The Role of Common Ground for Referential Expressions in Social Dialogues

Jaap Kruijt
Vrije Universiteit Amsterdam
j.m.kruijt@vu.nl

Piek Vossen
Vrije Universiteit Amsterdam
p.t.j.m.vossen@vu.nl

Abstract

In this paper, we frame the problem of co-reference resolution in dialogue as a dynamic social process in which mentions to people previously known and newly introduced are mixed when people know each other well. We restructured an existing data set for the Friends sitcom as a coreference task that evolves over time, where close friends make reference to other people either part of their common ground (inner circle) or not (outer circle). We expect that awareness of common ground is key in social dialogue in order to resolve references to the inner social circle, whereas local contextual information plays a more important role for outer circle mentions. Our analysis of these references confirms that there are differences in naming and introducing these people. We also experimented with the SpanBERT coreference system with and without fine-tuning to measure whether preceding discourse contexts matter for resolving inner and outer circle mentions. Our results show that more inner circle mentions lead to a decrease in model performance, and that fine-tuning on preceding contexts reduces false negatives for both inner and outer circle mentions but increases the false positives as well, showing that the models overfit on these contexts[1].

1 Introduction

People that have a long-term relationship develop an effective way of communication which also targets the relationship as such. We call such conversations "social dialogues" and we expect that common ground plays an important role and has an impact in the way reference is made. As two conversation partners develop a closer bond, they form conventions in how they refer to individuals that are often part of their shared experiences, and these references may become vague and ambiguous to others (Hawkins et al., 2021). How present and important a particular individual is within the common ground not only influences the ambiguity of the references used, it also influences how readily they can be introduced into the conversation. While less popular or newly introduced individuals (outer circle) need a more elaborate and explicit reference, well-known individuals (inner circle) can be referenced with their name or a short and vague description, which may be difficult to infer for outsiders. For instance, someone's grandmother could be brought up with the reference 'Nana'. We would thus expect to see a difference between less important or unknown individuals on the one hand and important individuals on the other hand with respect to their *co-reference chains*.

When agents become more and more part of our lives, we can expect that they also build up long-term relationships with us, just like people do. An agent that needs to engage in social conversation with a human should therefore be sensitive to these changes in the way reference is made as the common ground grows. Some parts of this common ground, such as observations within the visual scene, can be established based on the context of the shared environment (Gergle et al., 2013). However, references to individuals in their shared experience (i.e. individuals that have been mentioned before) belong to the common ground that is based on the more long-term shared world context that needs to be retained across interactions, and which can also change over time.

In this paper, we report on a first analysis of how inner and outer circle people are referred to in social dialogues in which common ground plays a role and we test how sensitive existing co-reference resolution models are to this. To test this, we use a data set consisting of episodes of the *Friends* TV series that has been annotated with mentions of individuals (Choi and Chen, 2018). This data set contains both social dialogue and long-term connections between mentions that go above the level

[1]Our code is available at: https://github.com/cltl/inner-outer-coreference

Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2022) , pages 99–110
Gyeongju, Republic of Korea, October 16–17, 2022.

of a single document, and therefore it could serve as a useful simulation of the buildup of common ground over time.

Our main contributions are: 1) we frame the problem of resolving co-reference in dialogues as a dynamic process in which common ground plays a role, introducing the concepts of inner and outer circle references, 2) we provide new insight into the way inner and outer circle references are made by "friends" with a lot of common ground, 3) we test the sensitivity of machine learning models to (long-term) common ground in dialogue, 4) we restructure an existing data set of social dialogue in such a way that the existing temporal and topical relations between the conversations are maintained, which can be used for investigating the buildup of common ground and the development of conventions in referencing.

This paper is structured as follows: in section 2 we discuss related work, and present our motivation and problem statement. In section 3 we analyze the data on the difference in mention patterns for well-known and lesser-known individuals, and we discuss our approach to testing model performance with respect to references with common ground, and in section 4 we present the results of our tests and perform some error analysis. Finally, in section 5 we interpret our results and link them to the broader question of achieving common ground in human-agent communication.

2 Related Work

Common ground is what we call the established shared information that speakers rely on within a conversation (Stalnaker, 2002). It is essential to successful communication. Consequently, an agent that communicates with a human also needs to establish common ground to overcome mismatched representations of the world (Chai et al., 2014). In a process called *grounding*, this common ground also needs to be continually updated (Clark and Brennan, 1991).

Various research has been done on grounding in human agent-interaction, for instance on grounding in the visual scene in relation to tasks (Brawer et al., 2018; Roesler and Nowé, 2019; Shridhar and Hsu, 2018; Chai et al., 2014). Agents which develop this common ground have been shown to perform better on well-known tasks and also adapt better to new tasks (Brawer et al., 2018). However, in these task-oriented dialogues, the references tend to be explicit and refer to objects in the shared environment. In social dialogues, which are open, not task-oriented and between people that established a long social relationship, references become more vague quickly, and also refer to objects or individuals which are not present, but part of past (shared) experiences. This makes the references harder to latch on to, and means the agent must relate them to background knowledge rather than to what it sees in front of him.

The more interactions the agent has had with a particular human, the more shared experiences and background knowledge it can potentially rely on. However, in natural dialogue, as the number of shared experiences increases, the references become also more conventionalized, and as a result, more ambiguous to outsiders. Researchers have shown in simulations and experiments in human-human communication how this conventionalization occurs, and how it leads to more efficient but also more vague expressions over time that nonetheless remain understandable for the conversation partners who share the common ground (Hawkins et al., 2021; Shih et al., 2021; Haber et al., 2019). However, these simulations have been limited to task-oriented dialogue. We add to this research an analysis of social dialogues between humans.

Existing end-to-end coreference resolution models can achieve high scores on well-established datasets such as Ontonotes (Pradhan et al., 2012). However, these datasets do not relate well to our use-case of social dialogue. They consist of snippets of text from news articles, telephone conversations or talkshows, often more formal in nature, which likely makes the references more explicit. Most importantly, though, since the data does not contain a temporal aspect in which the various documents relate to each other, it cannot be used to examine how common ground builds up over time and how a model could utilise that common ground. The CODI-CRAC shared task (Khosla et al., 2021) is aimed at improving coreference resolution performance in dialogue. They also provide a selection of data sets consisting of dialogue, such as the Switchboard corpus (Holliman et al., 1992) and the Persuasion for Good dataset (Wang et al., 2019). However, these data sets are not ideal for our case either. Although they do consist of (social) dialogue, the conversations are between speakers who were previously unacquainted, and who do not share a common background which can be built upon in

the conversation. This is a crucial part of the phenomenon that we aim to investigate. Therefore, we take an existing dataset containing social dialogue and temporal relations between documents (Choi and Chen, 2018) and adapt it to analyze the differences in referencing of inner-circle mentions, which are part of the common ground, and outer circle mentions, which are only relevant within the surrounding context. We also test to what extent a state-of-the-art end-to-end coreference resolution model utilizes background knowledge and how it resolves complex third-person references. We hypothesize that the model will perform worse on references that require common ground. A model failing to detect a vague introduction for an otherwise well-known individual could also have problems further on in the conversation, as third-person pronouns referring to this individual are instead linked to a different individual. Concretely, this means that the higher the amount of inner circle references in the test set, the lower the overall model performance will be for that set. We further investigate whether this performance can be improved by increasing the background knowledge by training on preceding interactions.

We believe that it is valuable to examine common ground buildup over time in the context of coreference resolution. To the best of our knowledge, this is a new approach to the problem of coreference resolution. In the next section, we describe how we created the dataset and how we tested the model.

3 Method

For our experiment, we take the current state-of-the-art model in co-reference resolution, SpanBERT (Joshi et al., 2019, 2020). We use an implementation of this model by Xu and Choi (2020). This model predicts co-reference chains by calculating scores for pairs of mention and antecedent span representations which have been contextualized using BERT. We test this model in two ways. First, we run the pre-trained model on the new data set without fine-tuning. We do this to examine what performance the model can already achieve without knowing anything about the background knowledge except for what may have been learned from public sources such as Wikipedia during pretraining. Next, we fine-tune the model on data of previous conversations that likely contain common ground information and discourse contexts specific to sitcom characters. We fine-tune three models, one on a small, one on a medium and on a large portion of the previous conversations (see Table 3). We then examine the impact of the level of conversational context knowledge on performance. Specifically, we analyze the performance for inner- and outer circle mentions separately. Crucially, the fine-tuning is only done on data which precedes the test set chronologically, since we want to investigate the effect of simulated buildup of conversational context over time, which may also represent common ground.

3.1 Data analysis

We use the data set from SemEval 2018 task 4 (Choi and Chen, 2018) which consists of transcripts from the first two seasons of *FRIENDS*. This show contains social multi-party and dyadic dialogue. The data set is formatted according to the CONLL-2012 standards (Pradhan et al., 2012) and contains gold mentions. The original task was described as 'character identification', which combined features from entity linking and co-reference resolution in one task (Choi and Chen, 2018). However, the format of the data set works just as well for a pure co-reference resolution task. Since the original task was aimed at the identification of characters though, the gold mentions only contain references to people, and not objects or other types of named entities such as companies or countries. For our task this is ideal, since we are only interested in references to individuals.

In the show, the main characters know each other well, and as such have developed certain ways to refer to the people that are in their common ground. These people are also referenced more often throughout the show, requiring less introduction. For instance, *Judy Geller*, mother of two of the main characters, is referenced with *mom, ... (your) mother , ... (my) mother* (among others), over the course of several episodes. Meanwhile, a minor character called *Debra* is only mentioned in one single scene, and is referenced with the references *(a) woman - her - Debra - she - she*, in succession while she is the topic of the conversation.

The original character identification task contained a list of all the characters mentioned for the entity linking part of the task, 401 in total. We use this list to categorize all of the characters into 'inner circle' and 'outer circle'. We took the following

approach to selecting which characters belong to the inner circle: first, all of the six main characters of the show belong to the inner circle. Secondly, all of the family members of the six main characters also belong to the inner circle, since they can be mentioned by vague and ambiguous kinship terms which need background knowledge to be resolved (Kemp et al., 2018). All of the real-life famous persons which are mentioned in the show also belong to the inner circle, since they are well-known to all of the main characters in the show. However, they are mostly referred to by name. Lastly, we selected a few characters which have an entry on the Wikidata page for *FRIENDS* [2] relating them to the main characters. Since only the most important characters in the show have entries on Wikidata, this serves as a good indication that they are characters which belong to the shared common ground within the show. In total, we end up with 50 characters in the inner circle. The remaining 351 characters belong to the outer circle.

Inner circle characters are referenced a bit more in total than outer circle people even though there are more than 7 times more outer circle characters in the data set: on average inner circle characters are referenced 91.8 times and outer circle characters 6.6 times. The average number of variants (unique tokens) used to make reference is 16.4 for inner circle entities and 1 for outer circle entities. Furthermore, 112 outer circle characters are only mentioned once, whereas the lowest number of mentions for inner circle entities is 3 ('dad':2, 'he':1).

In Table 1, we show the distribution of the part-of-speech of the mentions of the inner and outer circle people. Proportionally, inner circle characters are more often referenced by name (NNP) than by pronoun (PR) as compared to outer circle references, whereas both are referenced equally by noun phrase (NN).[3] Apparently, the inner circle references by name seem to preempt the use of pronouns: less than 30% of the references to the inner circle is made using a pronoun, while almost 45% of the outer circle references is a pronoun.

In order to get insight in the discourse sequences of the references, we counted the part-of-speech sequence pairs as shown in Table 2. The rows represent the first mention in a pair and the columns the next mention, where NULL marks the cases of a first introduction of the referent in a scene. The table shows trivial dependencies such as pronouns are often followed by other pronominal references and hardly used as the first reference. Use of a pronoun as the first reference still makes sense however because we are dealing with a multimodal setting in which people can be introduced visually and referenced with deictic pronouns. This happens twice more often proportionally for the outer circle (20.19%) than for the inner circle (11.19%). Further comparing inner and outer circle references, we indeed see more NNP-NNP sequences for inner circle people and NN-PR sequences for outer circle people. Inner circles are introduced by name in 58.56% of the cases, which is followed by a name again in 52.98% of the cases, compared to 42.72% and 46.34% for outer respectively. We can expect that NNP-NNP sequences are easier to resolve for systems, which is advantageous for inner circle references. NN introductions happen more often for outer (37.09%) than for inner circles (30.25%), which are mostly followed by pronouns for outer (53.57%) and another NN for inner (44.5%).

These statistics suggest that for inner circle references it would be better for the model to focus on NNP-NNP patterns whereas for outer circle references NN-PR patterns are more important. We expect the former to be less and the latter to be more discourse structure dependent.

3.2 Experiment

The aim of our experiment is to measure differences in performance of coreference models resolving inner and outer circle co-reference relations, given the different ways of making reference, the degree they rely on background knowledge and the potential of the preceding discourse to learn patterns for resolving coreference relations. We expect that inner circle references by pronouns and noun phrases are more difficult to resolve than their counter parts for outer circle references. On the other hand, the more frequent use of names referring to inner circle entities could make it easier to resolve inner circle co-references. We expect to measure the impact of these differences in the performance on test sets with different ratios of inner and outer circle references. Furthermore, we want to measure the effect of using the preceding conversational context on the performance as well. To what extent does this context contain knowledge and information to resolve either inner or outer cir-

[2] https://www.wikidata.org/wiki/Q79784

[3] Other parts-of-speech mostly result from annotation errors

	NNP		NN		PR		OTHER		Total	Avg ment. per ref.
Inner (50)	1075	0,384	674	0,241	838	0,299	212	0,076	2799	55.98
Outer (351)	530	0,261	493	0,243	908	0,447	100	0,049	2031	5.78

Table 1: Distribution of part of speech for inner and outer circle mentions and the average number of mentions per referent. The parts-of-speech listed are names/proper nouns (NNP), common nouns (NN), pronouns (PR) and an OTHER category for parts-of-speech not belonging to one of the previous.

Inner circle	NNP	NN	PR
NULL	58.56	30.25	11.19
NNP	**52.98**	19.04	27.98
NN	24.35	44.50	31.15
PR	16.99	13.89	69.12
Outer circle	NNP	NN	PR
NULL	42.72	37.09	20.19
NNP	46.34	14.33	39.33
NN	11.07	35.36	**53.57**
PR	10.79	15.25	73.96

Table 2: Overview of proportion of part-of-speech coreference pairs for inner and outer circle mentions. Rows represent the first mention and the columns the following mention part-of-speech. NULL signifies there was no preceding mention.

Set	Small	Medium	Large
Train	S1E1...E7	S1E1...E17	S1E1...2E12
N° tokens	22211	54138	110074
Dev	S1E08	S1E18	S2E13
N° tokens	2876	2356	2118

Table 3: Size of each of the train sets and their respective development set

cle co-reference relations? Does this knowledge represent discourse structures, background knowledge or simply frequency of names?

For our experiment, we adapted the data set by removing all first-person and second-person pronoun mentions. We are only interested in the resolution of third-person references, which can become part of the common ground (inner) and thus require background knowledge to resolve or are introduced in the discourse itself (outer). First-person and second-person pronouns can be resolved within the discourse by linking them to the speaker or hearer, and are therefore not relevant to our experiment.

In the original data set, the train, development and test set were randomly distributed. However, we want to maintain the temporal structure within the data. Therefore, we made new train/development sets and selected two new test sets, where the latter are chosen to follow the training data in time.

To investigate the effect of varying the prominence of inner circle mentions in the test data on the model performance, we calculated the amount of mentions in the gold data to inner circle and outer circle characters per episode. We use episodes as a base length, because we find that in the TV show minor characters belonging to the outer circle are usually only mentioned in at most one episode, while major characters belonging to the inner circle are mentioned throughout the show, in multiple episodes or even seasons. After categorizing the mentions into inner and outer circle, we calculated the ratio of inner circle/outer circle mentions per episode.

For the test set, we selected one episode which contains roughly 4 times as many mentions to the inner circle as to the outer circle (S2E14), and one with a roughly equal amount of inner and outer circle mentions (S2E24). In the Appendix, we show details for both test sets in Tables 7 and 8, respectively. Both tables list the identifiers for the different characters sorted for the number of mentions. They show that S2E14 is dominated by inner circle mentions (top 5) and S2E24 has mixed mentions for inner and outer circles. On the other hand, the non-coreferential mentions (mentioned once) in S2E24 are all outer circle entities, while they are mixed in SE14. Next, we made three different sizes of train sets: one small, one medium-size, and one large, to vary the degree of background in the model representations. Table 3 shows the sizes of the three train sets.

We test four models on the two sets: the SpanBERT-large co-reference resolution model (Joshi et al., 2020) pre-trained by Xu and Choi (2020) without any higher-order inferencing, which has not been fine-tuned on the new data, and three fine-tuned SpanBERT-large models trained on the small, medium and large train sets respectively. For testing and fine-tuning, we follow the procedure as described by Xu and Choi (2020)[4].

[4] https://github.com/lxucs/coref-hoi

Metric	Pretrained			Finetuned-small			Finetuned-medium			Finetuned-large		
	R	*P*	*F1*	*R*	*P*	*F1*	*R*	*P*	*F1*	*R*	*P*	*F1*
MUC	37.28	95.65	53.65	27.11	88.88	41.55	38.98	88.46	54.11	44.06	70.27	**54.16**
B^3	31.79	95.48	**47.70**	17.59	90.00	29.42	28.26	88.90	42.88	37.06	56.70	44.82
CEAFM	38.04	92.10	**53.84**	27.17	83.33	40.98	35.86	86.84	50.76	33.69	64.58	44.28
CEAFE	32.56	71.64	**44.77**	19.81	54.49	29.06	25.16	69.20	36.90	22.34	67.02	33.51
BLANC	26.40	93.14	**41.03**	12.28	82.72	21.35	22.86	86.60	35.97	22.86	62.56	30.84
- Coref	28.57	88.88	**42.24**	13.09	78.57	22.44	25.59	81.13	38.91	27.97	37.90	32.19
- Non-coref	23.23	97.39	**38.82**	11.47	86.88	20.26	20.12	92.07	33.03	17.74	87.23	29.49

Table 4: Performance for S2E14 (4/1 ratio). Recall *R*, precision *P* and F1 score are reported for each model and for each metric. BLANC coreference and non-coreference scores are provided separately.

Metric	Pretrained			Finetuned-small			Finetuned-medium			Finetuned-large		
	R	*P*	*F1*	*R*	*P*	*F1*	*R*	*P*	*F1*	*R*	*P*	*F1*
MUC	50.00	88.67	63.94	41.48	90.69	56.93	62.76	86.76	72.83	72.34	80.95	**76.40**
B^3	43.53	86.56	57.93	26.53	88.10	40.78	48.22	76.67	59.20	58.76	71.52	**64.52**
CEAFM	49.64	83.33	62.22	32.62	83.63	46.93	49.64	76.92	60.34	56.02	75.96	**64.48**
CEAFE	42.95	65.12	**51.76**	17.91	70.17	28.54	32.95	67.34	44.25	31.10	73.08	43.63
BLANC	37.44	86.64	51.57	18.96	75.50	29.38	43.06	72.03	50.37	47.82	73.30	**56.52**
- Coref	43.50	78.46	55.97	29.70	86.15	44.18	60.21	62.70	61.43	57.29	64.86	**60.84**
- Non-coref	31.39	94.82	47.16	8.21	64.86	14.58	25.91	81.36	39.30	38.35	81.75	**52.21**

Table 5: Performance for S2E24 (1/1 ratio). Recall *R*, precision *P* and F1 score are reported for each model and for each metric. BLANC coreference and non-coreference scores are provided separately.

4 Results

4.1 Preprocessing

Before analyzing, we converted the .jsonlines output into CONLL format using a third-party script[5]. However, we adapted this script to accommodate for the fact that the models ignore the gold mentions, leading to very low precision. To accurately compare the model performance on inner and outer circle mentions, we need to only analyze the mentions that the model found that are also a gold mention.

4.2 Model performance

We evaluated the models using the official CONLL-2012 scorer (Pradhan et al., 2012)[6]. Performance for the four models on S2E14 (4/1 ratio of inner/outer circle mentions) is shown in Table 4 and their performance on S2E24 (1/1 ratio) in Table 5.

Our prediction for this experiment was the models would perform worse on S2E14, which has a higher ratio of inner circle mentions compared to outer circle mentions, than on S2E24. The difference between S1E14 (Table 4) and S1E24 (Table 5) confirms our hypothesis both without and after fine-tuning, with especially recall being higher for S2E24. The only outlier is the non-coref recall of 8.21 on S2E24 using the fine-tuned model with least training data.

Another prediction was that the pre-trained model would have more trouble with resolving references than the fine-tuned model due to lack of background knowledge and relevant discourse information. The results show that this not the case for S2E14 (4/1), where the pretrained model outperforms all other models on almost all metrics. For S2E24 (1/1) however, we see that best results (on most metrics) are obtained for the model fine-tuned with most data. Fine-tuning with more data shows a trend of increasing scores for S2E24 as the training data grows, with the highest scores for large. However, this is not the case for S2E14 (4/1), as the medium model outperforms the large model for e.g. BLANC. Apparently, what the model learns by fine-tuning is more relevant for the outer circle cases (S2E24) than for the inner circle cases (S2E14).

4.3 Error analysis

To find out whether the models had more difficulty with inner- or outer circle mentions, we need to break down the model performances to each of

[5] https://github.com/boberle/corefconversion

[6] https://github.com/conll/reference-coreference-scorers

these circles separately. This means we have to use an entity-based analysis of the clusters. We cannot use a link-based approach, because we would then have to evaluate on a subset of the data which corresponds to only the mentions for one of the circles. This would change the problem, since this excludes errors which link inner circle mentions to outer circle clusters and vice versa. Our entity-based analysis, which is based on false and true positives and false negatives for each gold entity cluster, is most similar to the CEAFE metric in approach. Furthermore, we want to investigate to what extent errors are made by mentions of different part-of-speech, especially names, noun phrases and pronouns.

Table 6 shows proportions of false positives and false negatives for all the mentions of inner and outer circle entities for the different models on the two test sets. We divided the error counts by the total number of inner or outer circle mentions, respectively, per test set. We also split the error proportion by part-of-speech for the inner and outer circle references. Note that S2E14 has 77 mentions of inner circle characters and 21 mentions of outer circle characters, and S2E24 has 67 mentions of inner circle characters and 83 mentions of outer circle characters[7].

4.3.1 S2E14 VS. S2E24

We first consider the errors averaged over all models and compare the performance across S2E14 and S2E24 to examine the effect of the test set on the model behaviour. For this we look at the Average subtotals, which show the proportions of false positives and false negatives averaged over all four models. Remember that S2E14 has a 4/1 ratio of inner to outer mentions, while S2E24 has a 1/1 ratio. The proportion of false positives is higher in S2E24 for both for the inner and outer mentions, while the proportion of false negatives is higher in S2E14. The differences in proportion are larger for the outer circle than for the inner circle. If we compare the best-scoring model for S2E14, Pretrained (PreT), with the best-scoring model for S2E24, Finetuned-large (FTlarge), we observe that most of the errors for Pre in S2E14 are false negatives, while for FTlarge in S2E24 most errors are false positives. Again, these differences are bigger for the outer circle than for the inner circle. This suggests that for S2E14, the main challenge for the model was to detect the references to the outer circle in between the more abundant inner circle mentions. This could cause it to miss more of the outer mentions, increasing the false negative rate. For S2E24, where inner and outer circle mentions were more evenly distributed, the challenge might have been not to mix up more vague references to outer circle mentions with the inner circle mentions in between, causing the model to add the mention to the wrong entity cluster. This hypothesis is strengthened by the fact that most false positive errors are made with pronoun mentions (in bold), which are inherently ambiguous and easy to misinterpret. Of course, there is also a potential influence of finetuning for FTlarge which is not present for the pretrained model. We will look at this later.

4.3.2 Inner circle VS. outer circle

The Average subtotals show for S2E14 that more errors are made for inner circle mentions compared to outer circle mentions. This holds for both false positives and false negatives, although the difference is larger for the false positives. For S2E24, the proportion of false negatives is roughly equal for the inner and outer mentions, and here the proportion of false negatives for the inner mentions is quite large compared to the outer mentions. In general then, the models seem to have more difficulty with the inner circle mentions.

As for the errors for each part of speech, there is no strong difference between the inner and outer circle in terms of which part of speech error is most prominent for each. We see that most false negative errors (in bold) are made for names (NNPs) for all models for S2E14, both for the inner and outer circle. S2E24 shows a pattern where most false negatives occur in names (NNPs) for the inner circle, whereas common nouns (NN) make up most errors for the outer circle. Pronouns (PR) make up most of the false positive errors for all models, both for S2E14 and S2E24 and both for inner and outer circle mentions. This last point makes intuitive sense, since pronouns are highly ambiguous. However, we don't see a difference between the inner and outer mentions, despite the higher relative amount of pronouns for outer mentions. It makes less sense that most false negatives occur for names. Despite the fact that associating names to characters should be relatively easy, they are apparently

[7]Note that the subtotals for Fpos and Fneg may add up to over 100%. This is because a single mention can be a false positive for one cluster and a false negative for another, meaning that this mention appears twice.

		S2E14										S2E24									
		PreT		FTsmall		FTmedium		FTlarge		Average		PreT		FTsmall		FTmedium		FTlarge		Average	
	PoS	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
Fpos	NNP	4%				4%		19%	5%	7%	1%	3%	**7%**	1%	2%		6%	7%	14%	3%	8%
	NN	**5%**		1%	5%	3%		16%	5%	6%	2%	1%		1%	1%	10%	11%	**28%**	23%	10%	9%
	PR			**8%**		**10%**		**35%**	**33%**	**13%**	**8%**	**25%**	4%	**6%**	**7%**	**33%**	**27%**	16%	**36%**	**20%**	**18%**
	Other							1%	0%	0%	0%									0%	0%
	Subtot.	9%		9%	5%	17%		71%	43%	27%	12%	30%	11%	9%	11%	43%	43%	52%	73%	34%	35%
Fneg	NNP	17%	29%	**30%**	**29%**	**21%**	**29%**	17%	**29%**	**21%**	**29%**	24%	12%	**43%**	16%	**34%**	13%	**33%**	5%	**34%**	11%
	NN	**22%**	29%	19%	19%	19%	19%	17%	14%	19%	20%	25%	**23%**	19%	**24%**	13%	**16%**	9%	**10%**	17%	**18%**
	PR	3%		5%	10%	6%	5%	3%	0%	4%	4%	9%	2%	7%	16%	1%	1%	3%	2%	5%	5%
	Other	16%		14%		14%		14%	0%	15%	0%	1%		1%		1%		1%		1%	0%
	Subtot.	57%	57%	69%	57%	61%	52%	51%	43%	59%	52%	60%	37%	72%	55%	51%	30%	46%	17%	57%	35%

Table 6: Breakdown of proportions of false positives (Fpos) and false negative (Fneg) differentiated per part-of-speech for S2E14 and S2E24 and for the different models (PreT = pretrained, FT = finetuned) and averaged over all models. 'In' refers to the errors for the inner circle mentions, whereas 'out' refers to those for the outer circle mentions.

still missed to a large extent in establishing coreference relations. Finally, S2E14 shows a remarkable proportion of false negative errors for inner circle entities with part-of-speech **Other**. As these are mostly annotation errors in this specific episode, none of the models seems to detect these cases as they are non-representative. These annotation errors were originally found mostly in the training data, but some of them ended up in our test set due to our re-organization of the sets.

4.3.3 Effect of finetuning

We now investigate to what extent fine-tuning on previous conversations can learn to detect the correct coreference relations and whether there is a difference for inner and outer circle references. Figures 1 to 6 in the Appendix contain bar plots showing the effect of finetuning on more data on the proportion of false positives and negatives for both the inner and outer circle, per test set and per part of speech. Overall, it looks like with more finetuning data, false negatives decrease both for the inner and outer circle, with the notable exceptions of outer circle NNP's in S2E14 (Figure 1) and inner circle NNP's in S2E24 (Figure 4). For the outer circle in S2E14, this could mean that the model over-fitted on inner circle names, and together with the relatively high amount of inner circle mentions this causes the model to ignore most names referring to the outer circle. However, this does not explain the high proportion of false negatives for the inner circle in S2E24. In general, we also see false positives increase with more finetuning data, especially for the inner circle. Together with the general decrease in false negatives, this indeed seems to suggest that the model is over-fitting on a part of the training data. In Table 1, we showed that most inner mentions are names, whereas most outer mentions are pronoun mentions. As mentioned above, it looks like the model tends to prefer inner circle names, which are more present in the training data. However, for pronouns we see a remarkable increase in false positives both for the inner and outer circle. For pronoun mentions, the models might learn a different preference than for NNP mentions which is more based on discourse features rather than individual characters. In general we believe the model tends more towards learning discourse features, because the graphs do not show a much stronger effect of over-fitting for the inner or outer circle. Note that the fine-tuned models generate more errors than the pretrained model on S2E24 in entity-based evaluation, which correspond to the CEAFE scores given earlier in Table 5 but not with the other metrics.

4.3.4 Error examples

An example of a false negative for a common noun (NN) referring to the inner circle is *actor* in S2E24. It occurs in the sentence *Mr. Beatty comes up to me and says 'good **actor**'...* and refers to the inner circle character Joey, who utters the sentence. It could be that the model mis-identified this reference because it is uttered in direct speech, which makes it unclear that the speaker is the intended referent. Another curious case for a name (NNP) referring to the inner circle concerns *Rachel* and her nickname *Rach* in S2E24, where the first three occurrences of *Rachel* / *Rach* in the scene are not added to the same cluster as the latter three occurrences of *Rach* by the large model. Between these two sets of occurrences, another person is referenced, which could explain why they were assumed to be disjoint by the model. Possibly, this an effect of window size, which makes the earlier references unavailable to the system. While the

sliding window is in principle a good method to constrain the context that the model takes into account, in this case it leads to errors which could have been avoided. Some false positives for pronouns are the result of an introduction in the visual scene (such as a speaker pointing at a character).

5 Discussion

Our results showed differences between models and across test sets with different ratios. All models perform lower on S2E14 with more inner circle references than on S2E24 with less. For most metrics, the pretrained model performed best on S2E14 (4/1) and the largest fine-tuned model on S2E24 (1/1) ratio, except for entity-based evaluations in which pretrained performed best on both. When breaking down the errors per part-of-speech and across inner and outer references, we found some patterns but it remains difficult to relate these to the different part-of-speech statistics observed. Many errors are made for outer circle names and in the case of S2E24 also for inner circle name mentions. Remarkable are the false positive (and to some extent false negative) errors in pronominal inner and outer circle mentions in S2E24. Since the false positives tend to increase with fine-tuning, we suspect that the fine-tuned models are over-fitting. Our experiments do not allow us to draw conclusions towards the potential of more knowledge-rich approaches that incorporate built-up common ground. This is partly because fine-tuned language models are not transparent to what knowledge is picked up from the preceding conversations.

Clearly, more research on the role of common ground in referencing in social dialogue is necessary. Most co-reference resolution models continue to be trained and tested on well-established data sets which are not useful for exploring this phenomenon. Although the data set of episodes of *FRIENDS* that we used in this paper has the necessary properties, it too has its drawbacks. Most of the dialogues are multi-party dialogues, whereas dyadic dialogue would be a more controlled setting in which to explore the buildup of common ground. The dialogues also partly rely on visual cues, which the model cannot rely on and for which the necessary metadata is not provided in the data set. Furthermore, the show is a sitcom, and the many quips might have a detrimental effect on the naturalness of the conversations. Therefore, we encourage the further development of more data sets of social dialogue with multiple interactions over time, based on a more natural setting or with fewer speakers involved in the conversation.

In this work, we have made a distinction between well-known 'inner circle' and lesser known 'outer circle' referents. We believe it is relevant to be aware of such a distinction in referencing, since people rely on the established references to the inner circle to create a bond and distinguish their shared social circle from the outside world. If we want systems to become a part of this shared social circle and develop their own bond with humans, they too need to learn this way of referencing, and in long-term interaction it could help them reinforce this bond and improve communicative efficiency and enjoyment on the part of the human.

In future work, we will further explore how the buildup of common ground influences referential expressions to well-known individuals over time in dyadic social dialogue. This will be done in an interactive setting, where an artificial agent engages in conversation with a human and can use visual cues and human feedback to improve its representation of the common ground. Due to the interactive nature of the dialogue, the model will not be a pure co-reference resolution model, but it will build upon properties of both co-reference resolution and entity linking models. In addition, we will use a more explicit modeling of common ground, and include more knowledge-rich features in our model.

6 Conclusion

In this paper, we framed the problem of resolving third-person references in social dialogues as a dynamic process in which common ground plays a role. We made a difference between inner and outer circle references and hypothesized that the former are more difficult to resolve, which was partially confirmed by the model performances on data with more and less inner circle references. Training models on preceding data did not show a corresponding increase in performance on inner circle references, indicating that such models do not acquire common ground knowledge, but did improve the performance for outer circle mentions. We propose that co-reference resolution models for social dialogue could benefit from a more knowledge-rich approach in order to better adjust to the common ground, which in turn facilitates the resolution of complex third-person references.

References

Jake Brawer, Olivier Mangin, Alessandro Roncone, Sarah Widder, and Brian Scassellati. 2018. Situated human–robot collaboration: predicting intent from grounded natural language. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 827–833. IEEE.

Joyce Y. Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '14, page 33–40, New York, NY, USA. Association for Computing Machinery.

Jinho D. Choi and Henry Y. Chen. 2018. SemEval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 57–64, New Orleans, Louisiana. Association for Computational Linguistics.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition*, pages 127–149.

Darren Gergle, Robert E. Kraut, and Susan R. Fussell. 2013. Using visual information for grounding and awareness in collaborative tasks. *Human–Computer Interaction*, 28(1):1–39.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The photobook dataset: Building common ground through visually-grounded dialogue. *arXiv preprint arXiv:1906.01530*.

Robert D. Hawkins, Michael Franke, Michael C. Frank, Adele E. Goldberg, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. 2021. From partners to populations: A hierarchical bayesian account of coordination and convention.

E. Holliman, J. Godfrey, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520, Los Alamitos, CA, USA. IEEE Computer Society.

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.

Mandar Joshi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. 2019. BERT for coreference resolution: Baselines and analysis. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Charles Kemp, Yang Xu, and Terry Regier. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1):109–128.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40.

Oliver Roesler and Ann Nowé. 2019. Action learning and grounding in simulated human–robot interactions. *The Knowledge Engineering Review*, 34.

Andy Shih, Arjun Sawhney, Jovana Kondic, Stefano Ermon, and Dorsa Sadigh. 2021. On the critical role of conventions in adaptive human-ai collaboration. *arXiv preprint arXiv:2104.02871*.

Mohit Shridhar and David Hsu. 2018. Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831*.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533. Association for Computational Linguistics.

7 Appendix

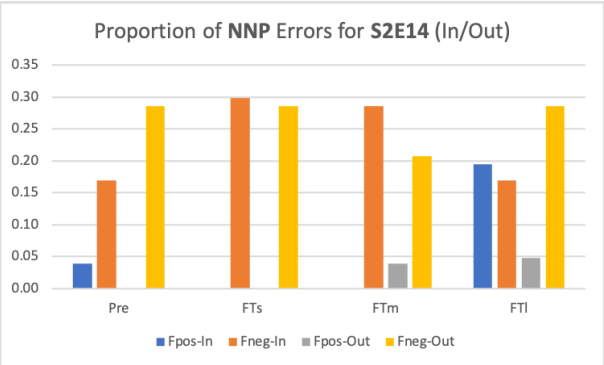

Figure 1

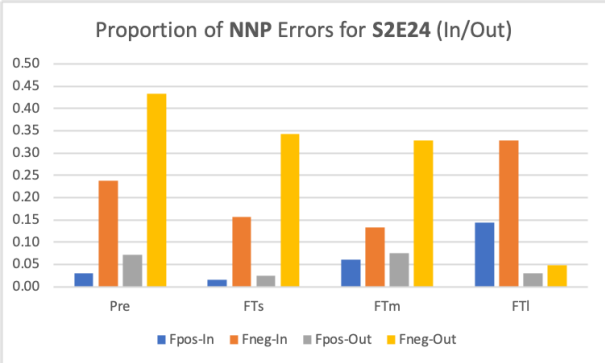

Figure 4

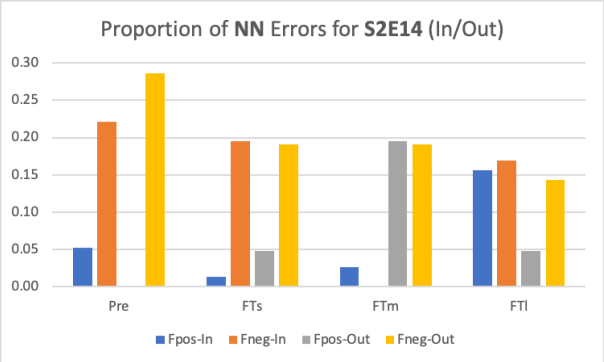

Figure 2

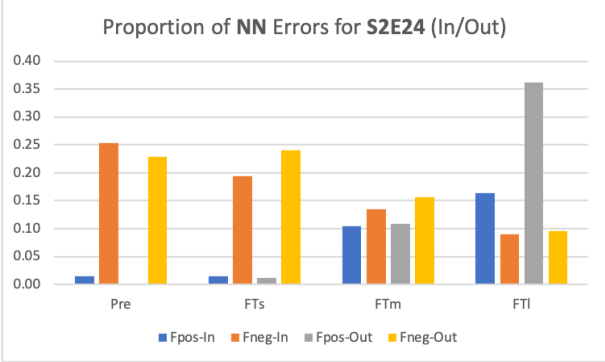

Figure 5

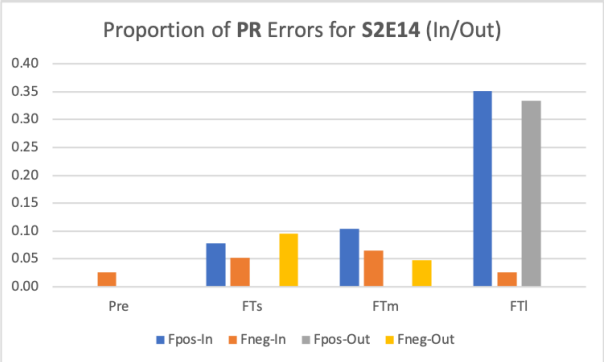

Figure 3

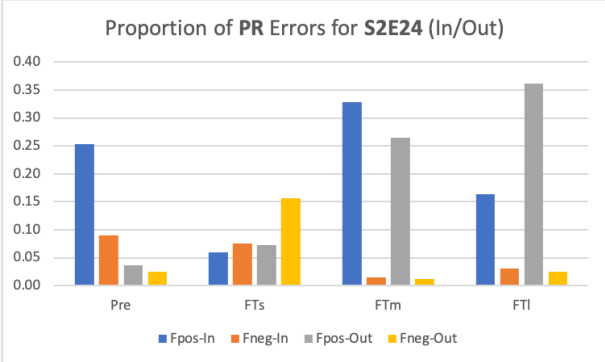

Figure 6

S2E14	Entity	#M	#V	#P	Variants
I	59	20	15	8	'Chandler': 5, 'man': 2, 'and': 1, 'an': 1, 'now': 1, 'even': 1, 'Well': 1, 'extra': 1, 'there': 1, "d'ya": 1, 'bud': 1, 'manager': 1, 'Bing': 1, 'Man': 1, 'Dude': 1
I	306	13	5	3	'Rachel': 5, 'she': 4, 'her': 2, 'waitress': 1, 'Rach': 1
I	335	13	9	3	'Ross': 5, 'dad': 1, 'man': 1, 'Geller': 1, 'yours': 1, 'date': 1, 'guy': 1, 'darling': 1, 'Jack': 1
I	183	12	7	6	'he': 3, 'man': 3, 'bud': 2, "n't": 1, 'it': 1, 'is': 1, ',': 1
I	248	10	7	4	'Monica': 3, 'her': 2, 'person': 1, 'darling': 1, 'sweetie': 1, 'she': 1, 'She': 1
O	55	6	3	3	'Casey': 3, 'he': 2, 'guy': 1
O	227	6	5	2	'he': 2, 'guy': 1, 'buddy': 1, 'him': 1, 'man': 1
O	78	3	3	2	'Dave': 1, 'Thomas': 1, 'founder': 1
I	271	3	2	2	'Judy': 2, 'mom': 1
I	30	2	2	1	'him': 1, 'he': 1
O	231	2	2	1	'Marcel': 1, 'Marceau': 1
O	352	2	2	1	'Steffi': 1, 'Graf': 1
I	51	1	1	1	'Carol': 1
O	137	1	1	1	'Gail': 1
I	145	1	1	1	'Gunther': 1
I	292	1	1	1	'she': 1
I	358	1	1	1	'Susan': 1
O	397	1	1	1	'woman': 1

Table 7: Statistics on the mentions, variants and their part-of-speech for the test case S2E14 with a 4/1 ratio for inner and outer entities. The first column differentiates inner circle (I) and outer circle (O) entities

S2E24	Entity	#M	#V	#P	Variants
O	60	33	8	3	'she': 10, 'her': 10, 'She': 4, 'girl': 3, 'guy': 3, 'person': 1, 'girlfriend': 1, 'woman': 1
I	306	22	7	4	'Rach': 6, 'Rachel': 6, 'she': 5, 'her': 2, 'honey': 1, 'Sweetie': 1, 'bride': 1
I	183	10	6	4	'Joey': 4, 'actor': 2, 'guy': 1, 'professional': 1, 'him': 1, 'Tribiani': 1
O	392	10	4	3	'Beatty': 4, 'guy': 3, 'Warren': 2, 'he': 1
O	29	9	4	4	'Barry': 5, 'him': 2, 'his': 1, 'Barr': 1
I	317	9	5	3	'Richard': 3, 'him': 3, 'sweetie': 1, 'He': 1, 'man': 1
O	215	7	6	4	'She': 2, 'Her': 1, 'Lola': 1, 'her': 1, 'she': 1, 'star': 1
O	242	7	6	2	'Min': 2, 'Mindy': 1, 'Mrs.': 1, 'Hunter': 1, 'Farber': 1, 'honey': 1
I	59	6	2	2	'Chandler': 5, 'guy': 1
I	30	5	3	3	'Benny': 2, 'he': 2, 'baby': 1
O	61	5	4	2	'husband': 2, 'his': 1, 'person': 1, 'guy': 1
I	168	5	2	1	'she': 4, 'her': 1'
I	335	5	4	4	'Ross': 2, 'his': 1, 'boyfriend': 1, 'She': 1
I	248	3	2	1	'Monica': 2, 'Honey': 1
O	252	3	3	2	'Mother': 1, 'Theresa': 1, 'mother': 1
O	228	2	2	2	'guy': 1, 'him': 1
I	292	2	2	2	'friend': 1, 'Phoebe': 1
O	17	1	1	1	'Angela': 1
O	32	1	1	1	'Man': 1
O	62	1	1	1	'secretary': 1
O	266	1	1	1	'Wineburg': 1
O	277	1	1	1	'Wineburg': 1
O	298	1	1	1	'friend': 1
O	372	1	1	1	'Tony': 1

Table 8: Statistics on the mentions, variants and their part-of-speech for the test case S2E24 with a 1/1 ratio for inner and outer entities. The first column differentiates inner circle (I) and outer circle (O) entities