Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 104–111
Marseille, 11–16 May 2020

MuDoCo: Corpus for Multidomain Coreference Resolution and Referring Expression Generation

Scott Martin, Shivani Poddar, and Kartikeya Upasani
Facebook, Inc.
1 Facebook Way, Menlo Park, CA 94025 USA
{scottfb, shivanip, kart}@fb.com

Abstract

This paper proposes a new dataset, *MuDoCo*, composed of authored dialogs between a fictional user and a system who are given tasks to perform within six task domains. These dialogs are given rich linguistic annotations by expert linguists for several types of reference mentions and named entity mentions, either of which can span multiple words, as well as for coreference links between mentions. The dialogs sometimes cross and blend domains, and the users exhibit complex task switching behavior such as re-initiating a previous task in the dialog by referencing the entities within it. The dataset contains a total of 8,429 dialogs with an average of 5.36 turns per dialog. We are releasing this dataset to encourage research in the field of coreference resolution, referring expression generation and identification within realistic, deep dialogs involving multiple domains. To demonstrate its utility, we also propose two baseline models for the downstream tasks: coreference resolution and referring expression generation.

Keywords: anaphora, coreference, dialog

1. Introduction

In this work we introduce a new dataset, *MuDoCo*, composed of a total of 8,429 dialogs with an average of 5.36 turns per dialog. The dataset contains authored dialogs between a ficional user and system, where the user asks the system to perform extensive tasks within multiple domains, as well as switch across a set of 6 task domains (calling, messaging, reminders, weather, news, and music). The users exhibit complex task switching behavior such as

- Re-initiating a previous task in the dialog by referencing the entities within it
- Re-using entities from a previous task as data for a new task

The re-use of entities across task domains introduces the aspect of resolving entities not only within the same domain, but across multiple ones. For instance, resolving the name of a song, which could also match a street name, to provide coherent information to the user and enable task completion. This work is distinguished from the existing literature in that, to the best of our knowledge, there does not exist any other dataset that is able to provide an insight into coreference and referring expression generation (REG) across domains. We also demonstrate that our dataset can be used for these valuable downstream dialog-focused tasks, and we propose baseline models for each, both built using this new dataset.

The work is organized in the following sections. In Section 2. we examine the various related datasets available in the area, as well as the downstream tasks that can draw value from our new dataset. In Section 3., we describe the proposed dataset MuDoCo, the authoring, annotation, and review processes as well as the various labels the resulting dataset exposes. Section 4. presents the two baseline experiments we undertook to probe the value of the *MuDoCo* data for the downstream tasks of coreference resolution and REG. Finally, in Section 5. we conclude and present the future directions we plan to explore as a result of this work. We hope that this new, publicly released dataset will provide a valuable and unique resource for future research in modeling multi-domain, multi-turn dialog and associated tasks.

2. Related Work

Many datasets for modeling coreference resolution focus on journalistic or document-based coreference chains. The most popular examples include MUC-7 (Chinchor, 2001), (discussed by Hirschman and Chinchor (1997)), ACE (Doddington et al., 2004), OntoNotes (Ralph Weischedel and Houston, 2013) (discussed by Pradhan et al. (2012)), and PreCo (Chen et al., 2018). Most work on coreference since OntoNotes was released has focused on that dataset (Clark and Manning, 2016a; Clark and Manning, 2016b; Lee et al., 2017; Peters et al., 2018). Notable exceptions include the ARRAU corpus (Massimo Poesio and Hitzeman, 2013), which contains a subset with dialogs containing coreference, and CoQA (Reddy et al., 2019), which has coreference annotations but is focused on the question answering task.

In contrast to these datasets, the MuDoCo dataset we propose here focuses entirely on the task of modeling entity mentions and coreference links in the setting of multi-turn, possibly multi-domain dialogs. This dialog setting differs from document-based coreference in the following important ways:

- It is straightforward to determine the speaker of a given utterance (user or system), and so more information is available for the resolution and generation of first- and second-person pronouns
- The majority of the references in the dataset (55.27%) involve simple pronoun references rather than the multi-word references often used in journalistic writing

- Entities may be re-used across multiple domains
- A single entity mention can change its cluster depending on domain (e.g., the song title *New York* versus the place name *New York*)
- Since they contain less descriptive content than full noun phrases, pronoun mentions may be more ambiguous in the dialog setting than in journalistic text, where a premium is placed on interpretability

MuDoCo explicitly identifies coreference across domains within a deep dialog of several turns. Within multiple domains, MuDoCo provides labels for three facets: references, named entities (which include all non-reference descriptions of entities), and coreference links between mentions (these labels are described in greater detail in Section 3.). In giving access to multiple domains, this dataset also exposes models to the greater challenge of resolving similarly named entities across domains. For example, *Play Game of Thrones* could mean that the user wants the system to play the soundtrack of that title or, alternatively, to play the video, based on the context in which the entity has been mentioned.

In particular, MuDoCo also compares favorably with OntoNotes in that singleton mentions (ones that are mentioned once but never referenced again by a pronoun) are explicitly labeled via either named entity or reference labels, or both. This aspect of the MuDoCo dataset also allows mention *detection* (the task of identifying mention spans in a text) to be separated from the task of mention *clustering* (the task of determining coreference links between mentions) because it is possible for a mention to receive a named entity or reference label without appearing in any mention cluster. Like the ARRAU dataset, MuDoCo also contains rich linguistic annotations for coreference but is considerably larger, containing 8,429 dialogs compared to ARRAU's 552 (Poesio and Artstein, 2008; Uryupina et al., 2016). And in contrast to CoQA, the focus of MuDoCo is solely on modeling coreference in a multi-turn dialog setting, rather than the related domain of question answering that CoQA seeks to address (Reddy et al., 2019).

Several works (Lee et al., 2017; Lee et al., 2018; Wu et al., 2019) explore approaches to model coreference resolution. Contextual embeddings from SpanBERT (Joshi et al., 2019) modify BERT's (Devlin et al., 2019) masking strategy to respect spans of entities instead of randomly masking tokens, and are hence better suited to coreference resolution. To keep the inference time of models short for deployment to production systems, we stick to simple feed-forward networks with feature engineering in this work.

3. Data

We solicited an English dataset that is designed to represent the kinds of short dialogs that might take place between a human user and a digital assistant, where the user might switch domains in the middle of a dialog. We then had the dataset labeled for references, named entities, and for coreference links. Our intention was to use this human-labeled dataset for coreference resolution, referring expression generation (REG), and for other related dialog modeling tasks such as slot carryover, which occurs when a user underspecifies a task by eliding an argument that occurred as a mention in the immediately preceding utterance(s). No higher-level structural aspects, such as discourse relations or turn taking, were annotated. The MuDoCo dataset is available for download at `https://github.com/facebookresearch/mudoco/`.

3.1. Dialog Authoring

We contracted a group of several dozen content specialists, all native speakers of English but not necessarily expert linguists, to generate the dialogs, which were focused around the following six task domains:

calling user initiates or manipulates a voice or video call

messaging user sends or reads messages, asks for information about their message queue

reminders user sets, modifies, queries or deletes reminders for a certain date or time

weather user asks about the current or future weather conditions in various locations

news user asks for information about current events related to a variety of topics

music user searches for music by a certain artist or in a certain genre, asks the system to play songs, etc.

Both roles in the dialog (user and system) were authored by a single content specialist, that is, each dialog was written in its entirety by a single individual. The content specialists were provided with guidelines that instructed them to create natural-seeming dialogs that made a point to both use references where appropriate and to switch domains when possible. An example domain switch would be a dialog like the one in Figure 1, where the dialog begins in the calling domain but then switches to the messaging domain.

> U: *Call Roberta*
> S: *I'm sorry, there was no answer*
> U: *Ok, cancel the call and send her a message that I'll be late*

Figure 1: Example of a domain switching dialog from the MuDoCo dataset.

The content specialists were asked to author dialogs between one and ten turns in length, but given the differing natures of the various domains, some domains have a greater average turn length than others. An example one-turn dialog is in Figure 2, in which the user responds to an incoming call. Importantly, this single-turn example in Figure 2

> U: *Can you accept the call from my boss and let him know I will be with him in a few minutes?*

Figure 2: Example of a single-turn dialog in the MuDoCo calling domain.

demonstrates that the corpus contains instances of intrasentential coreference.

The breakdown of the dataset by domain is given in Figure 3. For multi-domain dialogs, we chose the domain by the majority domain for the utterances in the dialog, with the first-used domain used in cases of a tie. We did some

domain	dialogs	turns	avg. turns
calling	4,714	24,245	5.14
messaging	1,110	6,185	5.57
reminders	704	3,799	5.40
weather	127	643	5.06
news	382	2,233	5.85
music	1,392	8,092	5.81
all	8,429	45,197	5.36

Figure 3: Number of dialogs and utterances in the dataset, by domain.

sanitization of the data after the authoring phase, namely tokenization and normalization. For example, punctuation is separated from neighboring words by a single space, and the possessive marker *'s* is also separated by a single space. The dataset is also split across domains between partitions for training (80%), evaluation (10%), and testing (10%), with sampling performed in a random way.

3.2. Annotation and Review

We then had all of the dialogs annotated with three different types of labels:

1. Reference mentions such as pronouns, ordinals, and definites
2. Named entity mentions of people, places, songs, movies, etc. (essentially, all non-reference mentions)
3. Coreference links between mentions

Reference mentions and named entity mentions are both represented as continguous spans that may be comprised of multiple words, and coreference links are modeled as simply a pair mentions. For each label type, we asked two trained linguists to separately annotate each dialog for reference and named entity mentions as well as for coreference links. We then had any disagreements between the linguist annotators manually resolved by a third annotator who was also an expert linguist.

Both reference mentions and named entity mentions are labeled as continguous character spans, identified by start and end character positions in the utterance where they occur. For a given dialog, each mention is uniquely identified by the triple $\langle t, s, e \rangle$, where t is the (1-based) turn number, and s and e the (0-based) character indices, respectively. References and named entities are labels for these span identifiers, whereas coreference links are represented as pairs of span identifiers. Importantly, reference and named entity annotations can overlap, so that the same text span can be annotated both with a reference type and with a named entity type.

3.2.1. Reference labeling

The annotators were asked to label references in the dialog utterances according to six different types:

personal pronoun *I, you, your, he, she, her, him*, etc.

location pronoun *here, there*

demonstrative pronoun *that, this*

definite determiner *the*

content reference e.g., *birthday* in *her birthday*

ordinal reference e.g., *third* in *the third one*

Since the annotation is at the level of mention spans, there are multi-word references that can be reconstructed from a sequence of reference labels, as in *her third message*, which would receive the label sequence **personal pronoun, ordinal reference, content reference**. An example reference labeling is given in Figure 4.

U: *Can* ***you*** *play Red by Taylor Swift ?*
S: ***This song*** *can be played from* ***your library*** *or Spotify .*
U: *Play from* ***my library*** *.*
S: *Okay , playing now .*

Figure 4: Example reference labeling for a short dialog in the music domain. Personal pronouns are labeled **orange**, demonstratives **olive**, and content references **magenta**.

3.2.2. Named entity labeling

We also asked a separate group of annotators to label mentions for named entity type, with the following four types:

entity a general entity with a non-specific type, such as a song or movie

location the name of a place or a location pronoun

person the name of a person or a personal pronoun

time an identifier of a date or time, or a datetime pronoun

The mention spans labeled for named entity type can and often do coincide with mention spans that have also been separately labeled for reference type, and so the notion of named entity labeling in MuDoCo goes beyond just proper nouns and place names. An example is a pronoun like *her*, which is labeled as both a pronoun and as a person. The named entity labeling for the dialog that was labeled for references in Figure 4 is given in Figure 5. Note that, in

U: *Can* ***you*** *play* ***Red*** *by* ***Taylor Swift*** *?*
S: *This* ***song*** *can be played from* ***your library*** *or* ***Spotify*** *.*
U: *Play from* ***my library*** *.*
S: *Okay , playing* ***now*** *.*

Figure 5: Example named entity labeling for the dialog in Figure 4. Person entities are labeled **teal**, general entities **violet**, and time entities **green**.

conjunction with the annotation in Figure 4, the mention span *your library* from the third turn of the dialog in Figure 5 is labeled both as a reference and as a named entity.

3.2.3. Coreference link labeling

Finally, we asked a third, separate group of expert linguist annotators to label any coreference links between mentions that are present in each dialog. This is accomplished by pairing any coreferent mention spans that occur either as a reference label or as a named entity label. For this reason, the coreference link labeling pass was undertaken once the other two labeling passes (for references and named entities) had been fully reviewed. The dataset represents coreferent links via a set of mention clusters, so that all of the mentions that are mutually coreferent in a given dialog are collected together into a single mention cluster.

The coreference link labeling for our running example from Figures 4 and 5 is given in Figure 6. Note that, as a simplification to avoid links spanning multiple mentions, coreference links are established between the head mention of a mention sequence. For example, in Figure 6 the content reference *song* from the second turn is linked to the song title *Red* from the first, but not the entire mention *This song*. If the reference were simply the demonstrative *This* without any content reference, then the demonstrative would have been chosen as the mention paired with *song*.

U: *Can you play **Red** by Taylor Swift ?*
S: *This **song** can be played from your library or Spotify .*
U: *Play from my library .*
S: *Okay , playing now .*

Figure 6: Example coreference link labeling for the dialog in Figures 4 and 5. Links are signaled by a shared coloring of both components of the mention pair.

4. Experiments

To demonstrate the utility of MuDoCo for modeling phenomena relevant to dialog, we experimented with models for coreference resolution and REG trained on its labeled dialogs. This section details those two experiments in turn.

4.1. Coreference resolution

To check how well the dataset models coreference resolution, we implemented a simple coreference model learned from the training set. Since the dataset represents coreference links with the head mention, we count all non-head mentions in a chain containing a head mention as though they pointed to the head mention.

4.1.1. Model

The model uses a feed-forward network with a mix of embedding-based features and features that are straightforward to compute from the mentions and labels for the context dialog. We start with a single concatenated vector representing a candidate mention pair with features representing both mentions and features relating the two mentions, similar to the mention-pair encoder used by Clark and Manning (2016b). The input feature vector contains

1. Embeddings and features related to the first mention,
2. Embeddings and features related to the second mention, and
3. Features about the suitability of the mention pair for coreference.

The feed-forward network uses two hidden layers, each with an associated ReLU activation (Nair and Hinton, 2010), and a final sigmoid prediction layer that outputs the probability that the two mentions are coreferent. Figure 7 shows a diagram of the network architecture.

Letting m and n represent two mentions from a given dialog, we construct the input vector as

$$x = [e_m, f_m, e_n, f_n, s_{m,n}],$$

where e_a represents the embeddings generated for mention a, f_a represents the mention-based features for a, and $s_{a,b}$ the features pertaining to the suitability of a and b as a coreferent pair. (These vector components are discussed in more detail in Section 4.1.2., below.) The model architecture then learns the weights associated with each layer and the implied final prediction as the probability $p(m, n)$ that m and n are coreferent.

$$\begin{aligned} h^{(1)} &= \mathbf{relu}(W^{(1)}x + b^{(1)}) && \text{(hidden 1)} \\ h^{(2)} &= \mathbf{relu}(W^{(2)}h^{(1)} + b^{(2)}) && \text{(hidden 2)} \\ p(m,n) &= \mathbf{sigmoid}(w^\top h^{(2)}) && \text{(prediction)} \end{aligned}$$

Where $W^{(i)}$ and $b^{(i)}$ respectively represent the weight matrix and biases for layer i, and w the weights for the final prediction layer. The model is trained by finding the parameters $\theta = \langle W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}, w \rangle$ that minimize a cross-entropy loss function over the labels in the training set.

4.1.2. Features

The embedding-based mention features come from 300-dimensional word embeddings, which were trained on multi-user threads from a commercial chat application and are not updated during the training of our model. We derive the following features from these embeddings:

mention start the embedding vector for the first word in the mention

mention end the embedding vector for the last word in the mention

window average an average over the embeddings for the mention's five preceding and five following words

The window size is adjusted as needed depending on the available words in the utterance, for example in *show me the first one now*, the preceding window for *the first one* is *show me* while the following window is *now*.

We also use four mention features that are not based on embeddings.

animacy whether the mention span has a named entity label **person**

plurality whether the mention is a plural pronoun (*we, our, ours, they, them, their, theirs, these, those*)

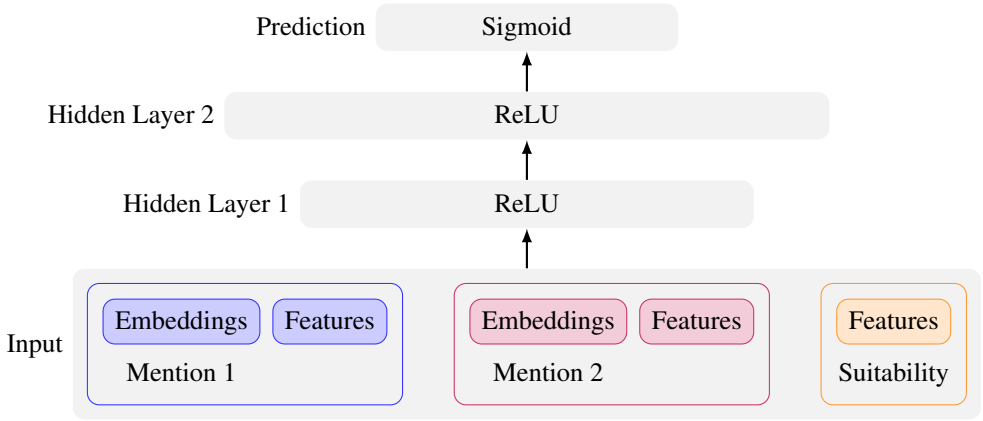

Figure 7: Diagram of the feed-forward network architecture. Dropout layers between the two hidden layers are suppressed.

pronominality whether the mention is a (personal, location, or demonstrative) pronoun

person the mention's grammatical person, determined as

- **first** if the mention is a first-person pronoun (*I, me, my, mine, we, our, ours*)
- **second** if the mention is a second-person pronoun (*you, your, yours*)
- **third** otherwise

For training, these features are represented numerically, with 0 or 1 for the boolean features animacy, plurality, and pronominality and 1, 2, or 3 for the grammatical person value.

Lastly, we derive several features having to do with how well suited the two mentions are for coreference.

type compatibility whether the named entity labels for both mentions match

string matching two features:

- **exact** both mentions are string-identical
- **partial** one mention is a substring of the other

certainly (not) coreferent features are based on the utterance source, user or system (determined by turn number, since the user turns are always odd)

- whether both mentions are first-person pronouns
- whether one mention is a first-person pronoun but the other is not

turn difference the number of turns separating the two mentions, binned (0, 1, 2–5, >5)

intervening mentions the number of mentions between the two mentions, binned (0, 1, 2–5, 5–10, >10)

The complete input vector for training the model contains the features for both mentions, the features relating the mentions, and lastly a boolean label representing whether the mentions occur together in a mention cluster.

4.1.3. Hyperparameters and training

Since our model makes decisions based on mention pairs, there is a strong class imbalance toward the negative (not coreferent) case in the training data. We correct this by randomly subsampling the negative instances in the training set until the negative class is roughly the same size of the positive (coreferent) class.

Through experimentation, we determined that the best sizes for the hidden layers of the network are 695 nodes and 997 nodes, respectively. We also add dropout layers to aid in training, and determined via experiment that the optimal rate for dropout is 0.82. The training proceeds over 25 epochs with a batch size of 390 samples, and we used ADAM (Kingma and Ba, 2014) for optimization with a learning rate of 0.001, also determined by experimentation.

4.1.4. Results on various subsets of the data

To see how well this model performs at the coreference linking task, we divided the data into subsets based on a **core** and an **extended** set for domains, reference types, and named entity types, as described in Figure 8. These subsets of the data are devised to gauge how well the coreference model does on the entire, more complex set of labels versus its performance on a simpler label set containing only pronouns and the less complex named entity labels **person** and **location**, which are often shorter than general entity descriptions like *the third song on my mom's playlist* or time descriptions such as *two Fridays from tomorrow*. We also subset the domains to exclude the more complex domains related to music and news from the core domains, since music and news often involve entity mentions that are not always a simple proper name, place name, or pronoun.

We then trained a model with the data partitioned on each of the eight core/extended subsets of the data. The results for each subset are given in Figure 9. These results show the pairwise precision, recall and F1 score for each of the eight ways of subsetting the dataset based on core and extended domains, reference labels, and named entity labels for every mention pair in an entire dialog. As expected, the model performs best on the core subsets of the domains and labels, with both the best precision and F1 obtained on the set containing the core domains, reference labels, and named entity labels. The subset with the core domains and named entity labels but the extended reference labels

	core	extended
domain	calling, messaging, reminders, weather	core + news, music
reference	personal pronouns, location pronouns, demonstratives	core + content references, definite determiners, ordinal references
named entity	person, location	core + entity, time

Figure 8: Data subset criteria.

domain	ref.	NE	P	R	F1
core	core	core	**78.70**	89.64	**83.81**
	ext	core	63.59	**90.88**	74.82
	core	ext	66.20	87.71	75.45
	ext	ext	62.22	89.62	73.45
ext	core	core	76.94	89.06	82.56
	ext	core	59.88	89.88	71.87
	core	ext	51.92	90.49	65.98
	ext	ext	59.15	85.04	69.77

Figure 9: Results for coreference models trained on each of the data subsets.

achieved the best recall, but recall was also high throughout (between 85.04 and 90.88). The performance of the core labelsets on the extended domains is also quite good, at 82.56 F1. In all cases the F1 score degrades quite noticeably when the extended label sets are added, owing to decreased precision, with about a 9 point drop for the core domains and around a 13 point drop for the extended domains. Interestingly, the full dataset (i.e., the one containing both the extended domains and extended labels) outperforms the subset with extended domains, core reference labels, and extended named entity labels on both precision (59.15 vs. 51.92) and F1 (69.77 vs. 65.98). Room for improvement by other approaches remains for both the core and extended subsets, however, since the model described here does no better than 83.81 F1 score for either.

4.2. Referring Expression Generation

Referring expression generation (REG) is a sub-task of natural language generation (NLG) which involves selecting the most natural form to refer to a noun based on context. If an entity is of high salience because of contextual factors, then using a pronoun rather than a longer-form referring expression may seem more human-like. Factors contributing to salience may include an entity being mentioned frequently or recently, or being the only entity in the thread. If a pronoun is used when an entity is not salient enough, the reference may be ambiguous, making it impossible to understand which entity the pronoun refers to.

We show that the dataset can also be used to learn a classifier for choosing natural referring expressions for person entities. Specifically, given some multi-turn dialog context and an entity, the task is to choose between the noun or pronoun form for expressing the entity. The formulation of problem is similar to that done previously in literature (McCoy and Strube, 1999; Henschel et al., 2000). Downstream components like traditional NLG systems or neural models can consume this information for choosing the right surface form of an entity.

The following properties make the dataset interesting for this task:

- Turns with pronouns: 24,980 / 45,197
- Threads with pronouns: 4,677 / 8,429
- Threads with >1 persons: 7,933 / 8,429

For a given entity, we encode information about its mentions in previous turns as features described in Table 1. Both user and assistant turns are used for this purpose. Only

feature	type
is last mentioned entity	binary
turns since mention	integer
num. entities in thread	integer
beginning of thread	integer
cluster size	integer
pronoun used before	binary

Table 1: Input features for the REG model in order of importance.

third person pronouns are considered (instances with like *me* for *you* are ignored for purposes for REG).

A gradient-boosted decision tree (Friedman, 2001) is fed the features and predicts whether pronoun use is appropriate (1) or not (0) for the next mention of the entity. The classifier achieves a precision of 87.9 at 67.8 recall (76.55 F1). We bias it towards precision since, in the setting of needing to generate a system utterance, it is better to err on the side of not using a pronoun than using one when any ambiguity may confuse the user.

Figure 10 shows an example with multiple entities where the model is correctly able to choose pronoun forms such that they are natural and not ambiguous. In Figure 11 we

U: *Did **Jerry** call ?*
S: ***Jerry Gergich** ?*
U: *No I meant **Larry** , **Larry Gingrich** .*
S: *Yes , **he** called at lunch .*
U: *Please tell **him** not to call me .*
S: *Message sent .*

Figure 10: Example for REG when multiple entities are involved in dialog. The model predicts the correct form for each reference in this case.

give an example where the model predicts the use of the pronoun *her* in the second turn whereas the ground truth contains the name *Kim*. In this instance, the pronoun use is

U: *Is that **Kim** calling ?*
S: *No , it is not **her** .*

Figure 11: Example for cases where pronoun is predicted correctly but is still penalized since ground truth contains name instead.

unambiguous since there is only one entity in the conversation. However, since the ground truth contains otherwise, we end up penalizing the model in calculating accuracy.

5. Conclusion and Future Work

We presented a new dataset for coreference resolution and generation that focuses on modeling reference in the dialog setting, in contrast to most reference datasets that focus on journalistic style text and clustering mentions across an entire document. This dataset includes dialogs from a variety of domains and also includes many instances of domain switching in the middle of a dialog. We demonstrated that a feed-forward neural coreference model based on word embeddings and a small, simple feature set performs fairly well on the mention clustering task, especially when the more complex domains and label types are excluded. In future work, we will experiment with more complex models of coreference such as the ones investigated by Clark and Manning (2016a; Clark and Manning (2016b) and Lee et al. (2017).

We also show that the dataset can be used to learn usage of the right referring expression for person entities in multi-turn dialog. Scope for future work includes learning referring expressions for other entities such as locations, song names, and artist names. This work also assumes that the right form of pronoun (*he* vs *him*) is known beforehand, leaving the task of learning to generate the specific form for future work.

6. Acknowledgements

We give thanks to Long Ma and Rebecca Silvert for their help with organizing and labeling the dataset, and to Ashish Baghudana for developing an initial prototype of the experimental coreference model we employed here. We also thank the anonymous reviewers for several helpful suggestions for related work.

7. Bibliographical References

Chen, H., Fan, Z., Lu, H., Yuille, A. L., and Rong, S. (2018). PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Clark, K. and Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Clark, K. and Manning, C. D. (2016b). Improving coreference resolution by learning entity-level distributed representations. In *Association for Computational Linguistics*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proceedings of the Language Resources and Evaluation Conference*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232.

Henschel, R., Cheng, H., and Poesio, M. (2000). Pronominalization revisited. In *Proceedings of the 18th Conference on Computational Linguistics, Volume 1*, pages 306–312. Association for Computational Linguistics.

Hirschman, L. and Chinchor, N. (1997). MUC-7 coreference task definition. In *Proceedings of MUC7*.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2019). SpanBERT: Improving pre-training by representing and predicting spans. *CoRR*, abs/1907.10529.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana, June. Association for Computational Linguistics.

McCoy, K. and Strube, M. (1999). Generating anaphoric expressions: Pronoun or definite description? In *ACL Workshop on Discourse and Reference Structure*, pages 63–71.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning (ICML)*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*.

Poesio, M. and Artstein, R. (2008). Anaphoric Annotation in the ARRAU Corpus. In *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL2012 shared task: Modeling

multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL)*.

Reddy, S., Chen, D., and Manning, C. (2019). CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7(0).

Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Rodriguez, K., and Poesio, M. (2016). ARRAU: Linguistically-Motivated Annotation of Anaphoric Descriptions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2058–2062, Portorož, Slovenia. European Language Resources Association (ELRA).

Wu, W., Wang, F., Yuan, A., Wu, F., and Li, J. (2019). Coreference resolution as query-based span prediction.

8. Language Resource References

Nancy Chinchor. (2001). *Message Understanding Conference (MUC) 7*. Linguistic Data Consortium, ISLRN 783-262-033-141-8.

Massimo Poesio, Ron Artstein, Olga Uryupina, Kepa Rodriguez, Francesca Delogu, Antonella Bristot and Janet Hitzeman. (2013). *The ARRAU Corpus of Anaphoric Information*. Linguistic Data Consortium, ISLRN 462-157-606-044-8.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchin, Mohammed El-Bachouti, Robert Belvin and Ann Houston. (2013). *OntoNotes Release 5.0*. Linguistic Data Consortium, 5.0, ISLRN 151-738-649-048-2.