Supplemental Materials of Single Model for Influenza Forecasting of Multiple Countries by Multi-task Learning

Taichi Murayama(✉)[1[0000−0003−1148−711X]], Shoko Wakamiya[1[0000−0002−9371−1340]], and Eiji Aramaki[1[0000−0003−0201−3609]]

Nara Institute of Science and Technology (NAIST), Japan
{taichi.murayama.mk1,wakamiya,aramaki}@is.naist.jp

A Correlation between time series of ILI rates in each country

Fig. 1 presents the flu time series from 2016/29th week to 2019/30th week in five countries. This represents that the flu time series in each country exhibits strong seasonality and therefore, holds strong similarity. Table 1 displays the Pearson correlations of the time series among the respective countries. These values suggest that the influenza-like illness (ILI) rates in the five countries with different cultures, locations, and languages have a moderate correlation with one another (almost all correlations are over 0.6). The high correlation gives us a strong motivation to addresses the challenge of flu forecasting for various countries in one model.

B Related Work

Flu forecasting is a type of time-series prediction. Time-series prediction tasks are mainly divided into univariate and multivariate types [16]. In research on time-series prediction, various models suitable for each task have been proposed over time. Owing to the rapid development of neural networks, many models have been based on these, particularly convolutional neural networks (CNNs) [21, 2, 5] and recurrent neural networks (RNNs) [23, 22, 14], which capture the temporal variation. In recent years, there has been an increase in time-series prediction models using "attention" (Transformer) to achieve state-of-the-art performance in multiple natural language processing applications [1, 24]. Attention generally aggregates temporal features using dynamically generated weights, thereby enabling the network to focus on significant time steps in the past directly. For example, [9, 15, 17] are time-series prediction models based on attention. Although numerous models and methods have been proposed to achieve higher prediction accuracy, it is difficult to apply them to the influenza forecasting problem in a simple manner. This is owing to the problem setting of flu forecasting; that is, the future flu volume is estimated from two major resources: historical ILI data and UGC data. This problem belongs to a multivariable problem with one

width=.95

Table 1. Pearson correlation between time series of ILI rates in each country (the only time series in AU shifts forward in 22 weeks to match the peak point of the US and AU).

	US	JP	UK	AU	FR
US		0.793	0.614	0.840	0.745
JP			0.592	0.751	0.527
UK				0.772	0.693
AU					0.681
FR					

width=.95

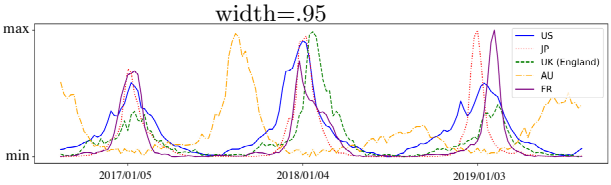

Fig. 1. Time series of ILI rates in five countries, applied to min-max normalization, from 2016/29th week to 2019/30th week.

objective variable and various explanatory variables, and most models are not designed for this problem type. Thus, we need to develop a time-series model suitable for flu and not apply state-of-the-art prediction models.

Numerous models relating to flu forecasting have been proposed to date. Prior to the emergence of online UGCs, compartmental models such as SIR [13] and IDEA [19], as well as statistical models such as autoregressive models [25] using historical ILI data, were used extensively for flu forecasting. With the development of the Internet, some researches [11, 7] revealed that online UGCs such as search queries and social media posts are useful resources, same as in the field of influenza prediction [10]. Although certain studies have used only one resource of either historical ILI data or online UGCs, the majority of studies proposed supervised methods using online UGCs together with historical ILI data simultaneously as input. Most of these approaches do not consider the characteristics of the data types, but simply simultaneous inputting, when learning the model. Several previous models [20, 18] based on statistical methods have been developed to exploit the characteristics of the different input data. However, these studies exhibit certain disadvantages, such as the necessity of long training terms or a small degree of improved accuracy. Our method based on neural networks to capture the latent features can achieve the best accuracy in flu forecasting by means of an appropriate combination method of two inputs: historical ILI data and search query data.

Moreover, we aim to learn a single flu forecasting model for multiple countries as multi-task learning. Multi-task learning, which was introduced by [6], improves the generalization, and achieves superior efficiency and prediction accuracy by using a shared representation from related tasks. It is used extensively in various areas, such as natural language processing (NLP) [3, 12] and time series [4, 8]. The fundamentals of multi-task learning were presented in detail in [6]. Zou et al. [26] tackled a similar problem to ours and proposed a multi-task model based on linear and Gaussian regression to forecast the flu volume in the following two problem settings: several states in the US, and two countries, namely the US and England. Our multi-task model and task further develop the above

Table 2. Model forecasting performances for JP.

		Input			1-week		2-week		3-week		4-week		5-week	
Term	Model	Multi	Historical	Query	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
	GRU w/o sq		✓		3.998	0.922	4.876	0.885	5.313	0.863	5.836	0.835	6.275	0.810
	Transformer		✓		3.217	0.949	4.202	0.936	4.864	0.895	5.715	0.849	6.353	0.819
	*Proposed w/o sq		✓		3.087	0.956	3.811	0.935	4.362	0.903	5.518	0.853	6.052	0.824
2017/30th	GRU		✓	✓	3.412	0.939	4.019	0.923	5.223	0.915	5.982	0.826	6.164	0.813
—	ARGO		✓	✓	3.317	0.941	—	—	—	—	—	—	—	—
2018/29th	Two-stage		✓	✓	4.727	0.898	4.866	0.851	5.653	0.822	6.301	0.818	6.814	0.787
	*Proposed_single		✓	✓	2.517	0.964	3.218	0.944	3.688	0.934	4.898	0.884	5.822	0.836
	MTEN	✓		✓	3.697	0.922	—	—	—	—	—	—	—	—
	GRU_multi	✓	✓	✓	3.261	0.944	3.901	0.882	5.072	0.820	6.552	0.776	6.752	0.756
	*Proposed_multi2	✓	✓	✓	2.665	0.965	3.163	**0.951**	3.569	0.938	4.298	0.910	4.652	0.895
	*Proposed_multi5	✓	✓	✓	**2.429**	**0.970**	**2.878**	**0.951**	**3.411**	**0.941**	**4.057**	**0.920**	**4.423**	**0.905**
	GRU w/o sq		✓		3.526	0.928	4.304	0.883	4.670	0.874	6.034	0.778	6.980	0.711
	Transformer		✓		3.211	0.934	3.985	0.909	4.733	0.881	6.287	0.790	6.526	0.710
	*Proposed w/o sq		✓		2.458	0.951	3.931	0.909	5.621	0.810	5.998	0.794	6.802	0.720
2018/30th	GRU		✓	✓	3.384	0.931	4.250	0.884	4.654	0.884	6.070	0.773	7.209	0.699
–	ARGO		✓	✓	4.034	0.920	—	—	—	—	—	—	—	—
2019/29th	Two-stage		✓	✓	3.222	0.934	4.022	0.899	5.044	0.834	6.174	0.780	6.516	0.715
	*Proposed_single		✓	✓	**2.390**	**0.962**	4.004	0.908	5.878	0.808	6.134	0.783	6.802	0.731
	MTEN	✓		✓	5.023	0.863	—	—	—	—	—	—	—	—
	GRU_multi	✓	✓	✓	3.913	0.905	4.567	0.858	4.988	0.823	6.001	0.769	6.949	0.705
	*Proposed_multi2	✓	✓	✓	2.800	0.955	3.557	**0.927**	**4.790**	**0.868**	5.694	0.814	6.483	0.789
	*Proposed_multi5	✓	✓	✓	2.897	0.952	**3.556**	0.926	4.800	0.854	**5.673**	**0.815**	**6.258**	**0.793**
	GRU w/o sq		✓		3.429	0.709	5.821	0.164	7.128	-0.252	7.436	-0.362	7.378	-0.340
	Transformer		✓		3.567	0.723	5.746	0.301	7.085	-0.320	7.515	-0.489	7.929	-0.702
	*Proposed w/o sq		✓		3.054	0.784	4.938	0.398	6.917	-0.179	7.005	-0.221	7.322	-0.317
2019/30th	GRU		✓	✓	3.303	0.722	5.655	0.297	7.360	-0.357	8.085	-0.514	8.622	-0.729
–	ARGO		✓	✓	3.411	0.740	—	—	—	—	—	—	—	—
2020/29th	Two-stage		✓	✓	3.569	0.731	4.585	0.401	6.812	-0.166	7.383	-0.333	7.890	-0.586
	*Proposed_single		✓	✓	3.326	0.726	4.251	0.419	6.600	0.005	7.240	-0.290	7.292	-0.311
	MTEN	✓		✓	3.922	0.563	—	—	—	—	—	—	—	—
	GRU_multi	✓	✓	✓	2.916	0.812	3.930	0.610	5.488	0.332	6.008	0.108	7.006	-0.153
	*Proposed_multi2	✓	✓	✓	2.936	0.830	3.724	0.649	5.400	0.312	6.135	0.072	6.972	-0.197
	*Proposed_multi5	✓	✓	✓	**2.800**	**0.858**	**3.566**	**0.715**	**5.004**	**0.407**	**5.839**	**0.282**	**6.849**	**-0.121**

***** indicates the variation in the proposed model. Bold indicates the best score in each metric and each term.

in two aspects: we tackle flu forecasting in five countries, each of which differs in terms of the area or language, and we apply not a simple model such as a statistical model, but our novel neural network-based model for multi-task learning to achieve higher accuracy and long-term forecasting.

C Experimental results for JP

The result for JP is presented in Tables 2. This result indicates that the proposed model (particularly our multi-task model) outperformed most baseline methods, confirming the benefits of the model architecture and multi-task learning, same as the US. The proposed models (**Proposed_single**, **Proposed_multi2**, and **Proposed_multi5**) achieved the best scores with respect to the terms, metrics, and any-ahead forecasts. These results reveal that the architecture in the proposed model is useful for flu forecasting. **Proposed_single** achieved the best score among the models without multi-task learning in almost all terms except for 2018 to 2019 in JP, in which it exhibited the best score in the near-ahead fore-

Table 3. Examples of selected search queries by translation-based and WT-based methods.

English query		Translation-based	WT-based
fever␣and␣flu	ja	発熱␣と␣インフルエンザ	熱␣インフル
	fr	fiévre␣et␣grippe	fiévre␣grippe
the␣flu	ja	インフルエンザ	インフル
	fr	la␣grippe	grippe
symptoms␣of␣flu	ja	インフルエンザ␣の␣症状	徴候␣インフル
	fr	symptômes␣de ␣la ␣grippe	infection␣grippe

"␣" indicates space for search.

cast, whereas it had a lower score in the far-ahead forecast than the GRU-based models. Same as the US, the high degree of the score improvement in **Proposed_multi2** and **Proposed_multi5** compared to Proposed_single demonstrated the usefulness of the multi-task learning, except in 1-week forecast.

In terms of the comparison of models without and with search queries, the experimental results for the flu forecast in JP indicate that the change from GRU w/o sq to GRU resulted in an average improvement of -0.026 points in the RMSE, and of -0.030 points in the R^2. However, the change from Proposed w/o sq to Propose_single resulted in an average improvement of 0.187 points in the RMSE, and of 0.012 points in the R^2. Same as the US, this suggests that the search query data resulted in the GRU-based models exhibits low or worse improvement scores by adding them, however the proposed model, with a well-crafted architecture for the search query data input, achieved a significantly improved score.

D Examples of search queries based on each selection method

Examples of search queries based on each selection method are presented in Table 3. Compared to the translation-based method, the WT-based method exhibited many similar points in the selection results, although several aspects differed. For example, in Japanese, the abbreviation representation "インフル (I-N-FU-LU)" is selected as the corresponding word for "flu," and not "インフルエンザ (I-N-FU-LU-E-N-ZA)." In French, "infection" is selected as the corresponding word for "symptoms."

E Effect of the country embedding

To examine the country embedding effectiveness, we validated the degree of improvement of the two proposed models without and with country embedding (Proposesd_multi5 w/o CE and Proposesd_multi5). The experimental results

Table 4. Comparison of forecasting performances of Proposesd_multi5 and Proposesd_multi5 without country embedding (CE) in US and JP from 2017/30th week to 2018/29th week.

Country	Method	1-week		2-week		3-week		4-week		5-week	
		RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
US	Proposesd_multi5 w/o CE	0.299	0.952	0.535	0.935	0.788	0.832	0.902	0.800	0.994	0.743
	Proposesd_multi5	**0.237**	**0.986**	**0.498**	**0.941**	**0.692**	**0.837**	**0.805**	**0.832**	**0.942**	**0.770**
JP	Proposesd_multi5 w/o CE	**2.400**	**0.970**	3.367	0.939	3.701	0.933	4.449	0.909	5.102	0.875
	Proposesd_multi5	2.429	**0.970**	**2.878**	**0.951**	**3.411**	**0.941**	**4.057**	**0.920**	**4.423**	**0.905**

for the flu forecast in US and JP are presented in Table 4. Proposesd_multi5 achieved relatively better scores than those without country embedding in two countries. This suggests that the country embedding, which is the initial latent representation of two GRUs regarding the time series of the search queries and deseasonalized component for each country, exhibits improvement scores.

F Effect of the attention in the proposed model

The attention mechanism in the proposed model not only successfully combines the search query data, but can also provide a broad understanding of which queries affect the forecast and in what manner. Fig. 2 presents the visualization of the attention weight of each search query for the forecast in the US from 2017/30th week to 2018/29th week. A large change occurred in the attention weight around 2018/8th week, at which time the flu became an epidemic, whereas the attention weight in each query was almost constant except during this period. This means that the information from search queries is useful in forecasting the flu during an epidemic period. In particular, "flu and fever" and "symptoms of flu" are useful in the flu forecasting model because they have a large attention weight. It is interesting that the weight of "flu and fever" was large despite that of "fever flu," with the same meaning, being small. Using attention to visualize useful search queries offers the potential to aid in determining the resources of the input for creating the forecasting model in situations where a useful search query has not been determined.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)

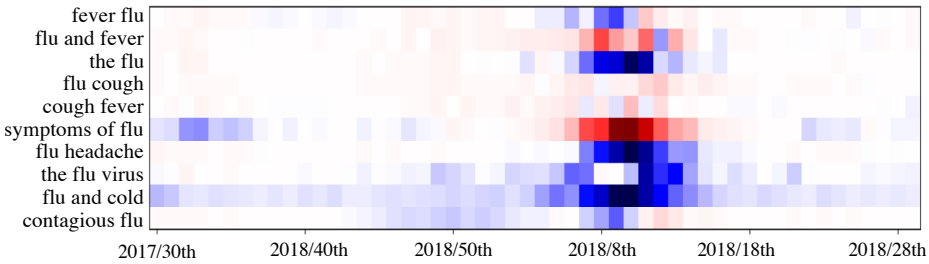

Fig. 2. Visualization of attention weight of each search query for forecast in US from 2017/30th week to 2018/29th week. Red indicates a large attention weight, whereas blue indicates a small attention weight.

3. Bohnet, B., Nivre, J.: A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In: Proc. of EMNLP. pp. 1455–1465 (2012)
4. Bonilla, E.V., Chai, K.M., Williams, C.: Multi-task gaussian process prediction. In: Proc. of NeurIPS. pp. 153–160 (2008)
5. Borovykh, A., Bohte, S., Oosterlee, C.W.: Conditional time series forecasting with convolutional neural networks. arXiv preprint arXiv:1703.04691 (2017)
6. Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)
7. Choi, H., Varian, H.: Predicting the present with google trends. Economic record **88**, 2–9 (2012)
8. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proc. of SIGKDD. pp. 109–117 (2004)
9. Fan, C., et al.: Multi-horizon time series forecasting with temporal attention learning. In: Proc. of SIGKDD. pp. 2527–2535 (2019)
10. Ginsberg, J., et al.: Detecting influenza epidemics using search engine query data. Nature **457**(7232), 1012–1014 (2009)
11. Goel, S., et al.: Predicting consumer behavior with web search. PNAS **107**(41), 17486–17490 (2010)
12. Hatori, J., et al.: Incremental joint approach to word segmentation, pos tagging, and dependency parsing in chinese. In: Proc. of ACL. pp. 1045–1053 (2012)
13. Hethcote, H.W.: The mathematics of infectious diseases. SIAM review **42**(4), 599–653 (2000)
14. Lai, G., et al.: Modeling long-and short-term temporal patterns with deep neural networks. In: Proc.of SIGIR. pp. 95–104 (2018)
15. Li, S., et al.: Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In: Proc. of NeurIPS. pp. 5243–5253 (2019)
16. Lim, B., Zohren, S.: Time series forecasting with deep learning: A survey. arXiv preprint arXiv:2004.13408 (2020)
17. Lim, B., et al.: Temporal fusion transformers for interpretable multi-horizon time series forecasting. arXiv preprint arXiv:1912.09363 (2019)
18. Murayama, T., et al.: Robust two-stage influenza prediction model considering regular and irregular trends. PloS one **15**(5), e0233126 (2020)
19. Nasserie, T., et al.: Seasonal Influenza Forecasting in Real Time Using the Incidence Decay With Exponential Adjustment Model. Open Forum Infectious Diseases **4**(3) (2017)

20. Ning, S., Yang, S., Kou, S.: Accurate regional influenza epidemics tracking using internet search data. Scientific reports **9**(1), 1–8 (2019)
21. Oord, A.v.d., et al.: Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
22. Rangapuram, S.S., et al.: Deep state space models for time series forecasting. In: Proc. of NeurIPS. pp. 7785–7794 (2018)
23. Salinas, D., et al.: Deepar: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting **36**(3), 1181–1191 (2020)
24. Vaswani, A., et al.: Attention is all you need. In: Proc. of NeurIPS. pp. 5998–6008 (2017)
25. Viboud, C., et al.: Prediction of the spread of influenza epidemics by the method of analogues. American Journal of Epidemiology **158**(10), 996–1006 (2003)
26. Zou, B., Lampos, V., Cox, I.: Multi-task learning improves disease models from web search. In: Proc. of WebConf. pp. 87–96 (2018)