SOCCER: An Information-Sparse Discourse State Tracking Collection in the Sports Commentary Domain

Ruochen Zhang and **Carsten Eickhoff**
Brown University
{ruochen_zhang, carsten}@brown.edu

Abstract

In the pursuit of natural language understanding, there has been a long standing interest in tracking state changes throughout narratives. Impressive progress has been made in modeling the state of transaction-centric dialogues and procedural texts. However, this problem has been less intensively studied in the realm of general discourse where ground truth descriptions of states may be loosely defined and state changes are less densely distributed over utterances. This paper proposes to turn to simplified, fully observable systems that show some of these properties: Sports events. We curated 2,263 soccer matches including time-stamped natural language commentary accompanied by discrete events such as a team scoring goals, switching players or being penalized with cards. We propose a new task formulation where, given paragraphs of commentary of a game at different timestamps, the system is asked to recognize the occurrence of in-game events. This domain allows for rich descriptions of state while avoiding the complexities of many other real-world settings. As an initial point of performance measurement, we include two baseline methods from the perspectives of sentence classification with temporal dependence and current state-of-the-art generative model, respectively, and demonstrate that even sophisticated existing methods struggle on the state tracking task when the definition of state broadens or non-event chatter becomes prevalent.

1 Introduction

State tracking, the task of maintaining explicit representations of user requests and agent responses, has long been a key component of dialogue systems (Williams et al., 2013; Henderson et al., 2014a,b; Kim et al., 2016). The same challenge arises during reading comprehension of procedural texts (recipes, how-to guides, etc.) where systems focus on predicting changes of object attributes at the entity-level (a car window may transition from foggy to clear) (Dalvi et al., 2018; Tandon et al., 2020). However, both of these state tracking variants rely on transaction-based or turn-based data such as transactional dialogues or procedure descriptions that are information-dense. Few works have studied state tracking tasks where state changes occur infrequently while a large proportion of messages are "chatter".

As an alternative to altogether unrestricted state tracking—a task that is daunting due to the complexity of even describing ground-truth states in a discrete manner—we resort to a simpler and more self-contained setting: sports competitions. Given the stream of natural language utterances with which a commentator describes the events in a real-world setting (here a sports competition), an ideal natural language understanding system would maintain and reason over a coherent and accurate representation of the match based on how the commentator described it. This representation can, in turn, be used for downstream tasks such as inference or language generation. Sports matches provide an ideal test bed for state tracking due to their self-contained, fully observable nature and their inherent interpretability in the form of the temporal evolution of scores. However, existing sports-related commentary collections such as described by Aull and Brown (2013) and Merullo et al. (2019) do not provide such within-match temporal information.

To this end, we collect temporally-aligned commentaries and live scores of soccer matches along with other meta information from the website goal.com and compile the dataset SOCCER. To the best of our knowledge, SOCCER is the first temporally-aligned collection of sports match commentary and state. It contains over 2,200 matches from tournaments such as the UEFA Champions League or the UK Premier League between 2016 and 2020. Across these matches, there are over
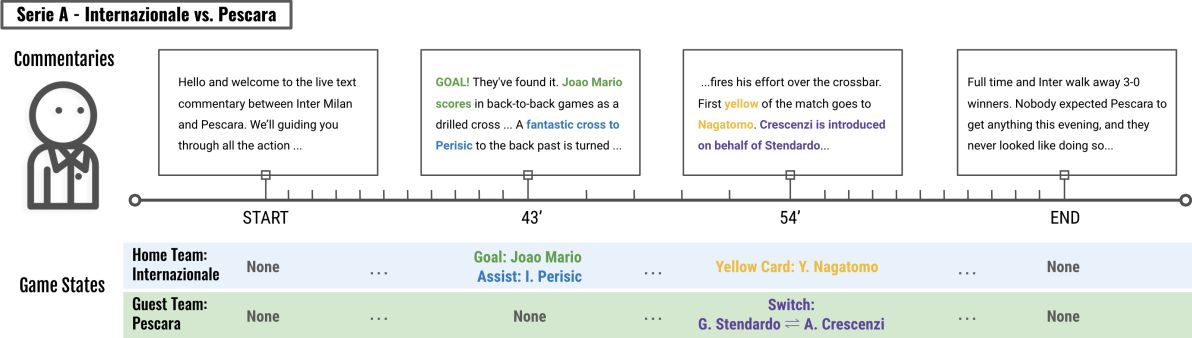
Figure 1: An overview of the state tracking task in sports commentary.

135,000 individual comments and approximately 31,000 events. A simplified example is shown in Figure 1.

To demonstrate the potential of state tracking for open-domain discourse, we use the proposed dataset to investigate to what degree state-of-the-art systems are able to track the progression of events described in the commentary. This overview includes two model classes: classification models that treat match events as different class labels, and generative language models such as GPT-2 (Radford et al., 2019) that model context and events in a causal manner. Our experiments show that both methods do not perform well on SOCCER and only slightly outperform distributional heuristics, leaving considerable room for improvement.

The novel contributions of this paper are three-fold: (1) we propose a new task of tracking event occurrences via state changes, (2) we create SOCCER, a general discourse state tracking dataset that contains temporally-aligned human-composed commentary and in-game events, serving as the training and evaluation dataset for this task, and (3) we provide two intuitive baselines demonstrating the difficulty of this task and presenting exciting opportunities for future research.

2 Related Work

Dialogue State Tracking (DST). Current DST collections and benchmarks tend to rely on transaction-centric dialogues with predefined domain-specific ontologies and slot-value pairs. Prominent examples include the DSTC2 (Henderson et al., 2014a) and MultiWOZ datasets (Budzianowski et al., 2018). Consequently, previous work focuses on picklist-based approaches (Mrkšić et al., 2017; Perez and Liu, 2017; Zhong et al., 2018; Ramadan et al., 2018; Gao et al., 2019)

to formulate state tracking as a series of classification tasks over candidate-value lists. A major difference between SOCCER and other DST datasets lies in its information density. As dialogues in DST are usually short conversations with direct transactional objectives such as booking hotels or reserving restaurant tables, frequent state changes are required to be captured within limited turns of the conversation. In sports commentary, on the contrary, in-game events occur at a comparatively low frequency and a considerable proportion of commentator utterances may not be related to any changes in the game state.

State Tracking in Procedural Text. State tracking in procedural text understanding focuses on the task of tracking changes in entity attributes (Tandon et al., 2020). A variety of procedural progresses have been proposed such as tracking entity presence and location in scientific processes (Dalvi et al., 2018), ingredients in cooking recipes (Bosselut et al., 2017), and character motivation and emotional reaction in simple stories (Rashkin et al., 2018). Yet, similar to DST settings, these highly specific tasks depend on small fixed ontologies covering limited ranges of entities and states. Another more recent dataset (Tandon et al., 2020) turns to an open-vocabulary setting when defining entity attributes. But since the dataset is comprised of how-to guides from WikiHow.com, the task still sees a high density of state changes per natural language instruction.

Information Density The concept of Information Density has been mainly used in the Uniform Information Density (UID) theory (Jaeger, 2010) to measure the amount of information per unit comprising an utterance. Levy and Jaeger (2007) demonstrated that speakers tend to maximize the uniformity of information via syntactic reduction.
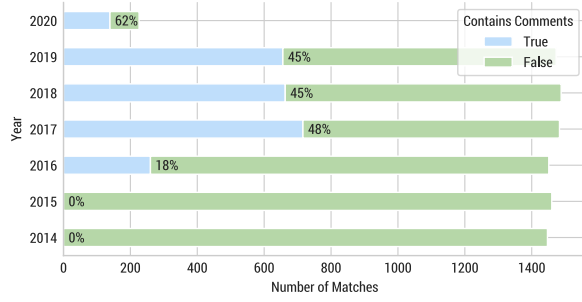
Figure 2: Frequency distribution of matches with and without commentary across available data years.

The notion of information density in our paper, however, focuses on quantifying the frequency of event occurrences on the corpus level instead of understanding syntactic choices on the utterance level.

Sports Event Datasets and Tasks. Commentary in the sports domain has been collected to study a variety of problems such as racial bias in football game reporting (Merullo et al., 2019) and gender construction in NBA/WNBA coverage (Aull and Brown, 2013). However, these datasets do not provide any information on the temporal alignment between commentary and events. Another similar dataset, BALLGAME (Keshet et al., 2011) is comprised of baseball commentary with annotated events and timestamps, but it contains less than 20 games and the annotation is unavailable online. Some work focuses on sports-related inference of player performance metrics (Oved et al., 2019) or game outcomes (Velichkov et al., 2019) that predict full-time results based on signals from pre-game player interviews. However, no in-game sequential contexts are provided in these datasets. Most similar to our work, Bhagat (2018) collected in-game commentaries for soccer player analytics, but their approach is restricted by classical machine learning methods and ignores the effect of information sparsity within the dataset.

3 Dataset Construction

We collect time-stamped commentary with key events of 2,263 soccer matches in total. The matches stem from four major soccer tournaments including the UEFA Champions League, UEFA Europa League, Premier League and Series A between 2016 and 2020. SOCCER consists of over 135,000 time-stamped pieces of commentary and 31,000 within-match events. This section describes our data collection and preparation process in detail.

Figure 3: A short snippet of a match in the dataset.

3.1 Data Processing

Commentaries, events, team lineups, match dates and other meta-information are gathered from match-specific pages. Out of a total of 9,028 matches covered on goal.com between 2014 and 2020, we retain only those 2,434 matches that list detailed event records and commentary. Any matches missing either of the two information streams are discarded. The retained matches belong to the four major tournaments mentioned above and all occurred starting 2016. Figure 2 shows the frequency distribution of included and overall matches across the years in which they took place. All commentaries are in English and available in text form, thus requiring no transcription. Pieces of commentary come pre-segmented and aligned to match-internal timestamps so that in-game events and commentary with the same timestamps can be linked. Comments whose temporal information is unavailable usually belong to the pre-game, intermission and post-game periods and are labeled as START, BREAK, END accordingly. The total number of commentary paragraphs within a game is the same as the number of timestamps. This number varies between matches as timestamps during which the commentator did not provide commen-

Event Name	Goal		Assist		Yellow Card		Red Card		Switch	
Team	*Home*	*Guest*	*Home*	*Guest*	*Home*	*Guest*	*Home*	*Guest*	*Home*	*Guest*
Event # per team	3582	2799	2434	1871	3948	4320	163	197	6111	6117
Event total #	6381		4305		8268		360		12228	
Player # per team	1001	924	882	774	1548	1613	145	183	2546	2575
Player total #	2915		1656		3161		328		5121	

Table 1: Event type and player name distribution.

tary are omitted. Finally, any templated sentences following the format "*team 1 score - score team 2*" are removed to avoid trivial leakage of the match state. All annotation and filtering processes are done programmatically and no manual efforts are involved.

Events are classified into five types: *goal, assist, yellow card, red card* and *switch*. We consider events as keys and the event-related players as the corresponding values. For example, if player B from the home team assists in scoring a goal, player B will be the value of the event *assist* for the home team. Hence, at each timestamp t, there are ten event-player pairs (five event types tracked for two teams). From this representation, we construct a comprehensive game state incorporating all the event-player pairs for each team as well as a cumulative score at each timestamp (See Figure 3). Special events such as penalty goals or own goals are not explicitly labeled, but can be derived from the evolution in cumulative score between neighboring timestamps. After processing, 171 games were found to have ill-formed commentary or misaligned end-game match scores compared to the goal records in the key events. These matches were eliminated from the original 2,434 games crawled with commentary, giving us a total of 2,263 games. Finally, the collected data is partitioned into distinct training (70%), validation (15%) and test (15%) sets.

4 State Definition and Task Proposal

For each match m in the dataset M, there is a set of timestamps $T_m = \{t\}$ accurate to a minute. As input, we are given a stream of commentaries $C_m = \{c_t\}_{t=1}^{T_m}$ and c_t represents the paragraph of commentary at time t. The output will be a set of general match states $S_m = \{s_t\}_{t=1}^{T_m}$ such that each s_t reflects the state change in the comment c_t at the same timestamp. s_t contains a set of events $e_{i,j}^{(t)}$, where i represents the event types ($i \in \{goal, assist, yellow\ card, red\ card, switch\}$) and j denotes the event actor ($j \in \{home, guest\}$).

Given the sparse distribution of s_t, we propose two alternative variants of the variable to assess the difficulty of state tracking at different granularity levels of state resolution.

Team Level. In this simplest notion of state, events are tracked at the team level. In other words, $e_{i,j}^{(t)} = \{yes, no\}$. Consider the event of the home team scoring a goal $e_{goal,\ home}^{(t)}$ at time t as an example: given the commentary c_t and other related meta-information, a model is tasked with determining the value of $e_{goal,\ home}^{(t)}$ to be *yes* if the home team indeed scored a goal in a given minute, or *no* otherwise.

Player Level. At this significantly increased level of resolution, all events are additionally associated with their player agents $p \in P$, where P denotes the collection of players. Concretely, the variable $e_{i,j}^{(t)}$ is mapped to either the related players' names p or a *none* answer to each event at time t. To facilitate this form of state, match meta-information includes lineups that associate present players with teams.

5 Analysis and Baseline Experiments

In the following, we provide descriptive statistics of the SOCCER dataset and include two model baselines for recognizing match events resulting in changes of states.

5.1 Dataset Statistics and Comparison

The SOCCER dataset covers 2,263 matches with 135,805 pieces of commentary and 31,542 in-game event records. In all event records, each event type of each team appears approximately 3,154 times on average. There are a total of 3,507 unique player names across all event types and an average 1,219 unique player names per event type per team. A more detailed overview of the distribution of event types and player names can be seen in Table 1.

Common state tracking datasets either in dialogue systems or procedural texts are designed to capture frequent state changes in the text. In
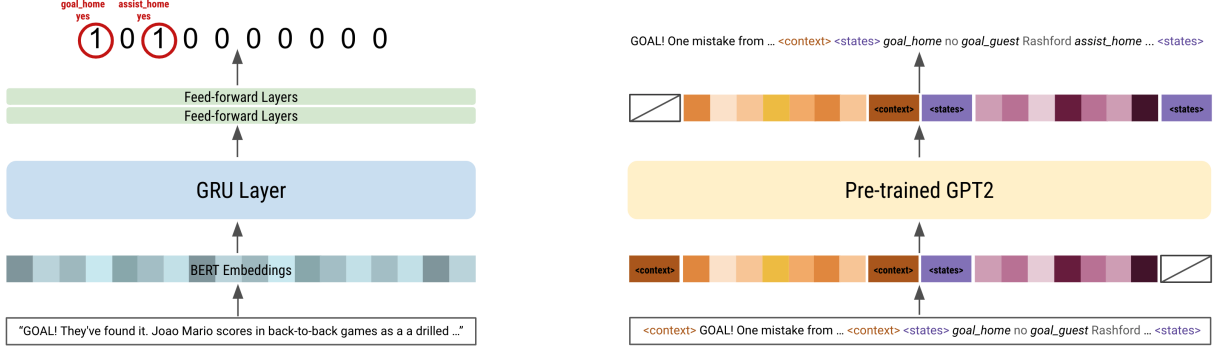
Figure 4: Model architecture of the GRU classifier and GPT-2 based variant.

	MultiWOZ2.1	OpenPI	SOCCER
Avg. turn length	11.75 - 15.12	13.32	50.28
Avg. turn #	13.46	5.00	60.01
ID	1.05	3.8 - 4.3	0.19

Table 2: Information density of SOCCER vs. other state tracking datasets.

SOCCER, we study a more general setting where the corpus is much less information dense due to an abundance of non-event related chatter. To quantify this difference, we define information density (*ID*) as:

$$ID = \frac{Total\ \#\ of\ state\ changes}{Total\ \#\ of\ turns/steps/timestamps}$$

As shown in Table 2, our dataset has a considerably lower information density with more turns of information. In SOCCER, the match state only gets updated every 5 timestamps, while in datasets such as MultiWOZ2.1 (Eric et al., 2019) and OpenPI (Tandon et al., 2020), there are between 1 and 4 state changes per turn or step on average.

5.2 Baseline Setup

SOCCER presents a new challenge to the state tracking community by introducing a more general corpus with an all-new state definition and a sparse information distribution. These properties render it difficult to directly apply some existing models such as TRADE used in DST tasks and ProLocal (Dalvi et al., 2018) proposed for procedural texts. Motivated by previous work on state tracking and based on the characteristics of the task, we use two baseline training and inference schemes: 1) a GRU (Cho et al., 2014) classifier with pre-trained BERT (Devlin et al., 2019) embeddings, and 2) a generative pre-trained GPT2 (Radford et al., 2019) variant.

GRU Classifier with BERT Embeddings. The GRU model is used as a preliminary baseline to

assess the difficulty level of the SOCCER dataset. Embeddings of the timestamped commentary c_t are obtained from the pretrained weights of BERT (Devlin et al., 2019), that then get fed into a 1-layer GRU (Cho et al., 2014) network followed by two feed-forward layers. We only tasked this model with team-level state tracking as the classification will be extremely difficult if each player name is treated as a distinct class. We map the 10 event variables $e_{i,j}^{(t)}$ as binary flags to a 10-bit scalar value in which each digit denotes the predicted value of a variable. For example, if the 0th position corresponds to the variable $e_{goal,\ home}^{(t)}$, then the predicted value at that position denotes whether the home team scores a goal (See Figure 4). Compared to converting the problem into ten binary classifications, this allows us to directly model the joint occurrence of events.

GPT-2 Based Variant. Recent approaches to state tracking (Kim et al., 2019; Hosseini-Asl et al., 2020; Tandon et al., 2020) have shown that generative models are competitive especially in open-vocabulary settings. Inspired by simpleTOD (Hosseini-Asl et al., 2020) and the OpenPI baseline (Tandon et al., 2020), we cast the player-level state tracking task as a sequence generation problem, allowing us to leverage the capabilities of causal language models such as GPT-2 (Radford et al., 2019). The training sequence consists of a concatenation of the commentary, event types and player names, allowing us to model the joint probability of the whole sequence. Event names are preprocessed as tokens like *goal_home* to avoid being tokenized into sub-word units. Commentary and event-player pairs are encapsulated in special tokens to help the model distinguish context from labels. See Figure 4 for a schematic overview of the model training input. In training, the model takes the concatenated

	Team Level		Player Level	
Metrics	*Acc.*	*Pos. Recall*	*Acc.*	*Pos. Recall*
GRU classifier	0.9775	0.3990	-	-
GPT-2 variant	0.9759	0.4855	0.9670	0.0775

Table 3: Team and player level test set performance.

sequence as input to perform next token prediction task. At inference time, greedy decoding is used to generate state predictions due to its superior performance compared to beam search and top-k sampling (Hosseini-Asl et al., 2020).

5.3 Implementation Details

During preprocessing, we find that 98.1% of comments in the collection are shorter than 200 words, therefore any outliers with a length of more than 200 words are truncated at that point. Then, the input text sequences are tokenized using byte-pair encoding (Sennrich et al., 2016) to avoid out-of-vocabulary words.

The sentence embeddings processed by the GRU classifier stem from the pretrained weights of HuggingFace's BERT model (Wolf et al., 2019). The GPT-2 model (Radford et al., 2019) is also obtained from HuggingFace with pretrained weights, which are then fine-tuned on SOCCER[1].

5.4 Evaluation

Accuracy, and recall for occurrences of all event-types are used to assess the performance of both models. Due to the sparsity of event occurrences, recall is crucial to track the models' ability to extract events given the full set of types. For convenience, we refer to event types with ground truth *none* answers as negative cases and positive cases otherwise. Therefore, recall among event occurrences is referred to as positive recall in the tables. More specifically, in Tables 3 and 5, accuracy and positive recall are measured on all labels (positive and negative combined). In Table 4, the performance is reported on positive labels only, and detailed metrics including precision, recall and F1 scores are provided.

6 Results

This section reports the results on the test set of SOCCER. As a naïve distributional baseline, we compute the ratio of negative cases in the test set to be 0.9766.

In Table 3, both models achieve an accuracy that is approximately equal to this majority class baseline due to the heavily imbalanced distribution of event positives and negatives. While accuracy scores are very high, positive recall is much lower, indicating that many event occurrences are missed by the models. When comparing the GPT-2 model's performance on both team level and player level event recognition[2], we notice that player level recall is substantially worse than that on team-level. This result suggests that complex state tracking involving broad ranges of possible slot values is a comparatively harder task that may require more sophisticated approaches.

6.1 Results Per Event Type

In addition to these general results, we break down model performance of positive cases by event-type and provide additional metrics including precision, recall and F_1 scores (see Table 4). When associating the scores with the event type distribution (see Table 1), we can observe that, generally, greater numbers of available data points result in better performance. Take the event type *goal* as an example. According to Table 1 there are about 800 more positive cases of the event $e_{goal,\ home}^{(t)}$ than $e_{goal,\ guest}^{(t)}$. A difference that is reflected in all the metrics in Table 4 for both models. Another interesting point to note is the performance gap between the GRU classifier and GPT-2 model on the event type *red card*. The *red card* event is extremely rare in SOCCER as illustrated in Table 1. Though we observe the performance of both models on *red card* events to be comparably lower than those of the other events, the GRU classifier is able to capture more positive cases while no occurrences are detected by GPT-2.

6.2 Results on Varying Information Densities

In Section 5.1, we have shown that a key difference between SOCCER and other state tracking datasets lies in its low information density (See Table 2 for a detailed comparison). It is conceivable that such differences in information density affect state tracking performance. To eliminate confounding effects introduced via direct comparison to other datasets, this section explores the connection between event density across pieces

---

[1]The SOCCER dataset as well as the code base used to collect it and run the experiments presented in the remainder of this paper are available here.

[2]The GRU classifier is only used in team-level tasks since treating each player in the ontology as a distinct class to classify is very difficult.

	Event Type	Goal		Assist		Yellow Card		Red Card		Switch	
Model	**Team**	*Home*	*Guest*	*Home*	*Guest*	*Home*	*Guest*	*Home*	*Guest*	*Home*	*Guest*
GRU Classifier	*precision*	0.54	0.50	0.51	0.33	0.46	0.54	0.13	0.19	0.55	0.53
	recall	0.83	0.01	0.84	0.00	0.53	0.28	0.04	0.33	0.50	0.62
	F_1	0.66	0.02	0.64	0.01	0.49	0.37	0.06	0.24	0.52	0.57
GPT-2 Variant	*precision*	0.56	0.32	0.42	0.27	0.48	0.39	0.00	0.00	0.53	0.46
	recall	0.89	0.08	0.92	0.06	0.95	0.06	0.00	0.00	0.93	0.01
	F_1	0.69	0.13	0.58	0.09	0.64	0.11	0.00	0.00	0.67	0.01

Table 4: Model performance of positive classes on team-level task per event type.

Comment Sparsity	0%		20%		40%		60%		80%	
Metrics	*Acc.*	*Pos. Recall*	*Acc.*	*Pos. Recall*	*Acc.*	*Pos. Recall*	*Acc.*	*Pos. Recall*	*Acc.*	*Pos. Recall*
Task Level					**Team Level**					
GRU Classifier	0.89	0.44	0.90	0.41	0.92	0.35	0.94	0.30	0.97	0.31
GPT-2 Variant	0.88	0.49	0.90	0.49	0.93	0.47	0.95	0.41	0.98	0.44
Task Level					**Player Level**					
GPT-2 Variant	0.83	0.06	0.87	0.06	0.90	0.04	0.94	0.04	0.98	0.02

Table 5: Model performance on team-level and player-level tasks with data of different information density.

of commentary and model performance. We begin by discarding all but the truly event related comments in each match to obtain a subset containing 0% negative cases. This subset contains 25,934 event related comments across all matches. Then, by randomly replacing positive comments [3] with negative ones from the same match at a sparsity ratio $r \in \{20\%, 40\%, 60\%, 80\%\}$, we keep the total number of comments at the same constant count of 25,934 and keep the temporal ordering of comments intact, while effectively reducing the level of information density. Table 5 reports accuracy and positive recall for both methods and task levels when training and evaluating on non-overlapping splits of the newly constructed subsets. Note that, despite our earlier discussion of information density, Table 5 reports a converse notion, sparsity. In this setting, 0% corresponds to the highest and 80% the lowest information density.

Comparing accuracy at different event sparsity levels, we notice that scores increase as events become more sparsely distributed. This effect stems from the fact that, when we are replacing event related comments with non-event chatter, chance agreement improves as the number of true negatives increases. Positive recall of event occurrences, however, demonstrates an opposing trend, suggesting that the task of recognizing true state updates becomes more challenging the sparser the discourse domain is. This assumption is further supported by the different degree of performance observed on SOCCER vs. existing collections such

as MultiWOZ2.1 (Eric et al., 2019), where recall scores of many models range in the mid-fifty percent range.

7 Conclusion

In this paper, we introduce SOCCER, the first discourse state tracking collection in the sports commentary domain. We propose two different levels of state granularity and provide two performance benchmarks for models ranging from GRU (Cho et al., 2014) for embedding temporal dependency to GPT-2 (Radford et al., 2019) for causal language modeling. The dataset shows a much lower information density than many existing resources on state tracking, making it considerably more challenging. We believe that, in conjunction with the wide vocabulary of player-level notions of state, this property makes SOCCER an exciting resource on which our community can advance discourse state tracking to a broader range of settings than have been studied previously.

Acknowledgement

---

[3] Positive comments here refer to comments with event occurrences.

References

Laura L Aull and David West Brown. 2013. Fighting words: a corpus analysis of gender representations in sports reportage. *Corpora*, 8(1):27–52.

Rahul Ashok Bhagat. 2018. *Towards Commentary-Driven Soccer Player Analytics*. Ph.D. thesis.

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2017. Simulating action dynamics with neural process networks. *arXiv preprint arXiv:1711.05313*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1595–1604, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–273, Stockholm, Sweden. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014a. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014b. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329. IEEE.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

T Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Ezra Keshet, Terrence Szymanski, and Stephen Tyndall. 2011. BALLGAME: A corpus for computational semantics. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, Matthew Henderson, and Koichiro Yoshino. 2016. The fifth dialog state tracking challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 511–517. IEEE.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2019. Efficient dialogue state tracking by selectively overwriting memory. *arXiv preprint arXiv:1911.03906*.

Roger Levy and T Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19:849.

Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating sports commentator bias within a large corpus of American football broadcasts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6355–6361, Hong Kong, China. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Nadav Oved, Amir Feder, and Roi Reichart. 2019. Predicting in-game actions from the language of nba players. *Computing Research Repository*, arXiv:1910.11292.

Julien Perez and Fei Liu. 2017. Dialog state tracking, a machine reading approach using memory network. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 305–314, Valencia, Spain. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Osman Ramadan, Paweł Budzianowski, and Milica Gašić. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 432–437, Melbourne, Australia. Association for Computational Linguistics.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi Mishra, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. *arXiv preprint arXiv:2011.08092*.

Boris Velichkov, Ivan Koychev, and Svetla Boytcheva. 2019. Deep learning contextual models for prediction of sport event outcome from sportsman's interviews. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1240–1246, Varna, Bulgaria. INCOMA Ltd.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1458–1467, Melbourne, Australia. Association for Computational Linguistics.