IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 May 2023, accepted 23 July 2023, date of publication 28 August 2023, date of current version 5 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3308916

RESEARCH ARTICLE

Japanese Event Factuality Analysis in the Era of BERT

HIROTAKA KAMEKO [1], YUGO MURAWAKI [2], SUGURU MATSUYOSHI [3], AND SHINSUKE MORI [1]

[1]Academic Center for Computing and Media Studies, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[2]Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[3]Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology, Hachioji, Tokyo 192-0982, Japan

Corresponding author: Hirotaka Kameko (kameko@i.kyoto-u.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant 18K11427 and Grant 19K20341.

ABSTRACT Recognizing event factuality is a crucial factor for understanding and generating texts with abundant references to possible and counterfactual events. Because event factuality is signaled by modality expressions, identifying modality expression is also an important task. The question then is how to solve these interconnected tasks. On the one hand, while neural networks facilitate multi-task learning by means of parameter sharing among related tasks, the recently introduced pre-training/fine-tuning paradigm might be powerful enough for the model to be able to learn one task without indirect signals from another. On the other hand, ever-increasing model sizes make it practically difficult to run multiple task-specific fine-tuned models at inference time so that parameter sharing can be seen as an effective way to reduce the model's size. Through experiments, we found: (1) BERT-CRF outperformed non-neural models and BiLSTM-CRF; (2) BERT-CRF did neither benefit from nor was negatively impacted by multi-task learning, indicating the practical viability of BERT-CRF combined with multi-task learning.

INDEX TERMS Event factuality, modality, sequence labeling, neural networks, multi-task learning.

I. INTRODUCTION

Identifying the factuality of an event mention is an important task in natural language processing (NLP), with a wide range of potential applications such as information extraction, recognizing textual entailment, reasoning and natural language understanding [1], [2], [3], [4], [5], [6]. Here we work on a recently published corpus on *shogi* (Japanese chess) commentaries in Japanese [7] to develop a system of event factuality analysis although the proposed method can readily be ported to other corpora following the same design principle. As an extensive-form game, *shogi* allows a computer to ground most event mentions in a game tree. Yet it is complex enough for its commentaries to exhibit a rich variety of factual statuses, for example, a possibility (Ex. (1)) and a counterfactual (Ex. (2)) (event mentions are marked with underlines):

(1) <u>居飛車</u>を採用するかもしれない
White may use <u>static rook strategy</u>

(2) <u>ゴキゲン中飛車</u>との予測は外れた
The prediction that white would use <u>cheerful central rook strategy</u> turned out to be false

Given these, we expect event factuality analysis to help automatic generation of human-like commentaries, among other applications.

The design principle this corpus adopts is to decompose event factuality analysis into a combination of several subtasks. Event mentions need to be detected to begin with. To assign factual statuses to them, we need to identify words and phrases that convey factuality information, which are a subset of *modality expressions*. Identifying grammaticalized verbs can be a useful filtering step because due to semantic bleaching, they are unsuitable for further factuality analysis. We also notice that event mentions have a substantial overlap with named entities (NEs) specially designed for the *shogi* domain [8]. The divide-and-rule strategy is useful for

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Ali.

TABLE 1. **The annotation layers of the *shogi* commentary corpus and the task definition. The glosses are not included in the corpus but added only for readers.**

	Layer																								
Input	Word	先手	は	美濃	囲い	が	崩れ	て	い	る	の	で	、	飛車	交換	は	後手	の	得	に	な	り	そう	だ	。
	gloss	black	TOP	*Mino*	castle	NOM	break	have PP			because		,	rook	change	TOP	white	's	good	to	be	be likely to			.
Output	NE	Tu-B	O	Ca-B	Ca-I	O	Ao-B	O	O	O	O	O	O	Mn-B	Mn-I	O	Tu-B	O	Ee-B	O	Ao-B	O	O	O	O
	Modality	O	O	O	O	O	MEn-B	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	MEa-B	O	O
	Event class	O	O	O	EVe	O	EVe	O	EVf	O	O	O	O	O	EVi	O	O	O	O	O	EVe	O	O	O	O
	Factuality	O	O	O	FNc	O	FPc	O	O	O	O	O	O	O	O	O	O	O	O	O	FPr	O	O	O	O

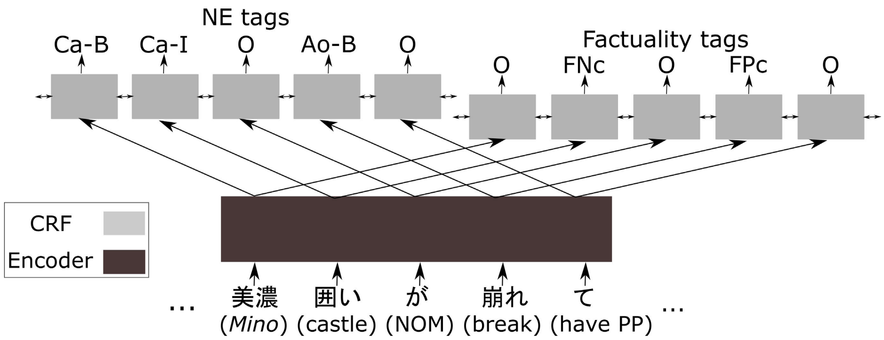


FIGURE 1. **The overall architecture of the proposed model. The networks for modality tags and event class tags are omitted to save space.**

corpus construction as well because it facilitates speedy and consistent annotation.

The question, then, is how to solve the closely related but different subtasks as a whole. Since manually writing rules to connect them [9] is daunting, it is desirable to make a computer automatically learn their relationships from data. While each subtask can be straightforwardly formalized as sequence labeling, how best to exploit dependencies among subtasks remains unknown. The creators of the annotated corpus only reported preliminary experiments where they independently tackled each subtask using a non-neural sequence labeling tool [7].

One apparently promising approach is multi-task learning. Unlike taggers supplied with hand-crafted features, neural networks have the ability of flexible knowledge sharing among related subtasks, which has proven to be effective in natural language analysis [10], [11]. For sequence labeling, knowledge sharing can be done by building subtask-specific taggers on top of a shared text encoder. The shared encoder transforms the input sentence into a sequence of vector representations, and each tagger uses them to predict labels. By sharing the encoder, the taggers implicitly exploit inter-task dependencies.

The situation has changed with the introduction of the powerful pre-training/fine-tuning paradigm [12], however. It has been shown that Transformer-based models pre-trained on a huge raw corpus outperform existing neural models with large margins and tend to retain good performance even if a small amount of training data are given for the target task. This raises the possibility that pre-trained models are powerful enough to overshadow indirect signals from related subtasks.

From a practical point of view, it is non-negligible that pre-trained models are huge in size, with their success driving a race to build even larger models. If the model is fine-tuned separately for each subtask, we end up running multiple variants of a huge model at inference time. For this reason, we observe that huge pre-trained models give a new significance to multi-task learning: an effective way to reduce the model's size when we have multiple related tasks.

We conducted experiments to identify NEs, modality expressions, event classes, and event factuality, either separately or jointly. We found that BERT-CRF consistently outperformed non-neural models and BiLSTM-CRF, reconfirming the power of pre-training. Multi-task learning brought neither increase nor decrease in performance for BERT-CRF. Thus we conclude that BERT-CRF with multi-task learning is a practical solution.

II. TASK DESIGN

As shown in Table 1, we adopt the task design proposed by Matsuyoshi et al. [7]. We assume that the input sentence is segmented into words. Our task is to perform sequence tagging for the following four layers:

A. NAMED ENTITIES

21 NE types are defined for the *shogi* domain [8]. With the BIO tagging scheme [13], each word is given one of 43 (= 21 × 2 + 1) tags. Note that many NEs happen to be event mentions. For example, moves (**Mn**) and defensive formations (**Ca**) are likely to be events.

B. MODALITY EXPRESSIONS

8 types are defined for words and multi-word expressions that express factuality and other kinds of modalities. With the same BIO tagging scheme, they are mapped to 17 tags. **MEn** and **MEa** in Table 1 indicate that the target events are counterfactual and possibly factual, respectively. As an agglutinative

TABLE 2. Corpus specifications.

	#Sentences	#Words	#NEs	#Modality Exps	#Class Tags	#Factuality Tags
Wikipedia (raw)	25,074,606	712,048,970	-	-	-	-
shogi (annotated)	2,041	34,188	10,287	1,622	5,014	3,092

TABLE 3. Feature templates for sparse. x_n denotes the word in the current position while pos_n refers to the corresponding part-of-speech tag.

$$x_{n-2},\ x_{n-1},\ x_n,\ x_{n+1},\ x_{n+2},$$
$$(x_{n-2}, x_{n-1}),\ (x_{n-1}, x_n),\ (x_n, x_{n+1}),\ (x_{n+1}, x_{n+2}),$$
$$(x_{n-2}, x_{n-1}, x_n),\ (x_{n-1}, x_n, x_{n+1}),\ (x_n, x_{n+1}, x_{n+2}),$$
$$pos_{n-2},\ pos_{n-1},\ pos_n,\ pos_{n+1},\ pos_{n+2},$$
$$(pos_{n-2}, pos_{n-1}),\ (pos_{n-1}, pos_n),\ (pos_n, pos_{n+1}),\ (pos_{n+1}, pos_{n+2}),$$
$$(pos_{n-2}, pos_{n-1}, pos_n),\ (pos_{n-1}, pos_n, pos_{n+1}),\ (pos_n, pos_{n+1}, pos_{n+2})$$

TABLE 4. Hyper-parameters for BiLSTM-CRF.

Dimension of word embeddings	128
Number of BiLSTM layers	1
Dimension of the LSTM hidden layer	128
Dropout rate	0.25
Initial learning rate	0.001
Mini-batch size	20
Number of epochs	100

TABLE 5. Hyper-parameters for BERT-CRF.

Pre-training step	
Dimension of word embeddings	768
Number of Transformer layers	12
Dimension of the hidden layer	768
Number of self-attention heads	12
Dropout rate	0.1
Initial learning rate	0.0001
Mini-batch size	16
Number of epochs	30
Fine-tuning step	
Dropout rate	0.25
Initial learning rate	0.00002
Mini-batch size	20
Number of epochs	100

TABLE 6. Model performance on the four subtasks. Best scores are marked in bold.

Layer	Model	F1	Prec.	Recl.
NE	**Linear CRF**	0.874	0.894	0.856
	PWNER	0.877	**0.902**	0.854
	BiLSTM-CRF	0.871	**0.902**	0.843
	+multi	0.865	0.892	0.839
	BERT-CRF	**0.901**	0.898	**0.903**
	+multi	0.891	0.885	0.897
Modality	**Linear CRF**	0.751	0.769	0.734
	PWNER	0.774	**0.844**	0.716
	BiLSTM-CRF	0.776	0.825	0.733
	+multi	0.770	0.806	0.737
	+MEF	0.765	0.819	0.718
	BERT-CRF	**0.828**	0.829	**0.828**
	+multi	0.812	0.831	0.795
	+MEF	0.823	0.827	0.819
Event class	**Linear CRF**	0.636	0.691	0.589
	PWNER	0.738	0.786	0.696
	BiLSTM-CRF	0.710	0.758	0.669
	+multi	0.695	0.757	0.642
	+MEF	0.692	0.755	0.640
	BERT-CRF	**0.810**	0.804	**0.817**
	+multi	0.809	**0.811**	0.807
	+MEF	0.807	0.810	0.805
Factuality	**Linear CRF**	0.554	0.598	0.517
	PWNER	0.728	0.793	0.674
	BiLSTM-CRF	0.667	0.779	0.587
	+multi	0.675	0.773	0.603
	+MEF	0.677	0.788	0.596
	BERT-CRF	0.807	0.834	0.795
	+multi	0.811	**0.840**	0.795
	+MEF	**0.814**	0.824	**0.815**

language, Japanese often uses complex sequences of function words as modality expressions. There are also some predicates that quantify the degree of factuality of their arguments, and hence modality expressions can simultaneously be event mentions ("break" in this example). For ease of annotation, modality expressions are not explicitly linked to the corresponding event mentions, not to mention their scopes.

C. EVENT CLASSES

One of 8 tags is assigned to the head word of an event mention and the **O** tag to other words. The purpose of this layer is to distinguish factuality-bearing event mentions (e.g., **EVe**) from others. For example, grammaticalized verbs that do not warrant factuality statuses are given **EVf** tags.

D. EVENT FACTUALITY

One of 6 tags, such as **FNc** (certain−) and **FPr** (probable+), is assigned to the head word of a factuality-bearing event mention while other words are given **O** tags.

III. PROPOSED METHOD

Fig. 1 shows an overview of the proposed neural network model. To solve the four related subtasks introduced in Section II, we adopt multi-task learning that enables parameter sharing. We build task-specific CRF taggers on top of a shared encoder.

The input word sequence, $x_1, x_2, \cdots, x_N$, is converted into a sequence of word embeddings, $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N$, using a lookup table. The vector sequence is fed into the encoder to obtain $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N$, or vector representations of the input sequence.

For the encoder, we test (1) BiLSTM and (2) BERT. BiLSTM is a combination of a forward LSTM and a backward LSTM. LSTM [14] is a powerful extension to recurrent neural networks and is capable of capturing long-distance dependencies. Combining two LSTM units, BiLSTM makes use of both left and right contexts. For brevity, let $\mathrm{LSTM_f}$ be the blackbox forward LSTM. At time t, it takes $\boldsymbol{e}_t$ and its previous output $\overrightarrow{\boldsymbol{h}}_{t-1}$ as input and outputs $\overrightarrow{\boldsymbol{h}}_t$. The backward LSTM is defined in an analogous way. Combining

TABLE 7. Tag-wise statistics of the performances on NEs. Linear, LSTM, and BERT refer to Linear CRF, BiLSTM-CRF, and BERT-CRF, respectively.

Tag	Freq.	Model	F1	Prec.	Recl
Tu	1664	**Linear**	0.998	0.998	0.998
		PWNER	0.998	0.998	0.999
		LSTM	0.996	0.996	0.997
		+multi	0.996	0.995	0.997
		BERT	0.993	0.998	0.990
		+multi	0.996	0.996	0.997
Po	1465	**Linear**	0.983	0.988	0.978
		PWNER	0.997	0.996	0.998
		LSTM	0.995	0.996	0.995
		+multi	0.995	0.994	0.997
		BERT	0.995	0.993	0.998
		+multi	0.995	0.992	0.997
Pi	1817	**Linear**	0.987	0.984	0.990
		PWNER	0.981	0.978	0.984
		LSTM	0.983	0.983	0.983
		+multi	0.981	0.976	0.986
		BERT	0.979	0.978	0.981
		+multi	0.982	0.978	0.987
Ps	30	**Linear**	0.787	0.851	0.764
		PWNER	0.543	0.602	0.528
		LSTM	0.706	0.778	0.694
		+multi	0.636	0.733	0.592
		BERT	0.759	0.769	0.815
		+multi	0.772	0.826	0.759
Mc	151	**Linear**	0.995	1.000	0.990
		PWNER	0.983	0.995	0.973
		LSTM	0.992	0.989	0.995
		+multi	0.992	0.990	0.995
		BERT	0.992	0.988	0.995
		+multi	0.989	1.000	0.978
Mn	124	**Linear**	0.628	0.738	0.570
		PWNER	0.558	0.691	0.480
		LSTM	0.540	0.590	0.509
		+multi	0.542	0.639	0.488
		BERT	0.611	0.612	0.633
		+multi	0.533	0.589	0.512
Me	70	**Linear**	0.312	0.387	0.281
		PWNER	0.259	0.467	0.196
		LSTM	0.315	0.630	0.237
		+multi	0.276	0.422	0.233
		BERT	0.416	0.459	0.416
		+multi	0.265	0.350	0.231

Tag	Freq.	Model	F1	Prec.	Recl
St	56	**Linear**	0.903	0.927	0.883
		PWNER	0.734	0.889	0.679
		LSTM	0.781	0.905	0.726
		+multi	0.587	0.695	0.592
		BERT	0.799	0.819	0.814
		+multi	0.810	0.876	0.797
Ca	39	**Linear**	0.907	0.919	0.925
		PWNER	0.787	0.892	0.780
		LSTM	0.713	0.760	0.740
		+multi	0.813	0.883	0.817
		BERT	0.847	0.861	0.908
		+multi	0.685	0.746	0.712
Ev	76	**Linear**	0.406	0.560	0.415
		PWNER	0.320	0.510	0.285
		LSTM	0.454	0.577	0.489
		+multi	0.295	0.463	0.333
		BERT	0.524	0.587	0.562
		+multi	0.419	0.569	0.421
Ee	153	**Linear**	0.416	0.505	0.361
		PWNER	0.298	0.419	0.236
		LSTM	0.468	0.532	0.431
		+multi	0.431	0.494	0.395
		BERT	0.507	0.520	0.516
		+multi	0.435	0.416	0.473
Re	83	**Linear**	0.794	0.897	0.741
		PWNER	0.771	0.883	0.704
		LSTM	0.818	0.911	0.763
		+multi	0.753	0.808	0.715
		BERT	0.822	0.823	0.831
		+multi	0.769	0.768	0.789
Ph	44	**Linear**	0.571	0.630	0.550
		PWNER	0.520	0.543	0.512
		LSTM	0.544	0.602	0.508
		+multi	0.518	0.569	0.508
		BERT	0.685	0.684	0.697
		+multi	0.524	0.590	0.492
Pa	53	**Linear**	0.627	0.694	0.595
		PWNER	0.415	0.588	0.345
		LSTM	0.545	0.613	0.526
		+multi	0.515	0.625	0.464
		BERT	0.608	0.675	0.574
		+multi	0.543	0.672	0.508

Tag	Freq.	Model	F1	Prec.	Recl
Pq	15	**Linear**	0.415	0.500	0.370
		PWNER	0.347	0.450	0.308
		LSTM	0.372	0.426	0.343
		+multi	0.125	0.125	0.125
		BERT	0.531	0.552	0.552
		+multi	0.425	0.542	0.385
Hu	900	**Linear**	0.899	0.921	0.878
		PWNER	0.902	0.933	0.873
		LSTM	0.888	0.919	0.860
		+multi	0.884	0.915	0.856
		BERT	0.940	0.944	0.936
		+multi	0.925	0.930	0.920
Ti	399	**Linear**	0.898	0.894	0.903
		PWNER	0.923	0.914	0.932
		LSTM	0.926	0.928	0.925
		+multi	0.897	0.894	0.902
		BERT	0.902	0.894	0.911
		+multi	0.906	0.899	0.914
Ac	1425	**Linear**	0.781	0.778	0.786
		PWNER	0.805	0.795	0.815
		LSTM	0.752	0.844	0.679
		+multi	0.753	0.813	0.702
		BERT	0.879	0.873	0.885
		+multi	0.865	0.846	0.887
Ap	87	**Linear**	0.325	0.490	0.272
		PWNER	0.379	0.599	0.299
		LSTM	0.325	0.415	0.297
		+multi	0.340	0.485	0.296
		BERT	0.622	0.623	0.662
		+multi	0.544	0.630	0.516
Ao	316	**Linear**	0.363	0.531	0.280
		PWNER	0.454	0.553	0.391
		LSTM	0.475	0.580	0.421
		+multi	0.417	0.543	0.348
		BERT	0.559	0.550	0.579
		+multi	0.552	0.557	0.579
Ot	1320	**Linear**	0.758	0.780	0.737
		PWNER	0.751	0.789	0.718
		LSTM	0.749	0.773	0.729
		+multi	0.736	0.758	0.716
		BERT	0.796	0.791	0.802
		+multi	0.781	0.751	0.818

the two, BiLSTM computes $\boldsymbol{h}_t$ as follows:

$$\begin{aligned}\boldsymbol{h}_t &= \overrightarrow{\boldsymbol{h}}_t \oplus \overleftarrow{\boldsymbol{h}}_t \\ &= \mathrm{LSTM}_{\mathrm{f}}(\boldsymbol{e}_t, \overrightarrow{\boldsymbol{h}}_{t-1}) \oplus \mathrm{LSTM}_{\mathrm{b}}(\boldsymbol{e}_t, \overleftarrow{\boldsymbol{h}}_{t+1}),\end{aligned} \tag{1}$$

where $\oplus$ is the vector concatenation operation.

BERT (Bidirectional Encoder Representations from Transformers) [12] is a modern pre-trained language representation model known for achieving state-of-the-art performance for a wide range of tasks. Since BERT is pre-trained on a large raw corpus, we expect it to complement small annotated data.[1]

For each subtask $m \in M$, the task-specific CRF [16] takes $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N$ as the input and produces tagging decisions $\boldsymbol{y}_m = y_{m,1}, y_{m,2}, \cdots, y_{m,N}$. $\boldsymbol{h}_t$ is first linearly transformed into $\boldsymbol{o}_{m,t}$, whose dimension equals the number of tag types.

$$\boldsymbol{o}_{m,t} = \mathrm{softmax}(\boldsymbol{W}_m \boldsymbol{h}_t + \boldsymbol{b}_m) \tag{2}$$

$\boldsymbol{o}_m$ is then used to calculate the probability of $\boldsymbol{y}_m$:

$$p(\boldsymbol{y}_m | \boldsymbol{o}_m, \boldsymbol{T}_m) = \frac{\prod_{t=1}^{N+1} \exp(o_{m,t}^{y_{m,t}} + T_m^{y_{m,t-1}, y_{m,t}})}{\sum_{y'_m \in Y_m} \prod_{t=1}^{N+1} \exp(o_{m,t}^{y'_{m,t}} + T_m^{y'_{m,t-1}, y'_{m,t}})}, \tag{3}$$

where $o_{m,t}^{y_{m,t}} \in \mathbb{R}$ is the score for the output tag $y_{m,t}$ according to $\boldsymbol{o}_m$, and $T_m^{y_{m,t-1}, y_{m,t}} \in \mathbb{R}$ is the score of transition from $y_{m,t-1}$ to $y_{m,t}$. At $t = 0$, the special token BOS (beginning of sentence) is assigned to $y_{m,t}$. Similarly, the special token EOS (end of sentence) is assigned to $y_{m,N+1}$.

Let D_m be the training data for task m. The task-specific objective function is defined as

$$\mathrm{NLL}_m = -\sum_{D_m} \log p(\boldsymbol{y}_m | \boldsymbol{o}_m, \boldsymbol{T}_m). \tag{4}$$

Finally, we define the objective function as a weighted sum of the task-specific objective functions:

$$\mathrm{NLL} = \sum_{m \in M} \alpha_m \mathrm{NLL}_m, \tag{5}$$

where $\alpha_m \geq 0$ and $\sum_{m \in M} \alpha_m = 1$. Here we employ the multiple gradient descent algorithm (MGDA) [17], and α_m is automatically tuned at each backward step.

IV. EVALUATION

A. EXPERIMENTAL SETTINGS

Table 2 summarizes the corpus specifications. We used Japanese Wikipedia for pre-training and the *shogi* commentary corpus [7], [8] for evaluation. We used automatic word segmentation by KyTea [18] for the former and gold standard word segmentation for the latter.

The *shogi* commentary corpus was annotated with event factuality and other linguistic phenomena. For evaluation, the dataset was partitioned into ten roughly equal-sized subsets. Out of these subsets, eight were employed for training,

[1] In preliminary experiments, we also tested transfer learning from the latest version of the BCCWJ modality corpus [15]. It was a balanced corpus covering multiple domains. Although it was annotated with event class and factuality tags that were fully compatible with those of Matsuyoshi et al. [7], no annotation was available for modality expressions. We found no significant improvement with transfer learning, however.

TABLE 8. Tag-wise statistics of the performances on modality expressions (left), event classes (center), and event factuality (right). Linear, LSTM, and BERT refer to Linear CRF, BiLSTM-CRF, and BERT-CRF, respectively.

Tag	Freq.	Model	F1	Prec.	Recl
MEy	49	**Linear**	0.130	0.140	0.129
		PWNER	0.000	0.000	0.000
		LSTM	0.155	0.283	0.115
		+multi	0.054	0.083	0.040
		+MEF	0.090	0.233	0.056
		BERT	0.190	0.226	0.179
		+multi	0.069	0.150	0.050
		+MEF	0.075	0.093	0.070
MEa	224	**Linear**	0.557	0.676	0.476
		PWNER	0.607	0.726	0.525
		LSTM	0.613	0.674	0.568
		+multi	0.625	0.681	0.582
		+MEF	0.656	0.757	0.585
		BERT	0.693	0.673	0.720
		+multi	0.692	0.717	0.676
		+MEF	0.686	0.696	0.688
ME0	158	**Linear**	0.648	0.588	0.729
		PWNER	0.604	0.635	0.581
		LSTM	0.602	0.671	0.559
		+multi	0.565	0.588	0.551
		+MEF	0.590	0.616	0.572
		BERT	0.708	0.702	0.721
		+multi	0.658	0.640	0.687
		+MEF	0.713	0.675	0.765
MEm	21	**Linear**	0.817	0.889	0.770
		PWNER	0.413	0.556	0.344
		LSTM	0.672	0.778	0.622
		+multi	0.706	0.778	0.659
		+MEF	0.649	0.778	0.581
		BERT	0.854	0.944	0.826
		+multi	0.758	0.833	0.733
		+MEF	0.743	0.833	0.715
MEn	269	**Linear**	0.729	0.681	0.797
		PWNER	0.791	0.852	0.747
		LSTM	0.806	0.830	0.795
		+multi	0.797	0.834	0.773
		+MEF	0.766	0.815	0.731
		BERT	0.860	0.876	0.851
		+multi	0.866	0.893	0.846
		+MEF	0.862	0.880	0.852
MEp	692	**Linear**	0.905	0.900	0.912
		PWNER	0.928	0.930	0.927
		LSTM	0.913	0.926	0.903
		+multi	0.913	0.909	0.918
		+MEF	0.904	0.907	0.903
		BERT	0.941	0.935	0.948
		+multi	0.936	0.935	0.938
		+MEF	0.941	0.935	0.947
MEf	59	**Linear**	0.383	0.680	0.305
		PWNER	0.662	0.867	0.557
		LSTM	0.628	0.757	0.562
		+multi	0.617	0.783	0.566
		+MEF	0.622	0.818	0.536
		BERT	0.618	0.746	0.570
		+multi	0.560	0.775	0.498
		+MEF	0.666	0.792	0.597
MEh	150	**Linear**	0.646	0.768	0.573
		PWNER	0.547	0.706	0.454
		LSTM	0.624	0.807	0.526
		+multi	0.635	0.739	0.569
		+MEF	0.581	0.777	0.490
		BERT	0.812	0.809	0.830
		+multi	0.689	0.695	0.703
		+MEF	0.720	0.698	0.752

Tag	Freq.	Model	F1	Prec.	Recl
EVa	39	**Linear**	0.100	0.167	0.071
		PWNER	0.022	0.050	0.014
		LSTM	0.156	0.250	0.119
		+multi	0.183	0.317	0.132
		+MEF	0.150	0.320	0.117
		BERT	0.300	0.501	0.241
		+multi	0.329	0.560	0.237
		+MEF	0.424	0.503	0.433
EVq	111	**Linear**	0.034	0.131	0.023
		PWNER	0.288	0.657	0.202
		LSTM	0.314	0.570	0.237
		+multi	0.283	0.452	0.219
		+MEF	0.219	0.384	0.173
		BERT	0.583	0.703	0.587
		+multi	0.606	0.643	0.629
		+MEF	0.668	0.688	0.687
EVi	707	**Linear**	0.164	0.500	0.102
		PWNER	0.361	0.549	0.271
		LSTM	0.427	0.533	0.361
		+multi	0.427	0.480	0.392
		+MEF	0.426	0.484	0.390
		BERT	0.616	0.618	0.623
		+multi	0.612	0.646	0.606
		+MEF	0.562	0.681	0.548
EVp	7	**Linear**	0.095	0.071	0.143
		PWNER	0.000	0.000	0.000
		LSTM	0.500	0.500	0.500
		+multi	0.429	0.429	0.429
		+MEF	0.571	0.571	0.571
		BERT	0.667	0.643	0.714
		+multi	0.286	0.286	0.286
		+MEF	0.381	0.357	0.429
EVs	4	**Linear**	0.000	0.000	0.000
		PWNER	0.000	0.000	0.000
		LSTM	0.000	0.000	0.000
		+multi	0.000	0.000	0.000
		+MEF	0.000	0.000	0.000
		BERT	0.000	0.000	0.000
		+multi	0.000	0.000	0.000
		+MEF	0.000	0.000	0.000
EVe	3092	**Linear**	0.686	0.671	0.703
		PWNER	0.805	0.795	0.815
		LSTM	0.757	0.791	0.730
		+multi	0.745	0.813	0.690
		+MEF	0.739	0.804	0.686
		BERT	0.862	0.858	0.873
		+multi	0.863	0.864	0.870
		+MEF	0.858	0.855	0.872
EVc	293	**Linear**	0.639	0.765	0.568
		PWNER	0.717	0.843	0.631
		LSTM	0.709	0.770	0.670
		+multi	0.720	0.785	0.681
		+MEF	0.704	0.759	0.666
		BERT	0.708	0.742	0.689
		+multi	0.700	0.730	0.682
		+MEF	0.732	0.794	0.698
EVf	761	**Linear**	0.769	0.853	0.701
		PWNER	0.802	0.863	0.750
		LSTM	0.803	0.815	0.795
		+multi	0.784	0.816	0.757
		+MEF	0.801	0.833	0.773
		BERT	0.858	0.843	0.878
		+multi	0.848	0.843	0.855
		+MEF	0.854	0.849	0.862

Tag	Freq.	Model	F1	Prec.	Recl
FPc	2645	**Linear**	0.600	0.603	0.600
		PWNER	0.777	0.805	0.751
		LSTM	0.720	0.813	0.653
		+multi	0.729	0.805	0.671
		+MEF	0.728	0.814	0.661
		BERT	0.851	0.869	0.843
		+multi	0.853	0.873	0.846
		+MEF	0.854	0.860	0.860
FPr	233	**Linear**	0.054	0.333	0.030
		PWNER	0.297	0.564	0.204
		LSTM	0.287	0.442	0.219
		+multi	0.313	0.475	0.243
		+MEF	0.326	0.492	0.250
		BERT	0.531	0.602	0.542
		+multi	0.540	0.571	0.535
		+MEF	0.573	0.581	0.585
FPs	35	**Linear**	0.000	0.000	0.000
		PWNER	0.465	0.667	0.373
		LSTM	0.244	0.400	0.189
		+multi	0.207	0.350	0.160
		+MEF	0.290	0.500	0.210
		BERT	0.514	0.675	0.443
		+multi	0.287	0.433	0.244
		+MEF	0.606	0.720	0.547
FNc	140	**Linear**	0.059	0.242	0.037
		PWNER	0.327	0.550	0.245
		LSTM	0.272	0.446	0.212
		+multi	0.235	0.356	0.179
		+MEF	0.289	0.517	0.207
		BERT	0.573	0.626	0.570
		+multi	0.616	0.658	0.584
		+MEF	0.598	0.609	0.636
FNr	34	**Linear**	0.000	0.000	0.000
		PWNER	0.000	0.000	0.000
		LSTM	0.000	0.000	0.000
		+multi	0.000	0.000	0.000
		+MEF	0.000	0.000	0.000
		BERT	0.137	0.190	0.150
		+multi	0.149	0.283	0.125
		+MEF	0.100	0.125	0.100
FNs	4	**Linear**	0.000	0.000	0.000
		PWNER	0.000	0.000	0.000
		LSTM	0.000	0.000	0.000
		+multi	0.000	0.000	0.000
		+MEF	0.333	0.333	0.333
		BERT	0.000	0.000	0.000
		+multi	0.000	0.000	0.000
		+MEF	0.000	0.000	0.000

one for development, and the remaining one for evaluation. Hyper-parameters were tuned using the development set. This procedure was repeated ten times, with a distinct subset chosen for evaluation in each run. We averaged micro F-1 scores of the ten outcomes.

B. MODELS

As a baseline non-neural model, we used **Linear CRF**, a CRF model with sparse hand-crafted features. It directly outputs tags for each of the four subtasks. As shown in Table 3, the features used were word and POS n-grams ($n \leq 3$) taking into account three words on the both sides as well as the target word itself. We used KyTea [18] to obtain POS tags. We also tested **PWNER**,[2] an off-the-shelf non-neural sequence labeling tool. **PWNER** was used by the creators of the annotated corpus to provide initial evaluations [7].

For the proposed neural network-based method, we tested **BiLSTM-CRF** and **BERT-CRF**. Their sentence encoders were BiLSTM and BERT, respectively. The models with multi-task learning (**+multi**) were compared against the models without it (unmarked). In the absence of multi-task learning, we obtained fine-tuned BERT models for individual subtasks, leading to practical challenges in their simultaneous execution. We also tested multi-task learning focusing on modality expressions, event classes, and event factuality (**+MEF**), or in other words, excluding named entity recognition.

In the pre-training step for BERT, we first segmented sentences to word sequences with KyTea [18] and then split each word into subwords by WordPiece [19], with the vocabulary size of 32,000. We used Adam [20] as the optimization algorithm.

[2] http://www.lsta.media.kyoto-u.ac.jp/resource/tool/PWNER/home-e.html

The details of network configurations are shown in Tables 4 and 5. For the **BiLSTM-CRF** model, 64 $\times$ 2 dimensional vectors are fed into a CRF layer for each task because the outputs of the forward and backward LSTMs are concatenated into one. For **BERT-CRF**, 768 dimensional vectors are fed into a CRF layer for each. In both models, dropout [21] was applied to each layer.

C. RESULTS AND DISCUSSION

The main results are shown in Table 6. Overall, **BERT-CRF** performed the best. It consistently beat **BiLSTM-CRF** with large margins. Non-neural **PWNER** worked surprisingly well, especially for event classes and event factuality.

Multi-task learning (**+multi**) yielded no clear gains or losses. **BERT-CRF+multi** performed relatively poorly for NE. As indicated by Table 2, the number of NE tags were much larger than the numbers of event-related tags. These motivated us to try **+MEF**, but it brought no consistent changes either.

For further analyses, we calculated tag-wise statistics. For the detailed description of tag types, please refer to Mori et al. [8] and Matsuyoshi et al. [7]. Tables 7 and 8 show the results of the four subtasks. In these tables, "Freq." indicates the number of instances for each tag type in the corpus. Most noticeable is that the frequencies are skewed toward some tag types. **BERT-CRF** performed relatively well for low-frequency tags, demonstrating the effectiveness of pre-training. Again, we observed no clear trend for the effect of **+multi**.

As we discussed in Section I, multi-task learning, or more precisely, parameter sharing among subtasks, has a practical advantage in computational efficiency because running multiple variants of fine-tuned BERT at inference time can be prohibitively expensive. The absence of any performance gain or decline due to multi-task learning leads us to the conclusion that **BERT-CRF** combined with multi-task learning stands as the pragmatic selection for event factuality analysis.

V. CONCLUSION

We proposed a deep neural network model for Japanese event factuality analysis. We combined pre-training, multi-task learning, and other techniques to achieve high performance for this important task. We reconfirmed that pre-training was highly effective in enhancing accuracy. While multi-task learning does not improve accuracy, it saves us from running multiple variants of huge fine-tuned models. Our experiments led us to conclude that BERT-CRF combined with multi-task learning represents the practical choice for performing event factuality analysis.

Although our experiments employed a *shogi* (Japanese chess) commentary corpus, the proposed method is applicable to other domains if the task is designed in a similar way. In the future, we will apply the proposed approach to other domains, possibly with knowledge transfer from the *shogi* domain. We would also like to use event factuality analysis to tackle the symbol grounding problem since *shogi* is characterized by multiple possible worlds.

REFERENCES

[1] R. Saurí and J. Pustejovsky, "Are you sure that this happened? Assessing the factuality degree of events in text," *Comput. Linguistics*, vol. 38, no. 2, pp. 261–299, Jun. 2012.

[2] M.-C. de Marneffe, C. D. Manning, and C. Potts, "Did it happen? The pragmatic complexity of veridicality assessment," *Comput. Linguistics*, vol. 38, no. 2, pp. 301–333, Jun. 2012.

[3] A. Lotan, A. Stern, and I. Dagan, "TruthTeller: Annotating predicate truth," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, Atlanta, Georgia, 2013, pp. 752–757.

[4] K. Lee, Y. Artzi, Y. Choi, and L. Zettlemoyer, "Event detection and factuality assessment with non-expert supervision," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 1643–1648.

[5] G. Stanovsky, J. Eckle-Kohler, Y. Puzikov, I. Dagan, and I. Gurevych, "Integrating deep linguistic features in factuality prediction over unified datasets," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, Vancouver, BC, Canada, 2017, pp. 352–357.

[6] R. Rudinger, A. S. White, and B. Van Durme, "Neural models of factuality," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, New Orleans, Louisiana, 2018, pp. 731–744.

[7] S. Matsuyoshi, H. Kameko, Y. Murawaki, and S. Mori, "Annotating modality expressions and event factuality for a Japanese chess commentary corpus," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga, Eds. Miyazaki, Japan, May 2018, pp. 2475–2481.

[8] S. Mori, J. Richardson, A. Ushiku, T. Sasada, H. Kameko, and Y. Tsuruoka, "A Japanese chess commentary corpus," in *Proc. 10th Int. Conf. Lang. Resour. Eval.*, Portoroz, Slovenia, 2016, pp. 1415–1420.

[9] Y. Kamioka, K. Narita, J. Mizuno, M. Kanno, and K. Inui, "Semantic annotation of Japanese functional expressions and its impact on factuality analysis," in *Proc. 9th Linguistic Annotation Workshop*, Denver, CO, USA, 2015, pp. 52–61.

[10] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.

[11] A. Søgaard and Y. Goldberg, "Deep multi-task learning with low level tasks supervised at lower layers," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, pp. 231–235.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[13] T. K. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proc. 7th Conf. Natural Lang. Learn. HLT-NAACL*, 2003, pp. 142–147. [Online]. Available: https://ifarm.nl/erikt/papers/

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[15] S. Matsuyoshi, M. Eguchi, C. Sao, K. Murakami, K. Inui, and Y. Matsumoto, "Annotating event mentions in text with modality, focus, and source information," in *Proc. 7th Int. Conf. Lang. Resour. Eval.*, Valletta, Malta, 2010, pp. 1456–1463.

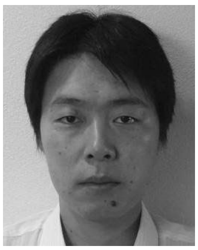
[16] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.* San Francisco, CA, USA: Morgan Kaufmann, 2001, pp. 282–289.

[17] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31. Long Beach, CA, USA: Curran Associates, 2018, pp. 525–536. [Online]. Available: https://www.proceedings.com/39083.html

[18] G. Neubig, Y. Nakata, and S. Mori, "Pointwise prediction for robust, adaptable Japanese morphological analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Human Lang. Technol.*, Portland, OR, USA, Jun. 2011, pp. 529–533.

[19] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, and J. Klingner, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.
[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

HIROTAKA KAMEKO received the B.E., M.E., and Ph.D. degrees from the University of Tokyo, in 2013, 2015, and 2018, respectively. He is currently an Assistant Professor with the Academic Center for Computing and Media Studies, Kyoto University. His research interests include natural language processing and game AI. He is a member of IPSJ and ANLP.



YUGO MURAWAKI received the B.S., M.S., and Ph.D. degrees from Kyoto University, in 2006, 2008, and 2011, respectively. From 2013 to 2015, he was an Assistant Professor with the Graduate School and the Faculty of Information Science and Electrical Engineering, Kyushu University. In 2016, he re-joined the Graduate School of Informatics, Kyoto University, as an Assistant Professor. Subsequently, he was a Senior Lecturer, in 2020, and further ascended to the role of an Associate Professor, in 2023, where he currently holds his position. His research interests include natural language processing and computational linguistics. He is a member of the Association for Natural Language Processing and the Information Processing Society of Japan.

SUGURU MATSUYOSHI received the B.S. degree in mathematics and the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2003, 2005, and 2008, respectively. Then, he was an Assistant Professor with the Nara Institute of Science and Technology, University of Yamanashi, and the University of Electro-Communications. Since 2021, he has been a Lecturer with the Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology. His research interests include computational linguistics and natural language processing. He is a member of Information Processing Society of Japan and the Association for Natural Language Processing.

SHINSUKE MORI received the B.S., M.S., and Ph.D. degrees in electrical engineering from Kyoto University, Kyoto, Japan, in 1993, 1995, and 1998, respectively. Then, he joined the Tokyo Research Laboratory of International Business Machines Company Ltd., (IBM). Since 2007, he has been an Associate Professor with the Academic Center for Computing and Media Studies, Kyoto University. He is currently a Professor. His research interests include computational linguistics and natural language processing. He received the IPSJ Yamashita SIG Research Award, in 1997, the IPSJ Best Paper Award, in 2010 and 2013, and the 58th OHM Technology Award from Promotion Foundation for Electrical Science and Engineering, in 2010. He is a member of the Information Processing Society of Japan, the Association for Natural Language Processing, the Database Society of Japan, and the Association for Computational Linguistics.