamee and Dang, 2009; Ji et al., 2010, 2011; Surdeanu, 2013; Surdeanu and Ji, 2014), and we argue that a relation extraction system should be aware of speakers in dialogues. In particular, most relational triples in DialogRE (89.9%) signify either an attribute of a speaker or a relation between two speakers. The same phenomenon occurs in an existing knowledge base constructed by encyclopedia collaborators, relevant to the same dialogue corpus we use for annotation (Section 3.2). Unfortunately, most previous work directly applies existing relation extraction systems to dialogues without explicitly considering the speakers involved (Yoshino et al., 2011; Wang and Cardie, 2012).

Moreover, traditional relation extraction methods typically output a set of relations only after they have read the entire document and are free to rely on the existence of multiple mentions of a relation throughout the text to confirm its existence. However, these methods may be insufficient for powering a number of practical real-time dialogue-based applications such as chatbots, which would likely require recognition of a relation at its first mention in an interactive conversation. To encourage automated methods to identify the relationship between two arguments in a dialogue as early as possible, we further design a new performance evaluation metric for the conversational setting, which can be used as a supplement to the standard F1 measure (Section 4.1).

In addition to dataset creation and metric design, we adapt a number of strong, representative learning-based relation extraction methods (Zeng et al., 2014; Cai et al., 2016; Yao et al., 2019; Devlin et al., 2019) and evaluate them on DialogRE to establish baseline results on the dataset going forward. We also extend the best-performing method (Devlin et al., 2019) among them by letting the model be aware of the existence of arguments that are dialogue participants (Section 4.2). Experiments on DialogRE demonstrate that this simple extension nevertheless yields substantial gains on both standard and conversational RE evaluation metrics, supporting our assumption regarding the critical role of tracking speakers in dialogue-based relation extraction (Section 5).

The primary contributions of this work are as follows: (**i**) we construct the first human-annotated dialogue-based relation extraction dataset and thoroughly investigate the similarities and differences between dialogue-based and traditional relation extraction tasks, (**ii**) we design a new conversational evaluation metric that features the timeliness aspect of interactive communications in dialogue, and (**iii**) we establish a set of baseline relation extraction results on DialogRE using standard learning-based techniques and further demonstrate the importance of explicit recognition of speaker arguments in dialogue-based relation extraction.

2 Data Construction

We use the transcripts of all ten seasons (263 episodes in total) of an American television situation comedy *Friends*, covering a range of topics. We remove all content (usually in parentheses or square brackets) that describes non-verbal information such as behaviors and scene information.

2.1 Relation Schema

We follow the slot descriptions[1] of the Slot Filling (SF) task in the Text Analysis Conference Knowledge Base Population (TAC-KBP) (McNamee and Dang, 2009; Ji et al., 2010, 2011; Surdeanu, 2013; Surdeanu and Ji, 2014), which primarily focuses on biographical attributes of person (PER) entities and important attributes of organization (ORG) entities. As the range of topics in *Friends* is relatively restricted compared to large-scale news corpora such as Gigaword (Parker et al., 2011), some relation types (e.g., PER:CHARGES, and ORG:SUBSIDIARIES) seldom appear in the texts. Additionally, we consider new relation types such as PER:GIRL/BOYFRIEND and PER:NEIGHBOR that

[1] http://surdeanu.info/kbp2014/def.php.

ID	Subject	Relation Type	Object	Inverse Relation	TR (%)
1	PER	per:positive_impression	NAME		70.4
2	PER	per:negative_impression	NAME		60.9
3	PER	per:acquaintance	NAME	per:acquaintance	22.2
4	PER	per:alumni	NAME	per:alumni	72.5
5	PER	per:boss	NAME	per:subordinate	58.1
6	PER	per:subordinate	NAME	per:boss	58.1
7	PER	per:client	NAME		50.0
8	PER	per:dates	NAME	per:dates	72.5
9	PER	per:friends	NAME	per:friends	94.7
10	PER	per:girl/boyfriend	NAME	per:girl/boyfriend	86.1
11	PER	per:neighbor	NAME	per:neighbor	71.2
12	PER	per:roommate	NAME	per:roommate	89.9
13	PER	per:children*	NAME	per:parents	85.4
14	PER	per:other_family*	NAME	per:other_family	52.0
15	PER	per:parents*	NAME	per:children	85.4
16	PER	per:siblings*	NAME	per:siblings	80.5
17	PER	per:spouse*	NAME	per:spouse	86.7
18	PER	per:place_of_residence**	NAME	gpe:residents_of_place	42.9
19	PER	per:place_of_birth**	NAME	gpe:births_in_place	100.0
20	PER	per:visited_place	NAME	gpe:visitors_of_place	43.0
21	PER	per:origin*	NAME		3.8
22	PER	per:employee_or_member_of*	NAME	org:employees_or_members	47.2
23	PER	per:schools_attended*	NAME	org:students	37.5
24	PER	per:works	NAME		27.0
25	PER	per:age*	VALUE		0.0
26	PER	per:date_of_birth*	VALUE		66.7
27	PER	per:major	STRING		50.0
28	PER	per:place_of_work	STRING		45.1
29	PER	per:title*	STRING		0.5
30	PER	per:alternate_names*	NAME/STRING		0.7
31	PER	per:pet	NAME/STRING		0.3
32	GPE	gpe:residents_of_place**	NAME	per:place_of_residence	42.9
33	GPE	gpe:births_in_place**	NAME	per:place_of_birth	100.0
34	GPE	gpe:visitors_of_place	NAME	per:visited_place	43.0
35	ORG	org:employees_or_members	NAME	per:employee_or_member_of	47.2
36	ORG	org:students*	NAME	per:schools_attended	37.5
37	NAME	unanswerable	NAME/STRING/VALUE		—

Table 2: Relation Types in DialogRE. Relation types with $\star$ represent the existing relation types defined in the TAC-KBP SF task, and we combine three SF fine-grained relation types about cities, states, and countries in a single relation type with $\star\star$. TR: Trigger ratio, representing the percentage of relational triples of a certain relation type that are accompanied by triggers.

frequently appear in *Friends*. We list all 36 relation types that have at least one relational instance in the transcripts in Table 2 and provide definitions and examples of new relation types in Appendix A.1.

2.2 Annotation

We focus on the annotation of *relational triples* (i.e., (*subject, relation type, object*)) in which at least one of the arguments is a named entity. We regard an uninterrupted stream of speech from one speaker and the name of this speaker as a *turn*.

As we follow the TAC-KBP guideline to annotate relation types and design new types, we use internal annotators (two authors of this paper) who are familiar with this task. For a pilot annotation, annotator A annotates relational triples in each scene in all transcripts and form a *dialogue* by extracting the shortest snippet of contiguous turns that covers all annotated relational triples and sufficient supportive contexts in this scene. The guidelines are adjusted during the annotation.[2] We prefer to use *speaker name* (i.e., the first word or phrase of a turn, followed by a colon) as one argument of a speaker-related triple if the corresponding full names or alternate names of the speaker name also appear in the same dialogue, except for relation PER:ALTERNATE_NAMES in which both mentions should be regarded as arguments. For an *argument pair* (i.e., (*subject, object*)), there may exist multiple relations between them, and we annotate all instances of all of them. For each

[2] As the pilot annotation only involves one annotator, we admit there may exist a certain degree of bias in defining new relation types and labeling argument pairs.

triple, we also annotate its *trigger*: the smallest extent (i.e., span) of contiguous text in the dialogue that most clearly indicates the existence of the relation between two arguments. If there exist multiple spans that can serve as triggers, we only keep one for each triple. For relation types such as PER:TITLE and PER:ALTERNATE_NAMES, it is difficult to identify such supportive contexts, and therefore we leave their triggers empty. For each relational triple, we annotate its inverse triple if its corresponding inverse relation type exists in the schema (e.g., PER:CHILDREN and PER:PARENTS) while the trigger remains unchanged.

In the second process, annotator B annotates the possible relations between candidate pairs annotated by annotator A (previous relation labels are hidden). Cohen's kappa among annotators is around 0.87. We remove the cases when annotators cannot reach a consensus. On average, each dialogue in DialogRE contains 4.5 relational triples and 12.9 turns, as shown in Table 3. See Table 1 for relational triple examples (R1, R2, and R3).

DialogRE	
Average dialogue length (in tokens)	225.8
Average # of turns	12.9
Average # of speakers	3.3
Average # of sentences	21.8
Average # of relational instances	4.5
Average # of no-relation instances	1.2

Table 3: Statistics per dialogue of DialogRE.

2.3 Negative Instance Generation, Data Split, and Speaker Name Anonymization

After our first round of annotation, we use any two annotated arguments associated with each dialogue to generate candidate relational triples, in which the relation between two arguments is unanswerable based on the given dialogue or beyond our relation schema. We manually filter out candidate triples for which there is "obviously" no relation between an argument pair in consideration of aspects such as argument type constraints (e.g., relation PER:SCHOOLS_ATTENDED can only exist between a PER name and an ORG name). After filtering, we keep 2,100 triples in total, whose two arguments are in "no relation", and we finally have 10,168 triples for 1,788 dialogues. We randomly split them at the dialogue level, with 60% for training, 20% for development, and 20% for testing.

The focus of the proposed task is to identify relations between argument pairs based on a dialogue, rather than exploiting information in DialogRE beyond the given dialogue or leveraging external knowledge to predict the relations between arguments (e.g., characters) specific to a particular television show. Therefore, we anonymize all speaker names (Section 2.2) in each dialogue and annotated triples and rename them in chronological order within the given dialogue. For example, S1 and S2 in Table 1 represent the original speaker names *"Rachel"* and *"Phoebe"*, respectively.

3 Data Comparisons and Discussions

3.1 Comparison Between DialogRE and SF

As a pilot study, we examine the similarities and differences between dialogue-based and traditional relation extraction datasets that are manually annotated. We compare DialogRE with the official SF (2013-2014) dataset (Surdeanu, 2013; Surdeanu and Ji, 2014) as 47.2% of relation types in DialogRE originate from the SF relation types (Section 2.1), and 92.2% of the source documents in it that contain ground truth relational triples are formally written newswire reports (72.8%) or well-edited web documents (19.4%) compared to the remaining documents from discussion fora. We show the relation distributions in DialogRE and SF in Figure 1 and Figure 2 (Appendix A.2), respectively. Half of the top ten relation types in DialogRE are newly defined (PER:GIRL/BOYFRIEND, PER:POSITIVE(NEGATIVE)_IMPRESSION, PER:FRIENDS, and PER:ROOMMATE), partially justifying the need for new relation types.

Argument Type: Based on the predefined SF and DialogRE relation types, a subject is expected to be an entity of type PER, ORG, or geo-political entity (GPE). Notably, subjects of most relational triples (96.8% vs. 69.7% in the SF dataset) in DialogRE are person names. The coarse-grained object type is entity, string, or value (i.e., a numerical value or a date). As shown in Table 4, we observe that a higher proportion (80.1%) of objects are entities in DialogRE compared to that in SF (65.3%).

	DialogRE	**SF**
Entity	80.1 (6,460)	65.3 (2,167)
String	18.9 (1,524)	25.4 (843)
Value	1.0 (84)	9.2 (306)

Table 4: Coarse-grained object type distributions (%) of DialogRE and SF with frequencies in brackets.

In particular, the subjects of 77.3% of relational triples are speaker names, and more than 90.0% of relational triples contain at least one speaker argument. The high percentage of "speaker-centric" relational triples and the low percentage of ORG and GPE arguments in DialogRE is perhaps because the transcripts for annotation are from a single situation comedy that involves a small group of characters in a very limited number of scenes (see more discussions in Section 5.3).

Distance Between Argument Pairs: It has been shown that there is a longer distance between two arguments in the SF dataset (Surdeanu, 2013; Huang et al., 2017) compared to that in many widely used human-annotated relation extraction datasets such as ACE (Doddington et al., 2004) and SemEval (Hendrickx et al., 2010). However, it is not trivial to compute an accurate distance between two arguments in a dialogue, especially for cases containing arguments that are speaker names. We instead consider different types of distances (e.g., average and minimum) between two argument mentions in a dialogue. We argue that DialogRE exhibits a similar level of difficulty as SF from the perspective of the distance between two arguments. 41.3% of arguments are separated by at least seven words even considering the minimum distance, and the percentage can reach as high as 96.5% considering the average distance, contrast with 46.0% in SF (Huang et al., 2017) and 59.8% in a recently released cross-sentence relation extraction dataset DocRED, in which Wikipedia articles serve as documents (Yao et al., 2019). Note that the provenance/evidence sentences in SF and DocRED are provided by automated systems or annotators. Also, 95.6% of relational triples from an annotated subset of DialogRE (Section 5.2) require reasoning over multiple sentences in a dialogue, compared with 40.7% in DocRED (Table 7). See Figure 3 in Appendix A.3 for more details.

3.2 Comparison Between DialogRE and Existing Relational Triples

We also collect 2,341 relational triples related to *Friends*, which are summarized by a community of contributors, from a collaborative encyclopedia.[3] We remove triples of content-independent relation types such as DIRECTED_BY, GUEST_STARS, and NUMBER_OF_EPISODES.

[3]https://friends.fandom.com/wiki/Friends.

We find that 93.8% of all 224 relation types in these triples can be mapped to one of the 36 relation types in our relation schema (e.g., HUSBAND, EX-HUSBAND, and WIFE can be mapped to PER:SPOUSE) except for the remaining relatively rare or implicit relation types such as PROM_DATE and GENDER, and KISSED, demonstrating the relation schema we use for annotation is capable of covering most of the important relation types labeled by the encyclopedia community of contributors.

On the other hand, the relatively small number of the existing triples and the moderate size of our annotated triples in DialogRE may suggest the low information density (Wang and Liu, 2011) in conversational speech in terms of relation extraction. For example, the average annotated triple per sentence in DialogRE is merely 0.21, compared to other exhaustively annotated datasets ACE (0.73) and KnowledgeNet (Mesquita et al., 2019) (1.44), in which corpora are formal written news reports and Wikipedia articles, respectively.

3.3 Discussions on Triggers

As annotated triggers are rarely available in existing relation extraction datasets (Aguilar et al., 2014), the connections between different relation types and trigger existence are under-investigated.

Relation Type: In DialogRE, 49.6% of all relational triples are annotated with triggers. We find that argument pairs are frequently accompanied by triggers when (1) arguments have the same type such as PER:FRIENDS, (2) strong emotions are involved (e.g., PER:POSITIVE(NEGATIVE)_IMPRESSION), or (3) the relation type is related to death or birth (e.g., GPE:BIRTHS_IN_PLACE). In comparison, a relation between two arguments of different types (e.g., PER:ORIGIN and PER:AGE) is more likely to be implicitly expressed instead of relying on triggers. This is perhaps because there exist fewer possible relations between such an argument pair compared to arguments of the same type, and a relatively short distance between such an argument pair might be sufficient to help the listeners understand the message correctly. For each relation type, we report the percentage of relational triples with triggers in Table 2.

Argument Distance: We assume the existence of triggers may allow a longer distance between argument pairs in a text as they help to decrease ambiguity. This assumption may be empirically

validated by the longer average distance (68.3 tokens) between argument pairs with triggers in a dialogue, compared to the distance (61.2 tokens) between argument pairs without any triggers.

4 Task Formulations and Methods

4.1 Dialogue-Based Relation Extraction

Given a dialogue $D = s_1 : t_1, s_2 : t_2, \ldots, s_m : t_m$ and an argument pair (a_1, a_2), where s_i and t_i denote the speaker ID and text of the i^{th} turn, respectively, and m is the total number of turns, we evaluate the performance of approaches in extracting relations between a_1 and a_2 that appear in D in the following two settings.

Standard Setting: As the standard setting of relation extraction tasks, we regard dialogue D as document d. The input is a_1, a_2, and d, and the expected output is the relation type(s) between a_1 and a_2 based on d. We adopt F1, which is the harmonic mean of precision (P) and recall (R), for evaluation.

Conversational Setting: Instead of only considering the entire dialogue, here we can regard the first $i \leq m$ turns of the dialogue as d. Accordingly, we propose a new metric F1_c, the harmonic mean of conversational precision (P_c) and recall (R_c), as a supplement to the standard F1. We start by introducing some notation that will be used in the definition of F1_c. Let O_i denote the set of predicted relation types when the input is a_1, a_2, and the first i turns (i.e., $d = s_1 : t_1, s_2 : t_2, \ldots, s_i : t_i$). For an argument pair (a_1, a_2), let L denote its corresponding set of relation types that are manually annotated based on the full dialogue. R represents the set of 36 relation types. By definition, $O_i, L \subseteq R$. We define that auxiliary function $\jmath(x)$ returns m if x does not appear in D. Otherwise, it returns the index of the turn where x first appears.

We define auxiliary function $\imath(r)$ as: (i) For each relation type $r \in L$, if there exists an annotated trigger for r, $\imath(r) = \jmath(\lambda_r)$ where λ_r denotes the trigger. Otherwise, $\imath(r) = m$. (ii) For each $r \in R \backslash L$, $\imath(r) = 1$. We define the set of relation types that are evaluable based on the first i turns by E_i:

$$E_i = \{r \,|\, i \geq \max\{\jmath(a_1), \jmath(a_2), \imath(r)\}\} \quad (1)$$

The interpretation of Equation 1 is that given d containing the first i turns in a dialogue, relation type r associated with a_1 and a_2 is evaluable if a_1, a_2, and the trigger for r have all been mentioned in d. The definition is based on our assumption that we can roughly estimate how many turns we require to predict the relations between two arguments based on the positions of the arguments and triggers, which most clearly express relations. See Section 5.2 for more discussions.

The conversational precision and recall for an input instance D, a_1, and a_2 are defined as:

$$\text{P}_\text{c}(D, a_1, a_2) = \frac{\sum_{i=1}^{m} |O_i \cap L \cap E_i|}{\sum_{i=1}^{m} |O_i \cap E_i|} \quad (2)$$

$$\text{R}_\text{c}(D, a_1, a_2) = \frac{\sum_{i=1}^{m} |O_i \cap L \cap E_i|}{\sum_{i=1}^{m} |L \cap E_i|} \quad (3)$$

We average the conversational precision/recall scores of all instances to obtain the final conversational precision/recall.

$$\text{P}_\text{c} = \frac{\sum_{D', a_1', a_2'} \text{P}_\text{c}(D', a_1', a_2')}{\sum_{D', a_1', a_2'} 1} \quad (4)$$

$$\text{R}_\text{c} = \frac{\sum_{D', a_1', a_2'} \text{R}_\text{c}(D', a_1', a_2')}{\sum_{D', a_1', a_2'} 1} \quad (5)$$

and $\text{F1}_\text{c} = 2 \cdot \text{P}_\text{c} \cdot \text{R}_\text{c} / (\text{P}_\text{c} + \text{R}_\text{c})$.

4.2 Baselines

Majority: If a given argument pair does not appear in the training set, output the majority relation type in the training set as the prediction. Otherwise, output the most frequent relation type associated with the two arguments in the training set.

CNN, LSTM, and BiLSTM: Following previous work (Yao et al., 2019), we adapt three baselines (Zeng et al., 2014; Cai et al., 2016) that use different document encoders. We refer readers to Yao et al. (2019) for more details.

BERT: We follow the framework of fine-tuning a pre-trained language model on a downstream task (Radford et al., 2018) and use BERT (Devlin et al., 2019) as the pre-trained model. We concatenate the given d and (a_1, a_2) with classification token `[CLS]` and separator token `[SEP]` in BERT as the input sequence `[CLS]`d`[SEP]`a_1`[SEP]`a_2`[SEP]`. We denote the final hidden vector corresponding to `[CLS]` as $C \in \mathbb{R}^H$, where H is the hidden size. For each relation type i, we introduce a vector $W_i \in \mathbb{R}^H$ and obtain the probability P_i of the existence of i between a_1 and a_2 based on d by $P_i = \text{sigmoid}(CW_i^T)$. The cross-entropy loss is used.

Method	Dev		Test	
	F1 (σ)	$F1_c$ (σ)	F1 (σ)	$F1_c$ (σ)
Majority	38.9 (0.0)	38.7 (0.0)	35.8 (0.0)	35.8 (0.0)
CNN	46.1 (0.7)	43.7 (0.5)	48.0 (1.5)	45.0 (1.4)
LSTM	46.7 (1.1)	44.2 (0.8)	47.4 (0.6)	44.9 (0.7)
BiLSTM	48.1 (1.0)	44.3 (1.3)	48.6 (1.0)	45.0 (1.3)
BERT	60.6 (1.2)	55.4 (0.9)	58.5 (2.0)	53.2 (1.6)
$BERT_S$	63.0 (1.5)	57.3 (1.2)	61.2 (0.9)	55.4 (0.9)

Table 5: Performance of relation extraction methods on DialogRE in both the standard and conversational settings.

$BERT_S$: We propose a modification to the input sequence of the above BERT baseline with two motivations: (1) help a model locate the start positions of relevant turns based on the arguments that are speaker names, and (2) prevent a model from overfitting to the training data. Formally, given an argument pair (a_1, a_2) and its associated document $d = s_1 : t_1, s_2 : t_2, \ldots, s_n : t_n$, we construct $\hat{d} = \hat{s}_1 : t_1, \hat{s}_2 : t_2, \ldots, \hat{s}_n : t_n$, where $\hat{s}_i$ is:

$$\hat{s}_i = \begin{cases} \texttt{[S}_1\texttt{]} & \text{if } s_i = a_1 \\ \texttt{[S}_2\texttt{]} & \text{if } s_i = a_2 \\ s_i & \text{otherwise} \end{cases} \quad (6)$$

where $\texttt{[S}_1\texttt{]}$ and $\texttt{[S}_2\texttt{]}$ are two newly-introduced special tokens. In addition, we define $\hat{a}_k$ ($k \in \{1, 2\}$) to be $\texttt{[S}_k\texttt{]}$ if $\exists i (s_i = a_k)$, and a_k otherwise. The modified input sequence to BERT is $\texttt{[CLS]}\hat{d}\texttt{[SEP]}\hat{a}_1\texttt{[SEP]}\hat{a}_2\texttt{[SEP]}$. In Appendix A.4, we investigate in three alternative input sequences. It is worth mentioning that a modification that does not disambiguate speaker arguments from other arguments performs substantially worse than the above speaker-aware modification.

5 Experiment

5.1 Implementation Details

CNN, LSTM, and BiLSTM Baselines: The CNN/LSTM/BiLSTM encoder takes as features GloVe word embeddings (Pennington et al., 2014), mention embeddings, and type embeddings. We assign the same mention embedding to mentions of the same argument and obtain the type embeddings based on named entity types of the two arguments. We use spaCy[4] for entity typing.

Language Model Fine-Tuning: We use the uncased base model of BERT released by Devlin et al. (2019). We truncate a document when the input sequence length exceeds 512 and fine-tune BERT using a batch size of 24 and a learning rate of 3×10^{-5} for 20 epochs. Other parameters remain unchanged. The embeddings of newly-introduced special tokens (e.g., $\texttt{[S}_1\texttt{]}$) are initialized randomly.

[4]https://spacy.io/.

5.2 Results and Discussions

We report the performance of all baselines in both the standard and conversational settings in Table 5. We run each experiment five times and report the average F1 and $F1_c$ along with standard deviation (σ). The fine-tuned BERT method already outperform other baselines (e.g., BiLSTM that achieves 51.1% in F1 on DocRED (Yao et al., 2019)), and our speaker-aware extension to the BERT baseline further leads to 2.7% and 2.2% improvements in F1 and $F1_c$, respectively, on the test set of DialogRE, demonstrating the importance of tracking speakers in dialogue-based relation extraction.

Conversational Metric: We randomly select 269 and 256 instances, which are associated with 50 dialogues from each of the dev and test sets, respectively. For each of relational instances (188 in total) that are previously labeled with triggers in the subsets, annotator A labels the smallest turn i^* such that the first i^* turns contain sufficient information to justify a relation. The average distance between i^* and our estimation $\max\{g(a_1), g(a_2), i(r)\}$ in Equation (1) (Section 4.1) is only 0.9 turn, supporting our hypothesis that the positions of arguments and triggers may be good indicators for estimating the minimum turns for humans to make predictions.

For convenience, we use BERT for the following discussions and comparisons.

Ground Truth Argument Types: Methods in Table 5 are not provided with ground truth argument types considering the unavailability of this kind of annotation in practical use. To study the impacts of argument types on DialogRE, we report the performance of four methods, each of which additionally takes as input the ground truth argument types as previous work (Zhang et al., 2017; Yao et al., 2019). We adopt the same baseline for a direct comparison

except that the input sequence is changed.

In **Method 1**, we simply extend the original input sequence of BERT (Section 4.2) with newly-introduced special tokens that represent argument types. The input sequence is `[CLS]` d `[SEP]` $\tau_1 a_1$ `[SEP]` $\tau_2 a_2$ `[SEP]`, where τ_i is a special token representing the argument type of a_i ($i \in \{1, 2\}$). For example, given a_1 of type PER and a_2 of type STRING, τ_1 is `[PER]` and τ_2 is `[STRING]`. In **Method 2**, we extend the input sequence of BERT_S with τ_i defined in Method 1 (i.e., `[CLS]` $\hat{d}$ `[SEP]` $\tau_1 \hat{a}_1$ `[SEP]` $\tau_1 \hat{a}_2$ `[SEP]`). We also follow the input sequence of previous single-sentence relation extraction methods (Shi and Lin, 2019; Joshi et al., 2020) and refer them as **Method 3** and **4**, respectively. We provide the implementation details in Appendix A.5. As shown in Table 6, the best performance achieved by Method 2 is not superior to that of BERT_S, which does not leverage ground truth argument types. Therefore, we guess that ground truth argument types may only provide a limited, if at all positive, contribution to the performance on DialogRE.

	Method 1	Method 2	Method 3	Method 4
Dev	60.6 (0.4)	**62.9** (1.2)	55.6 (2.4)	61.9 (1.4)
Test	59.1 (0.7)	**60.5** (1.9)	52.3 (3.2)	59.7 (0.6)

Table 6: Performance (F1 (σ)) comparison of methods with considering the ground truth argument types.

Ground Truth Triggers: We investigate what performance would be ideally attainable if the model could identify all triggers correctly. We append the ground truth triggers to the input sequence on the baseline, and the F1 of this model is 74.9%, a 16.4% absolute improvement compared to the BERT baseline. In particular, through the introduction of triggers, we observe a 22.9% absolute improvement in F1 on relation types whose inverse relation types are themselves (e.g., PER:ROOMMATE and PER:SPOUSE). These experimental results show the critical role of triggers in dialogue-based relation extraction. However, trigger identification is perhaps as difficult as relation extraction, and it is labor-intensive to annotate large-scale datasets with triggers. Future research may explore how to identify triggers based on a small amount of human-annotated triggers as seeds (Bronstein et al., 2015; Yu and Ji, 2016).

5.3 Error Analysis and Limitations

We analyze the outputs on the dev set and find that BERT tends to make more mistakes when there exists an asymmetric inverse relation of the relation to be predicted compared to those that have symmetric inverse relations. For example, the baseline mistakenly predicts S2 as the subordinate of S1 based on the following dialogue: *"... S2: Oh. Well, I wish I could say no, but you can't stay **my assistant** forever. Neither can you Sophie, but for different reasons. S1: God, I am so glad you don't have a problem with this, because if you did, I wouldn't even consider applying..."*. Introducing triggers into the input sequence leads to a relatively small gain (11.0% in F1 on all types with an asymmetric inverse relation) perhaps because inverse relation types share the same triggers (e.g., *"my assistant"* serves as the trigger for both PER:BOSS and PER:SUBORDINATE). One possible solution may be the use of directed syntactic graphs constructed from the given dialogue, though the performance of coreference resolution and dependency parsing in dialogues may be relatively unsatisfying.

A major limitation in DialogRE is that all transcripts for annotation are from *Friends*, which may limit the diversity of scenarios and generality of the relation distributions. It may be useful to leverage existing triples in knowledge bases (e.g., *Fandom*) for thousands of movies or TV shows using distant supervision (Mintz et al., 2009), considering the time-consuming manual annotation process. In addition, dialogues in *Friends* presents less variation based on linguistic features (Biber, 1991) than natural conversations; nonetheless, compared to other registers such as personal letters and prepared speeches, there are noticeable linguistic similarities between natural conversations and television dialogues in *Friends* (Quaglio, 2009).

6 Related Work

Cross-Sentence Relation Extraction Datasets Different from the sentence-level relation extraction (RE) datasets (Roth and Yih, 2004; Hendrickx et al., 2010; Riedel et al., 2010; Zhang and Wang, 2015; Zhang et al., 2017; Han et al., 2018), in which relations are between two arguments in the same sentence, we focus on cross-sentence RE tasks (Ji et al., 2011; Surdeanu, 2013; Surdeanu and Ji, 2014) and present the first dialogue-based RE dataset, in which dialogues serve as input contexts instead of formally written sentences or documents.

Task	style/source of doc	# rel	cross rate$^{\circ}$	# doc	# triples$^{\bullet}$
	—— distant supervision ——				
Peng et al. (2017)	written/PubMed	4	75.2	960,000	140,661
DocRED (Yao et al., 2019)	written/Wikipedia	96	n/a	101,873	881,298
T-REx (Elsahar et al., 2018)	written/Wikipedia	353	n/a	3 million	11 million
	—— human annotation ——				
BC5CDR (Li et al., 2016)	written/PubMed	1	n/a	1,500	2,434
DocRED (Yao et al., 2019)	written/Wikipedia	96	40.7	5,053	56,354
KnowledgeNet (Mesquita et al., 2019)	written/Wikipedia and others	15	n/a	4,991	13,425
DialogRE (this work)	**conversational**/Friends	36	**95.6**	1,788	8,068

Table 7: Statistics of publicly available cross-sentence relation extraction datasets ($\circ$: the percentage (%) of relational triples involving multiple sentences; $\bullet$: not include no-relation argument pairs).

We compare DialogRE and existing cross-sentence RE datasets (Li et al., 2016; Quirk and Poon, 2017; Yao et al., 2019; Mesquita et al., 2019) in Table 7. In this paper, we do not consider relations that take relations or events as arguments and are also likely to span multiple sentences (Pustejovsky and Verhagen, 2009; Do et al., 2012; Moschitti et al., 2013).

Relation Extraction Approaches Over the past few years, neural models have achieved remarkable success in RE (Nguyen and Grishman, 2015b,a; Adel et al., 2016; Yin et al., 2017; Levy et al., 2017; Su et al., 2018; Song et al., 2018; Luo et al., 2019), in which the input representation usually comes from shallow neural networks over pre-trained word and character embeddings (Xu et al., 2015; Zeng et al., 2015; Lin et al., 2016). Deep contextualized word representations such as the ELMo (Peters et al., 2018) are also applied as additional input features to boost the performance (Luan et al., 2018). A recent thread is to fine-tune pre-trained deep language models on downstream tasks (Radford et al., 2018; Devlin et al., 2019), leading to further performance gains on many RE tasks (Alt et al., 2019; Shi and Lin, 2019; Baldini Soares et al., 2019; Peters et al., 2019; Wadden et al., 2019). We propose an improved method that explicitly considers speaker arguments, which are seldom investigated in previous RE methods.

Dialogue-Based Natural Language Understanding To advance progress in spoken language understanding, researchers have studied dialogue-based tasks such as argument extraction (Swanson et al., 2015), named entity recognition (Chen and Choi, 2016; Choi and Chen, 2018; Bowden et al., 2018), coreference resolution (Chen et al., 2017; Zhou and Choi, 2018), emotion detection (Zahiri and Choi, 2018), and machine reading comprehension (Ma et al., 2018; Sun et al., 2019; Yang and Choi, 2019). Besides, some pioneer studies focus on participating in dialogues (Yoshino et al., 2011; Hixon et al., 2015) by asking users relation-related questions or using outputs of existing RE methods as inputs of other tasks (Klüwer et al., 2010; Wang and Cardie, 2012). In comparison, we focus on extracting relation triples from human-human dialogues, which is still under investigation.

7 Conclusions

We present the first human-annotated dialogue-based RE dataset DialogRE. We also design a new metric to evaluate the performance of RE methods in a conversational setting and argue that tracking speakers play a critical role in this task. We investigate the performance of several RE methods, and experimental results demonstrate that a speaker-aware extension on the best-performing model leads to substantial gains in both the standard and conversational settings.

In the future, we are interested in investigating the generality of our defined schema for other comedies and different conversational registers, identifying the temporal intervals when relations are valid (Surdeanu, 2013) in a dialogue, and joint dialogue-based information extraction as well as its potential combinations with multimodal signals from images, speech, and videos.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments and suggestions.

References

Heike Adel, Benjamin Roth, and Hinrich Schütze. 2016. Comparing convolutional neural networks to

traditional models for slot filling. In *Proceedings of NAACL-HLT*, pages 828–838, San Diego, CA.

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53, Baltimore, MD.

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Improving relation extraction by pre-trained language representations. In *Proceedings of AKBC*, Amherst, MA.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of ACL*, pages 2895–2905, Florence, Italy.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Kevin Bowden, Jiaqi Wu, Shereen Oraby, Amita Misra, and Marilyn Walker. 2018. SlugNERDS: A named entity recognition tool for open domain dialogue systems. In *Proceedings of LREC*, pages 4462–4469, Miyazaki, Japan.

Ofer Bronstein, Ido Dagan, Qi Li, Heng Ji, and Anette Frank. 2015. Seed-based event trigger labeling: How far can event descriptions get us? In *Proceedings of ACL-IJCNLP*, pages 372–376, Beijing, China.

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of ACL*, pages 756–765, Berlin, Germany.

Roberta Catizone, Alexiei Dingli, and Robert Gaizauskas. 2010. Using dialogue corpora to extend information extraction patterns for natural language understanding of dialogue. In *Proceedings of LREC*, pages 2136–2140, Valletta, Malta.

Henry Y Chen, Ethan Zhou, and Jinho D Choi. 2017. Robust coreference resolution and entity linking on dialogues: Character identification on tv show transcripts. In *Proceedings of CoNLL*, pages 216–225, Vancouver, Canada.

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of SIGDIAL*, pages 90–100, Los Angeles, CA.

Jinho D. Choi and Henry Y. Chen. 2018. SemEval 2018 Task 4: Character identification on multiparty dialogues. In *Proceedings of SemEval*, pages 57–64, New Orleans, LA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proeedings of NAACL-HLT*, pages 4171–4186, Minneapolis, MN.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of EMNLP-CoNLL*, pages 677–687, Jeju Island, Korea.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of LREC*, pages 837–840, Lisbon, Portugal.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frédérique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of LREC*, pages 3448–3452, Miyazaki, Japan.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of EMNLP*, pages 4803–4809, Brussels, Belgium.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of SemEval*, pages 33–38, Uppsala, Sweden.

Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning knowledge graphs for question answering through conversational dialog. In *Proceedings of NAACL-HLT*, pages 851–861, Denver, CO.

Lifu Huang, Avirup Sil, Heng Ji, and Radu Florian. 2017. Improving slot filling performance with attentive neural networks on dependency structures. In *Proceedings of EMNLP*, pages 2588–2597, Copenhagen, Denmark.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the TAC2011 Knowledge Base Population Track. In *Proceedings of TAC*, Gaithersburg, MD.

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC 2010 knowledge base population track. In *Proceedings of TAC*, Gaithersburg, MD.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Tina Klüwer, Hans Uszkoreit, and Feiyu Xu. 2010. Using syntactic and semantic based relations for dialogue act recognition. In *Proceedings of COLING*, pages 570–578, Beijing, China.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of CoNLL*, pages 333–342, Vancouver, Canada.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, pages 2124–2133, Berlin, Germany.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of EMNLP*, pages 3219–3232, Brussels, Belgium.

Fan Luo, Ajay Nagesh, Rebecca Sharp, and Mihai Surdeanu. 2019. Semi-supervised teacher-student architecture for relation extraction. In *Proceedings of the Third Workshop on Structured Prediction for NLP*, pages 29–37, Minneapolis, MN.

Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In *Proceedings of NAACL-HLT*, pages 2039–2048, New Orleans, LA.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the TAC 2009 knowledge base population track. In *Proceedings of TAC*, Gaithersburg, MD.

Filipe Mesquita, Matteo Cannaviccio, Jordan Schmidek, Paramita Mirza, and Denilson Barbosa. 2019. KnowledgeNet: A benchmark dataset for knowledge base population. In *Proceedings of EMNLP-IJCNLP*, pages 749–758, Hong Kong, China.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th ACL and the 4th IJCNLP of the AFNLP*, pages 1003–1011, Suntec, Singapore.

Alessandro Moschitti, Siddharth Patwardhan, and Chris Welty. 2013. Long-distance time-event relation extraction. In *Proceedings of the IJCNLP*, pages 1330–1338, Nagoya, Japan.

Thien Huu Nguyen and Ralph Grishman. 2015a. Combining neural networks and log-linear models to improve relation extraction. *arXiv preprint*, cs.CL/1511.05926v1.

Thien Huu Nguyen and Ralph Grishman. 2015b. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the First Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, CO.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition, linguistic data consortium. *Linguistic Data Consortium*.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, Doha, Qatar.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, New Orleans, LA.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of EMNLP-IJCNLP*, pages 43–54, Hong Kong, China.

James Pustejovsky and Marc Verhagen. 2009. SemEval-2010 task 13: Evaluating events, time expressions, and temporal relations (TempEval-2). In *Proceedings of SEW*, pages 112–116, Boulder, Colorado.

Paulo Quaglio. 2009. *Television dialogue: The sitcom Friends vs. natural conversation*, volume 36. John Benjamins Publishing.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of EACL*, pages 1171–1182, Valencia, Spain.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *Preprint*.

Farzana Rashid and Eduardo Blanco. 2018. Characterizing interactions and relationships between people. In *Proceedings of EMNLP*, pages 4395–4404, Brussels, Belgium.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML-PKDD*, pages 148–163, Barcelona, Spain.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of CoNLL at HLT-NAACL*, pages 1–8, Boston, MA.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. *arXiv preprint*, cs.CL/1904.05255v1.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. N-ary relation extraction using graph-state lstm. In *Proceedings of EMNLP*, pages 2226–2235, Brussels, Belgium.

Yu Su, Honglei Liu, Semih Yavuz, Izzeddin Gür, Huan Sun, and Xifeng Yan. 2018. Global relation embedding for relation extraction. In *Proceedings of NAACL-HLT*, pages 820–830, New Orleans, LA.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association of Computational Linguistics*, 7:217–231.

Mihai Surdeanu. 2013. Overview of the TAC2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *Proceedings of TAC*, Gaithersburg, MD.

Mihai Surdeanu and Heng Ji. 2014. Overview of the english slot filling track at the TAC2014 knowledge base population evaluation. In *Proceedings of TAC*, Gaithersburg, MD.

Kumutha Swampillai and Mark Stevenson. 2010. Inter-sentential relations in information extraction corpora. In *Proceedings of LREC*, pages 2637–2641, Valletta, Malta.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of SIGDIAL*, pages 217–226, Prague, Czech Republic.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of EMNLP-IJCNLP*, pages 5788–5793, Hong Kong, China.

Dong Wang and Yang Liu. 2011. A pilot study of opinion summarization in conversations. In *Proceedings of ACL*, pages 331–339, Portland, OR.

Lu Wang and Claire Cardie. 2012. Focused meeting summarization via unsupervised relation extraction. In *Proceedings of SIGDIAL*, pages 304–313, Seoul, South Korea.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of EMNLP*, pages 1785–1794, Lisbon, Portugal.

Zhengzhe Yang and Jinho D Choi. 2019. FriendsQA: Open-domain question answering on tv show transcripts. In *Proceedings of SIGDIAL*, pages 188–197, Stockholm, Sweden.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of ACL*, pages 764–777, Florence, Italy.

Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schtze. 2017. Comparative study of cnn and rnn for natural language processing. *arXiv preprint*, cs.CL/1702.01923v1.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. 2011. Spoken dialogue system based on information extraction using similarity of predicate argument structures. In *Proceedings of SIGDIAL*, pages 59–66, Portland, OR.

Dian Yu and Heng Ji. 2016. Unsupervised person slot filling based on graph mining. In *Proceedings of ACL*, pages 44–53, Berlin, Germany.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Proceedings of the AAAI Workshop on Affective Content Analysis*, pages 44–51, New Orleans, LA.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*, pages 1753–1762, Lisbon, Portugal.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING*, pages 2335–2344, Dublin, Ireland.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint*, cs.CL/1508.01006v2.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of EMNLP*, pages 35–45, Copenhagen, Denmark.

Ethan Zhou and Jinho D Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of COLING*, pages 24–34, Santa Fe, NM.

A Appendices

A.1 Definitions of New Relation Types

We follow the original guideline to annotate relation types in the TAC-KBP SF task (marked with $\star$) unless stated otherwise and define new relation types as follows except for self-explainable ones (e.g., PER:MAJOR, PER:FRIENDS, and PER:CLIENT). In this section, we keep the original speaker names in examples for better readability.

∘ **per:alternate_names***: Names used to refer a person that are distinct from speaker names or the first name mention in the given dialogue. It is possible to provide correct objects for this relation type without any contextual information such as triggers. Alternate names may include nicknames, first name, aliases, stage names, alternate transliterations, abbreviations, alternate spellings, full names, and birth names. However, if the full name mention appears first, we do not regard a first/last name alone as a valid value. An alternate name can also be a single word or a noun phrase.

∘ **per:positive_impression**: Have a positive impression (psychological) towards an object (e.g., a person, a book, a team, a song, a shop, or location). A named entity is expected here.

∘ **per:negative_impression**: Have a negative impression (psychological) towards an object. A named entity is expected here.

∘ **per:acquaintance**: A person one knows slightly (e.g., name), but who is not a close friend.

∘ **per:alumni**: Two persons studied in the same school, college, or university, not necessarily during the same period. Two persons can be in different majors. Classmates or batchmates also belong to this relation type.

∘ **per:boss**: In most cases, we annotate B as the boss of A when A directly reports to B and is managed by B at work. In the meantime, A is the subordinate of B. For example, we label (*"Rachel"*, per:boss, *"Joanna"*) and its corresponding trigger *"assistant"* based on dialogue D1.

D1	
Rachel:	Oh, uh, Joanna I was wondering if I could ask you something. There's an opening for an assistant buyer in Junior Miss...
Joanna:	Okay, but that would actually be a big step down for me.
Rachel:	Well, actually, I meant for me. The hiring committee is meeting people all day and...
Joanna:	Oh. Well, I wish I could say no, but you cant stay my assistant forever. Neither can you Sophie, but for different reasons.

∘ **per:girl/boyfriend**: A relatively long-standing relationship compared to PER:POSITIVE_IMPRESSION and PER:DATES, including but not limited to ex-relationships, partners, and engagement. The fact that two people dated for one or several times alone cannot guarantee that there exists a PER:GIRL/BOYFRIEND relation between them; we label PER:DATES for such an argument pair, instead.

∘ **per:neighbor**: A neighbor could be a person who lives in your apartment building whether they are next door to you, or not. A neighbor could also be in the broader sense of a person who lives in your neighborhood.

∘ **per:roommate**: We regard that two persons are roommates if they share a living facility (e.g., an apartment or dormitory), and they are not family or romantically involved (e.g., per:spouse and per:girl/boyfriend).

∘ **per:visited_place**: A person visits a place in a relatively short term of period (vs. PER:PLACE_OF_RESIDENCE). For example, we annotate (*"Mike"*, per:visited_place, *"Barbados"*) in dialogue D2 and its corresponding trigger *"coming to"*.

D2	
Phoebe:	Okay, not a fan of the tough love.
Precious:	I just can't believe that Mike didn't give me any warning.
Phoebe:	But he didn't really know, you know. He wasn't planning on coming to Barbados and proposing to me...
Precious:	He proposed to you? This is the worst birthday ever.

∘ **per:works**: The argument can be a piece of art, a song, a movie, a book, or a TV series.

∘ **per:place_of_work**: A location in the form of a string or a general noun phrase, where a person works such as *"shop"*.

∘ **per:pet**: We prefer to use named entities as arguments. If there is no name associated with a pet, we keep its species (e.g., dog) mentioned in a dialogue.

A.2 Relation Type Distribution

A.3 Distance Between Argument Pairs

A.4 Other Input Sequences

We also experiment with the following three alternative input sequences on the BERT baseline: (1) `[CLS]`$d^{\#}$`[SEP]`, (2) `[CLS]`$d^{\#}$`[SEP]`a_1`[SEP]`a_2`[SEP]`, and (3) `[CLS]`d''`[SEP]`, where $d^{\#}$ is obtained by

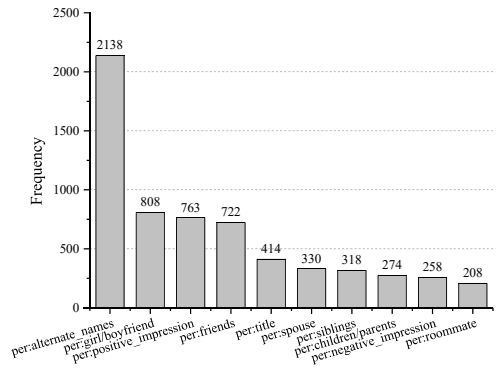

Figure 1: Relation type distribution in DialogRE.

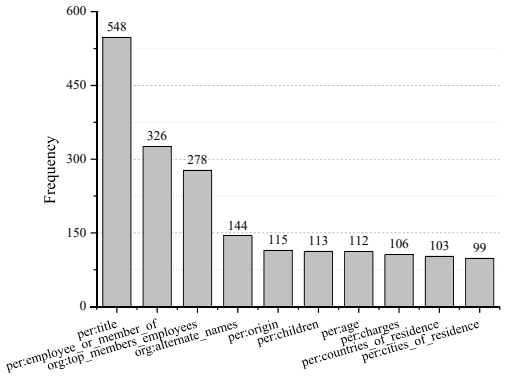

Figure 2: Relation type distribution in SF (2013-2014).

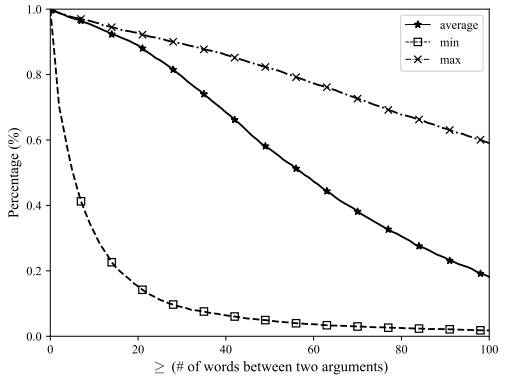

Figure 3: Number of words between two arguments within a dialogue in DialogRE.

replacing subject/object mentions in d with special tokens `[SUBJ]` and `[OBJ]`, and d'' is obtained by surrounding each mention of a_i ($i \in \{1, 2\}$) in d with special tokens `[A_i]` and `[/A_i]` (Baldini Soares et al., 2019). The F1 of them is 50.9%, 58.8%, and 57.9%, respectively, substantially lower than that of BERT$_S$ (61.2%).

A.5 Ground Truth Argument Type

Method 3 follows the input sequence employed by Joshi et al. (2020). Specifically, we replace the argument mentions in document d with newly-introduced special tokens that represent the subject/object and argument types. For example, if the subject type is PER and the object is STRING, we replace every subject mention in d with `[SUBJ-PER]` and every object mention with `[OBJ-STRING]`. Let d' denote the new document. The input sequence is `[CLS]` d' `[SEP]`.

Method 4 takes as input the sequence employed by Shi and Lin (2019). The input sequence is `[CLS]` d' `[SEP]` a_1 `[SEP]` a_2 `[SEP]`, where d' is defined in Method 3.