Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4302–4308
Marseille, 11–16 May 2020

Annotating Event Appearance for Japanese Chess Commentary Corpus

Hirotaka Kameko, Shinsuke Mori
Academic Center for Computing and Media Studies, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, Japan
{kameko, forest}@i.kyoto-u.ac.jp

Abstract

In recent years, there has been a surge of interest in natural language processing related to the real world, such as symbol grounding, language generation, and non-linguistic data search by natural language queries. Researchers usually collect pairs of text and non-text data for research. However, the text and non-text data are not always a "true" pair. We focused on the *shogi* (Japanese chess) commentaries, which are accompanied by game states as a well-defined "real world". For analyzing and processing texts accurately, considering only the given states is insufficient, and we must consider the relationship between texts and the real world. In this paper, we propose "Event Appearance" labels that show the relationship between events mentioned in texts and those happening in the real world. Our event appearance label set consists of temporal relation, appearance probability, and evidence of the event. Statistics of the annotated corpus and the experimental result show that there exists temporal relation which skillful annotators realize in common. However, it is hard to predict the relationship only by considering the given states.
Keywords: game commentary, modality, symbol grounding

1. Introduction

These days, the interest in the *symbol grounding* problems becomes larger and larger. A large number of datasets are now available in such as image and text (Young et al., 2014; Chen et al., 2015). These corpora are widely used for description generation (Vinyals et al., 2015; Xu et al., 2015), non-text information retrieval (Ushiku et al., 2017), and so on.

Some of these corpora consist of pairs of non-text data and human-writing texts. We usually treat the pairs that the text and non-text data are compatible. However, the assumption is not always true if the corpora are automatically collected. For example, non-text data sometimes reminds humans of some associated things, and they mention to them instead of the exactly given data. To reproduce intelligent systems which work like a human, we must analyze the gap between the text and non-text data.

We focused on commentary for extensive-form games like chess. The rules of these games are well-defined, and we can treat them with computer programs easier than the real world. With this well-defined world, computers, can access to the previous states and search the future states for predicting easily. Human players also consider the previous states and future predictions, and they often refer to these states when they comment on the game. For accurate analyzing and processing, we must analyze the relationship between texts and the real world for the first step.

In this paper, we propose "Event Appearance" label. Event appearance shows the relationship between events that are mentioned in texts and events that happen in the real world. Event appearance consists of temporal relation, appearance probability, and evidence of event. For annotation, we called annotators with high skill of *shogi*. In addition, we evaluated the annotated corpus with statistics and experiments.

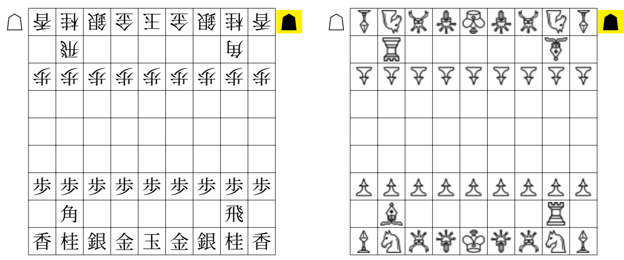


Figure 1: Initial state of *shogi*. Left: normal (Japanese Kanji style) depiction. Right: chess-like depiction.

2. Game and Commentary

2.1. *Shogi*: Japanese Chess

Shogi, which is known as Japanese chess, is a two-player board game similar to chess. As shown in Figure 1, each piece is presented by Japanese kanji characters. Same as chess, the goal of the *shogi* is to capture the opponent's king. Each player chooses one action, called move, from legal actions alternately. The most significant difference is that the players can use the captured pieces by dropping them to the vacant cells. For more detail, please refer to (Leggett, 2009).

The rule of *shogi* is well-defined and more comfortable to tackle for computers than real-world tasks. Constructing strong computer players is one of the milestones of artificial intelligence, and there exist some computer AI players that defeat top human players (Silver et al., 2016). Some of them are available as applications and users can access them easily. They use these applications not only as an opponent player but also as an evaluation tool.

2.2. *Shogi* and Commentary

Some of the professional *shogi* games are broadcast with human expert commentaries. For example, in Meijin-sen and Jun'i-sen, one of the largest professional *shogi* tournament and title match, all games are broadcast via the website for a fee[1]. These commentaries mainly explain the current game by reasoning the actions, evaluating the states, and predicting the next actions. These commentaries help spectators understand the games. On the other hand, commentators sometimes mention other games such as the players' previous games. These commentaries are not related to the current games.

2.3. *Shogi* Game Corpus

We constructed Shogi Game Corpus (SGC) (Mori et al., 2016) by collecting the commented *shogi* records of Meijin-sen and Jun'i-sen tournament. The game record is a sequence of game states from the initial state to the end of the game. Each comment is mapped to one of the states in the sequence. We also defined *shogi* named entities (s-NEs) and annotated commentaries. We built an automatic named entity recognition (NER) tool of s-NEs by using the annotated corpus, and the experimental result showed that we could obtain high-quality NER by using our corpus. We also augmented the SGC with modality expressions and event factuality annotation (Matsuyoshi et al., 2018).

These labels are annotated by using natural language information. The commentaries are tied up to each state, and we usually treat the pairs of comments and states as if they are correctly mapped. However, the commentaries are sometimes about the past or future states of the game, or outside of the current game. For an accurate evaluation of symbol grounding, we should construct a corpus which shows the relationship between natural language texts and non-text world states.

Table 1 shows the list of s-NEs that we defined on (Mori et al., 2016). In this work, we selected Strategy (St), Castle (Ca), and Move Name (Mn) as target s-NEs. Castle is a defensive piece formation to protect a king. These three NEs are strongly related to the states.

Tag	Meaning
Tu	Turn
Po	Position
Pi	Piece
Ps	Piece specifier
Mc	Move compliment
Mn*	Move name
Me	Move evaluation
St*	Strategy
Ca*	Castle
Ev	Evaluation: entire
Ee	Evaluation: part
Re	Region
Ph	Phase
Pa	Piece attribute
Pq	Piece quantity
Hu	Human
Ti	Time
Ac	Player action
Ap	Piece action
Ao	Other action
Ot	Other notion

Table 1: *Shogi*-specific named entity tags. The marked "*" indicates that the tag is selected for event appearance annotation.

3. Event Appearance

We define "*Event Appearance*" (EA) and annotate SGC with it. EA consists of temporal relation, appearance probability, and evidence of event. Each s-NE has one EA label.

3.1. Temporal Relation

Temporal relation has four types of labels. The descriptions and examples of each label are below (underlined phrases $_{\mathsf{Tag}}$ shows the target s-NEs.):

Present The event which the s-NE indicates appears in the current state.

ex.) 対する先手は美濃囲い$_{\mathsf{Ca}}$ に組んだ。

(Black player constructed Mino castle $_{\mathsf{Ca}}$.)

Note that some events appear in both past and current states. In this case, we annotate the events as Present.

Past The event which the s-NE indicates appeared in the past states.

ex.) 85 手目の突き捨て$_{\mathsf{Mn}}$ は、この変化でも生きてくる。

(Push Sacrifice $_{\mathsf{Mn}}$ at 85th move also shows an effect on this position.)

Future The event which the s-NE indicates will appear in the future states.

ex.) このあと居飛車$_{\mathsf{St}}$ に組む可能性が高そうだ。

(The probability that the player will choose Static-Rook strategy $_{\mathsf{St}}$ is high.)

Not The event which the s-NE indicates does not appear in the current and past states, and will not appear in the future states.

ex.) (Player Name) は居飛車$_{\mathsf{St}}$ 党。

((Player Name) tends to adopt Static Rook strategy $_{\mathsf{St}}$.)

Undecidable Annotators cannot decide from the four categories. This label is unexpected to be chosen.

[1]http://www.meijinsen.jp/ (in Japanese).

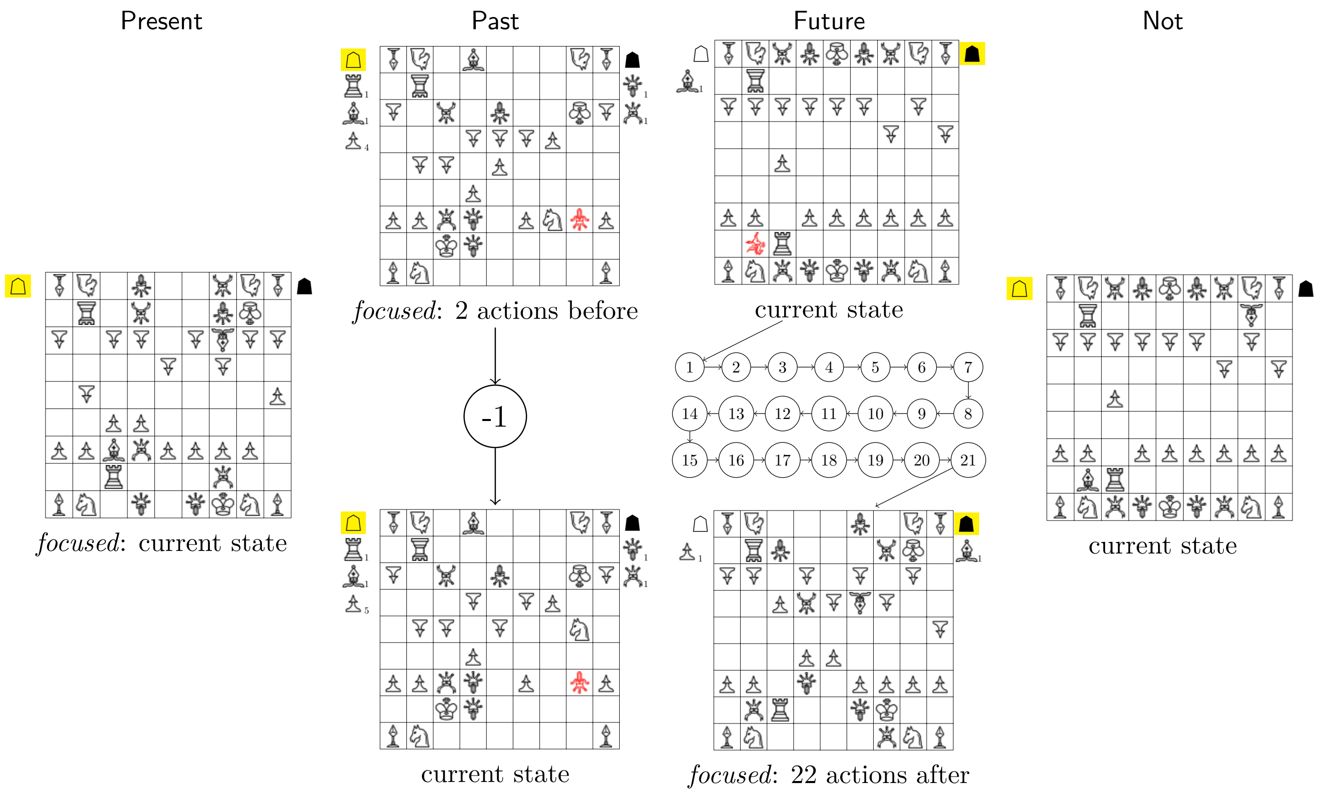


Figure 2: Example positions for each temporal relation labels. For Past and Future, we also show the *focused* positions which the events happen on.

Figure 2 shows the positions for each comments. Note that annotators decide labels by considering given states.

3.2. Appearance Probability

Future states are indefinite when the commentaries are written. We define appearance probability for s-NEs if the temporal relation is Future. Appearance probability has four types of labels:

High The event which the s-NE indicates will appear with high probability.

One-of There are more than one possible future state, and the state with the event which the s-NE indicates is one of those.

Low There are more likely future states than that with the event which the s-NE indicates, and the future state will appear with low probability.

For-Comment The event which the s-NE indicates will not appear in the proper future states because the comment is not for future prediction but for the explanation.

In *shogi* commentaries, experts sometimes explain bad actions that amateur players tend to choose. Professional players less likely to choose these actions, and the event will not appear in the proper future states.

We add one of the possible sequences of actions if the temporal relation is Future. A sequence of move expressions represents the sequence of actions. Annotators show one of the possible sequences, even if there are more than one sequence.

3.3. Evidence of Event

We define "Evidence of Event". When commentators explain the reason that they choose the s-NEs for explaining the states, they will abstract the states. For example, *4th-File-Rook strategy* (四間飛車) is defined as a strategy that the rook is swung to the 4th file from the left (Kawasaki, 2013). In this case, commentators may explain the evidence of *4th-File-Rook strategy* that the rook is swung to the 4th file from the left even if the reason is not only the position of the rook but also the whole states.

States are sometimes in the events over a long sequence of states. However, annotating all over the states may cost huge annotation efforts. Thus, in this paper, we define the evidence of event that a set of elements on exactly one state. We are mainly interested in "when the event happens" and "when the event ends". To acquire that information, we define the states of evidence for Present and Past as follows. If the temporal relation is Present, the state of evidence is the state in which the event happens and there is no state which is not in the event between the evidence state and the current state. For example, if the sequence of states is "XXOOOXOOO" where "O" and "X" are the states with and without the event, respectively, the state of evidence is the 7th one. If the temporal relation is Past, the state of evidence is the state just before the

	shogi club 24	*shogi wars*
Anttr. 1	7 *dan* (1.12%)	5 *dan* (1.1%)
Anttr. 2	6 *dan* (2.76%)	5 *dan* (1.1%)
Anttr. 3	3 *dan* (14.73%)	3 *dan* (6.9%)
Anttr. 4	3 *dan* (14.73%)	3 *dan* (6.9%)
1st author		2 *kyu* (69.9%)

Table 2: Player ranks (*dan*) of annotators (Anttr.). *Shogi club 24* and *shogi wars* are online *shogi* game servers and all players are rated by the results of games. Values in parentheses show that the rank is in the top X percentile provided by `http://shogi-naka.com/cgi-bin/24_ranking_checker.cgi` and `https://nandemoplus.com/shogiwars_occupancy/` (in Japanese). *dan* is a name of higher rank and higher number shows higher rank. *kyu* is a name of lower rank and lower number shows higher rank similar to negative numbers.

state in which the event ends. For example, if the sequence of states is "XXOOOXOOOXXX", the state of evidence is the 9th one. Annotators choose evidence from the following elements on states:

- Last action,
- Captured pieces, and
- Cells on the board.

4. Annotated Corpus

4.1. Annotation Process

First, we prepared nine games and commentaries for the games which are written in human experts. These commentaries are already annotated with s-NEs.

Four annotators annotated the same corpus. The annotators are native Japanese speakers as well as amateur *shogi* players. Table 2 shows the player ranks of annotators. The player ranks are provided by *shogi club 24*[2] and *shogi wars*[3]. The table shows that the annotators in this work have strong skill in playing *shogi*. For comparison between the higher and lower skill of *shogi*, the first author also annotated for the same data.

Figure 3 shows the web interface of annotation. The interface shows players' names of the game, a sequence of states from the initial state to the current state as a game record, and a comment with a target s-NE. Each state consists of the number of actions, the last action, the position of pieces, and the captured pieces. The annotators can choose the last action, each cell of the position, and each captured piece as the evidence of the labels. The annotators can look at only the current state and the previous states, but cannot look at the future states. The target s-NE is provided with the comment, highlighted in red.

St	Ca	Mn	Sum
56	39	124	219

Table 3: Statistics of the corpus: the number of *shogi* NEs of each type.

	Present	Past	Future	Not	Undec.
Anttr. 1	100	68	10	41	0
Anttr. 2	91	82	13	33	0
Anttr. 3	93	91	14	21	0
Anttr. 4	100	78	15	26	0
1st author	83	63	16	56	1

Table 4: Statistics of the corpus: the number of labels of each annotator.

4.2. Statistics

Tables 3 and 4 show the statistics of the corpus. No annotator chose Undecidable as temporal relation.

There are few events with Future label. It shows that commentators rarely mention to St, Ca, or Mn on future states.

Table 5 shows Cohen's kappa score (Cohen, 1960) of temporal relation labels between each pair of two annotators. It shows that the agreement between annotators is high. It suggests that annotation by annotators with high domain skills is important to analyze the texts. Fleiss' kappa score (Fleiss, 1971) of four annotators (except for the first author) is 0.74. These results show that the agreement of temporal relation is substantial.

We calculated the F-score of evidence of event by the following step:

- If the annotators chose different states, F-score = 0.
- If both annotators chose the same states but 0 elements, F-score = 1.
- Otherwise, calculate F-score by using the sets of elements of evidence.

Table 6 shows the F-score of evidence of event between each pair of two annotators. The F-scores seem a little bit low, considering the high agreement of temporal relation. There are two possible reasons. One is that the annotators have their decision process, and the evidence of the decisions is different. The other is that making the evidence of the decisions is hard, and there exists some error. We are confident that collecting the evidence of decisions is important for analyzing the humans' decision-making process.

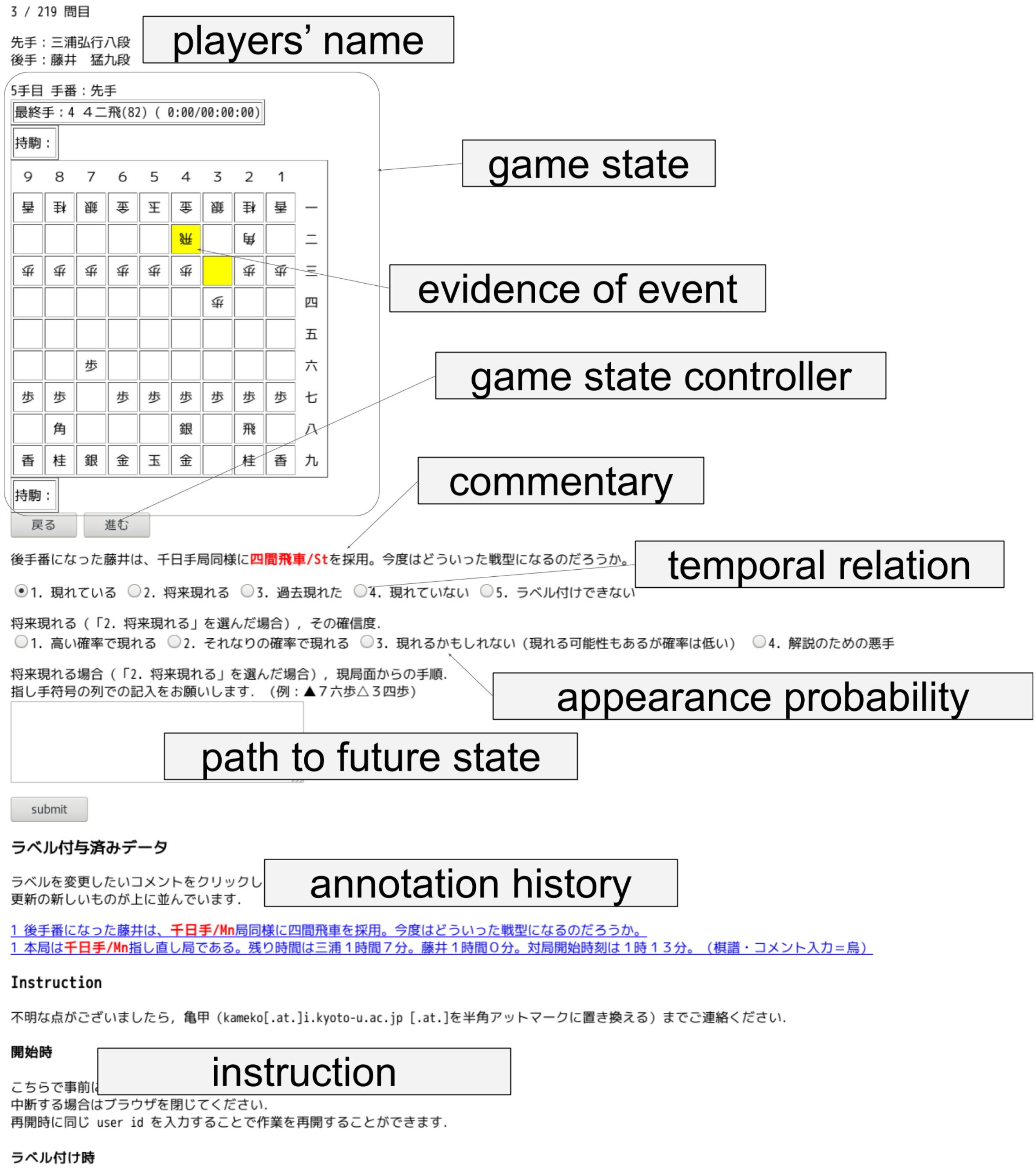
4.3. Experiments

We evaluated the *Temporal Relation* labels by comparing them and predicted s-NEs from states.

First, we trained an s-NE prediction model that predicts s-NEs that will appear in the comments for given

[2] `https://www.shogidojo.net/`

[3] `https://shogiwars.heroz.jp/`



players' name	players' name of the game
game state	focused game state (last action, positions of pieces, captured pieces)
game state controller	button to select previous/next state
commentary	target commentary (red phrase shows the focused s-NE)
temporal relation	radio button for temporal relation label
appearance probability	radio button for appearance probability
path to future state	text box for the sequence of actions to the future state
evidence of event	clickable cells for evidence of event label (yellow cells are selected)
annotation history	annotated comments (click them to revise)
instruction	annotation instruction

Figure 3: Web Interface of Annotation. The web page is written in Japanese.

states. Note that we used unlabeled comments as a training corpus, so we treated comments as positive examples if the comments include the phrases of s-NEs. We used the positions of all the pieces on the board and the captured pieces as the features of a state. We trained binary classifiers for each s-NE by using linear support vector machines, which are implemented by (Fan et al., 2008).

	1	2	3	4	A
Anttr. 1		0.70	0.71	0.68	0.70
Anttr. 2			0.78	0.78	0.72
Anttr. 3				0.79	0.66
Anttr. 4					0.68
1st author					

Table 5: Cohen Kappa score of each pair of annotators. "A" means first author.

	1	2	3	4	A
Anttr. 1		0.48	0.50	0.39	0.38
Anttr. 2			0.50	0.44	0.44
Anttr. 3				0.53	0.37
Anttr. 4					0.66
1st author					

Table 6: Pairwise F-score of evidence of event between each annotator. "A" means the first author.

We evaluated the relationship between temporal relation labels and the prediction of the trained model in two ways. One is to compare the mean of the probabilities that are calculated by the model for each label. The other is the area under the curve (AUC). We calculated AUC for each label by the following steps:

- Convert the labels into binary labels (1 if the label is focused else 0).
- Calculate AUC score by the converted labels and the probabilities that are calculated by the model.
- If AUC score < 0.5, output $1.0 -$ AUC score.

Table 7 shows the result of the experiment. The mean value of the probabilities for Present is higher than the other labels. It suggests that there is a stronger relationship between the current states and the event with Present labels than those with other labels. Present label means that the events are in the current states, so the result is expected. However, the AUC scores are about 0.5, and it shows that it is hard to classify the temporal relation by the probabilities. It suggests that considering only the current states is insufficient for analyzing the commentary.

4.4. Availability

We plan to distribute our corpus except for text data on our website [4] `http://www.ar.media.kyoto-u.ac.jp/data/game/home-e.html`. For detailed explanations, readers may visit it.

[4]The game records and the commentary sentences are distributed on the website: `http://www.meijinsen.jp` (in Japanese) for a fee. We provide a helper script to download the records and the text at `https://github.com/hkmk/shogi-comment-tools`.

		mean	AUC
Anttr. 1	Present	0.055	0.61
	Past	0.031	0.53
	Future	0.020	0.57
	Not	0.026	0.61
Anttr. 2	Present	0.045	0.50
	Past	0.041	0.54
	Future	0.029	0.55
	Not	0.029	0.55
Anttr. 3	Present	0.048	0.53
	Past	0.034	0.53
	Future	0.029	0.54
	Not	0.042	0.52
Anttr. 4	Present	0.051	0.56
	Past	0.028	0.57
	Future	0.033	0.55
	Not	0.037	0.52

Table 7: Relationship between annotators' labels and probabilities by trained *shogi* NE predictor. mean: Mean value of probabilities which the trained predictor outputs for each class. AUC: Area under the curve of one-versus-the-rest classification.

5. Applications

5.1. Module Evaluation of Symbol Grounding

Some multimodal models are trained by end-to-end architecture to reproduce the outputs by humans. A large amount of data improves these systems with the advance of technologies such as deep neural networks. However, when the performance of trained models is low, specifying the problems is sometimes hard. Our high-quality corpus, which is annotated by annotators with high skill of target domain, may help to evaluate the module evaluation of symbol grounding.

5.2. Symbol Grounding to Search Tree

For describing the machine thought, symbol grounding to the searching tree is an important factor. Searching trees is one of the results of intelligent machine thought, but it is too hard for humans to understand the huge searching tree of computer game players. Hence there is a demand for the representing system of searching trees in human-readable media such as natural language texts. The temporal relation and appearance probability, including the path to the future states, are the positive examples for symbol grounding.

5.3. Evaluation of *Explainable AI*

These days, *Explainable AI* is one of the hot topics (Costabello et al., 2019). One goal is to explain the reason for the decisions of AI by natural language since natural language is one of the easiest protocols for humans to understand.

We built an automatic game commentary system (Kameko et al., 2015). This system outputs comments for given states. For more detailed explanation, to show the evidence of the decision is a good way.

We consider that we can use our proposed corpus to evaluate such explainable abilities of the system.

6. Conclusion

In this paper, we proposed event appearance labels and augmented the SGC corpus by annotating the labels. We called annotators with high skill of the target domain *shogi*. In addition, we evaluated the hardness of predicting the temporal relation. The experimental result suggests that considering only the current states is insufficient for analyzing.

Expanding the target for event appearance labels is one of the future work. In this paper, we focused on strategy, castle, and move name because we expected that they are strongly related to states and easy for annotating and evaluating. The agreement of the labels shows that we can annotate meaningful event appearance labels.

7. Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 18K11427, 19K20341. We are also grateful to the annotators for their contribution.

8. Bibliographical References

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Costabello, L., Giannotti, F., Guidotti, R., Hitzler, P., Lecue, F., Minervini, P., and Sarker, M. K. (2019). AAAI 2019 tutorial on explainable AI: From theory to motivation, applications and limitations.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kameko, H., Mori, S., and Tsuruoka, Y. (2015). Learning a game commentary generator with grounded move expressions. In *2015 IEEE Conference on Computational Intelligence and Games*, pages 177–184.

Kawasaki, T. (2013). *HIDETCHI Japanese-English SHOGI Dictionary*. NEKOMADO, 1.1 edition.

Leggett, T. (2009). *Japanese chess : the game of shogi*. Tuttle Publishing.

Matsuyoshi, S., Kameko, H., Murawaki, Y., and Mori, S. (2018). Annotating modality expressions and event factuality for a Japanese chess commentary corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Mori, S., Richardson, J., Ushiku, A., Sasada, T., Kameko, H., and Tsuruoka, Y. (2016). A Japanese chess commentary corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Ushiku, A., Mori, S., Kameko, H., and Tsuruoka, Y. (2017). Game state retrieval with keyword queries. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 877–880, New York, NY, USA. ACM.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generation. In *Computer Vision and Pattern Recognition*, pages 3156–3164.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *arXiv*.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.