# BNU: A BALANCE-NORMALIZATION-UNCERTAINTY MODEL FOR INCREMENTAL EVENT DETECTION

*Jia Li*[1]*, Yunyan Zhang*[2]*, Yifan Yang*[2]*, Zhicheng An*[3]*, Yefeng zheng*[2,†]

[1] EECS, Beijing, Peking University
[2]Jarvis Lab, Shenzhen, Tencent
[3] Tsinghua-Berkeley Shenzhen Institute, Shenzhen, Tsinghua University
lijiaa@pku.edu.cn, {yunyanzhang,tobyfyang,yefengzheng}@tencent.com, azc19@mails.tsinghua.edu.cn

## ABSTRACT

Event detection is challenging in real-world application since new events continually occur and old events still exist which may result in repeated labeling for old events. Therefore, incremental event detection is essential where a model continuously learns new events and meanwhile prevents performance from degrading on old events. Although existing incremental event detection models achieve impressive performance, they face the data imbalance problem between old classes and new classes, and have the knowledge transfer problem which cannot adequately utilize the knowledge provided by the previous model and data. To this end, we propose a **B**alance-**N**ormalization-**U**ncertainty (**BNU**) model to address above problems. Specifically, in order to mitigate the adverse effects of data imbalance, we incorporate a balanced fine-tuning stage and a cosine normalization module. Meanwhile, we consider aleatoric uncertainty to preserve previous knowledge while training for new events. Experimental results show that our proposed method resolves the above challenges effectively and achieves consistent and significant performance on ACE and TAC KBP datasets.

***Index Terms*—** Incremental Learning, Event Detection, Balanced Fine-tuning, Cosine Normalization, Uncertainty.

## 1. INTRODUCTION

Event detection (ED) aims to detect event triggers from sentences and classify them into specific types. For example, **Death** event triggered by "**execute**" should be detected in the following sentence: Some 30 policemen were captured, tied up and **executed** in cold blood against the walls.

Until now, ED is limited in practical applications since new events continuously occur and old events still exist in the real world. An ideal model needs to not only learn new events incrementally but also prevent the model performance from degrading on old events. The simplest solution is to add new events into training dataset and retrain the model from scratch with updated data. However, it is impractical due to high requirements for computing resources. Another option is to label old events contained in new event data, but this procedure is time-consuming for repetitively labeling.

Therefore, incremental event detection is required to address the multiple event problem. A straightforward method for incremental ED is to optimize the model on new event data, but after the adaptation to the new training set, the model usually achieves poor performance on old events, which is called catastrophic forgetting [1, 2, 3]. Recently, many complicated methods have been developed to avoid catastrophic forgetting. One common way is to maintain significant parameters of the old model that is trained on previous classes [4, 5, 6]. Another way is to reserve some representative examples in each old class and retrains the model on it [7, 8, 9, 10]. The state-of-the-art method KCN [8] for incremental ED belongs to the latter way, but faces two issues. (i) **Imbalance Problem:** the training set size of old events is much smaller than the new event data, which results in bias for the new event; (ii) **Knowledge Transfer:** it can not effectively utilize the knowledge from old event data.

To address the aforementioned challenges, we propose a **B**alance-**N**ormalization-**U**ncertainty (**BNU**) model, which is shown in Figure 1. To reduce the adverse effect of the imbalance problem, we introduce a balanced fine-tuning stage and a cosine normalization module, which constrains the bias for new data. To confine the knowledge transfer problem, we consider an aleatoric uncertainty loss to transfer information of the previous model to the current model which represents how much the model is uncertain about the prediction due to the data. We evaluate our method on ACE and TAC KBP datasets. Experiments show that our BNU model can significantly improve the performance on both benchmarks.

Our main contributions are as follows:

- We propose a novel model BNU to address the data imbalance and the knowledge transfer problems for incremental ED.
- To mitigate the adverse effects of data imbalance, we introduce balanced fine-tuning and cosine normalization. Meanwhile, we consider an aleatoric uncertainty loss to

† Corresponding authors.

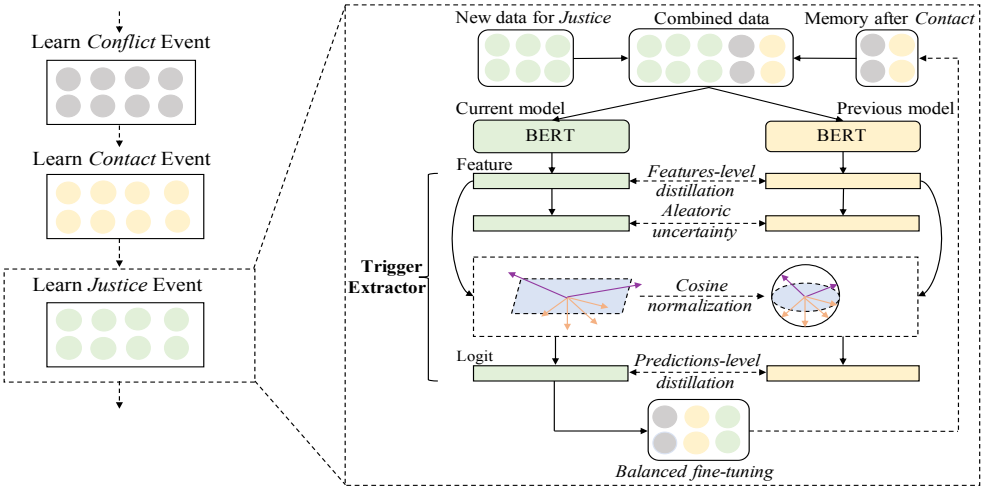



**Fig. 1**: Illustration of our Balance-Normalization-Uncertainty (BNU) model.

alleviate the knowledge transfer problem.

- Experimental results show that our model can achieve consistent and significant improvements on ACE and TAC KBP datasets.

## 2. THE PROPOSED METHOD

### 2.1. Problem Formalization

ED aims to detect an event triggered by an event trigger which is a word or phrase in sentences. As described above, new events continually occur and old events still exist. Therefore, a practical ED method should adapt to the incremental learning scenario. We formulate incremental ED as a sequence tagging task. To be specific, given an input sequence, the ED method labels each token in the sequence with pre-defined event classes. Formally, given a model trained on an old dataset $\mathbb{Z}_o = \{\mathbb{Z}_1, \mathbb{Z}_2, ..., \mathbb{Z}_{n-1}\}$, we aims to learn a unified method for both old event classes $\mathbb{Z}_o$ and new event class $\mathbb{Z}_n$, based on a new dataset $\mathbb{Z} = \hat{\mathbb{Z}}_o \cup \mathbb{Z}_n$, where $\mathbb{Z}_i$ represents the $i$-th event class set. $\mathbb{Z}$ is a large dataset that covers the new event set $\mathbb{Z}_n$, meanwhile $\hat{\mathbb{Z}}_o \in \mathbb{Z}_o$ reserves a small quantity of representative old event samples. After training the first $n$-th events, we evaluate the model on the test dataset that contains all old event classes.

### 2.2. BNU Model

To address above challenges, we propose a Balance-Normalization-Uncertainty (**BNU**) model for incremental ED. BNU consists of four important modules as follows.

**Event Trigger Extractor.** Recently, pre-trained language models, such as BERT [11] and RoBERTa [12], have shown impressive performance in many NLP tasks [13, 14]. In this work, we select BERT as encoder since it is pre-trained on a large amount of unlabeled corpus and shows strong ability in language representation and understanding. Given an updated dataset $\mathbb{Z} = \{(s_i, y_i), 1 \leq i \leq M\}$, where $M$ is the number of labeled sentences, $s_i = \{w_{i,1}, ..., w_{i,N_i}\}$ is a sentence with $N_i$ words, and $y_i = \{y_{i,1}, ..., y_{i,N_i}\}$ is the sequence of labels. For each example, we feed it into BERT and add a multi-perception layer on BERT to acquire the predicted score $\hat{y}_{i,j}$. Finally, the objective function of the event trigger extractor is formulated as:

$$\mathcal{L}_{ed} = -\frac{1}{|M|}\frac{1}{|N_j|}\sum_{i=1}^{M}\sum_{j=1}^{N_j} y_{i,j} log(\hat{y}_{i,j}) \tag{1}$$

Following [15, 8], we utilize a hierarchical distillation (i.e., feature-level and prediction-level distillations) for incremental ED since knowledge distillation is a common way to alleviate forgetting of the previous knowledge. The feature-level distillation loss function is formulated as:

$$\mathcal{L}_{fl} = \frac{1}{|M|}\sum_{i=1}^{M}\sum_{j=1}^{N_j} 1 - \langle \hat{f}_{i,j}, f_{i,j} \rangle \tag{2}$$

where $\hat{f}_{i,j}$ and $f_{i,j}$ are $l_2$-normalized features extracted by previous model and current model, and $\langle \hat{f}_{i,j}, f_{i,j} \rangle$ means cosine similarity. We compute prediction-level distillation loss as follows:

$$\mathcal{L}_{pl} = -\frac{1}{|M|}\sum_{i=1}^{M}\sum_{j=1}^{N_j}\sum_{z=1}^{n-1} \hat{\tau}_{i,j,z} log(\tau_{i,j,z}), \tag{3}$$

$$\hat{\tau}_{i,j,z} = \frac{e^{\hat{p}_{i,j,z}/T}}{\sum_{q=1}^{n-1} e^{\hat{p}_{i,j,q}/T}}, \tau_{i,j,z} = \frac{e^{p_{i,j,z}/T}}{\sum_{q=1}^{n-1} e^{p_{i,j,q}/T}}, \tag{4}$$

$$p_{i,j,z} = \theta_z^T f_{i,j} + b_z, \hat{p}_{i,j,z} = \hat{\theta}_z^T \hat{f}_{i,j} + \hat{b}_z \tag{5}$$

where $\hat{p}_{i,j,z}$ and $p_{i,j,z}$ are the output logits (i.e., the outputs before softmax) of the previous and current model for the $j$-th token, respectively. $n-1$ represents the number of observed event classes in the old model and $T$ is the temperature scalar.

**Balanced Fine-tuning.** Although reserving a small amount of old class data is helpful for incremental learning, the number of old class samples is smaller than that of the new class samples, which may bias the model towards the new class. To deal with this unbalanced training scenario, we introduce a balanced fine-tuning [16] stage with a balanced set of samples that has the same number of representative

examples for peer class, including old and new event classes. To select representative examples, we first compute the mean representation of each class:

$$x_u = \frac{1}{|Q_u|} \sum_{i=1}^{Q_u} s_i \tag{6}$$

where $Q_u$ is the total number of the $u$-th class set and $s_i$ is the representation of the $i$-th sentence belonging to the $u$-th class. In this work, we utilize the representation of [CLS] token as the sentence representation. We then rank all examples according to their distances to the mean representation, and select the first $w$ examples as the representative samples for each class.

**Cosine Normalization.** Besides balanced fine-tuning, we further introduce a cosine normalization module, which is inspired by [16], to address the imbalance problems further. As described in event trigger extractor, the predicted probability of $\hat{y}_{i,j}$ is computed as follows:

$$\hat{y}_{i,j} = \frac{e^{\theta_z^T f_{i,j} + b_z}}{\sum_{z=1}^{n-1} e^{\theta_z^T f_{i,j} + b_z}} \tag{7}$$

where $\theta$ and $b$ are the weights and the bias vectors in the last layer, respectively. However, because of the imbalance problem, the magnitudes of the embedding and the biases for the new class are significantly higher than those for the old classes, which results in the bias for the new class. Therefore, we introduce cosine normalization in the last layer:

$$\hat{y}_{i,j} = \frac{e^{\eta \langle \theta_z^*, f_{i,j}^* \rangle}}{\sum_{z=1}^{n-1} e^{\eta \langle \theta_z^*, f_{i,j}^* \rangle}} \tag{8}$$

where $v^* = v / \|v\|_2$ denotes the $l_2$-normalized vector. The learnable scalar $\eta$ controls the range of softmax distribution since the $\langle v_1^*, v_2^* \rangle$ is in range of $[-1, 1]$.

After the cosine normalization, the scores before softmax in the last layer can mimic the scores after softmax for two reasons. First, due to cosine normalization, the scores before softmax are range in $[-1, 1]$ which is the same range as the scores after softmax. Second, the old and new models have different scalars $\eta$. Therefore, the prediction-level distillation loss in Eq. 3 can be updated as:

$$\mathcal{L}_{pl}^* = -\sum_{z=1}^{n-1} \left\| \langle \theta_z^*, f_{i,j}^* \rangle - \langle \hat{\theta}_z^*, \hat{f}_{i,j}^* \rangle \right\|. \tag{9}$$

**Aleatoric Uncertainty.** The objective of incremental learning is to preserve the information of old classes while training for the new class. We introduce an aleatoric uncertainty loss to further transfer the information of old classes to the new model. The aleatoric uncertainty describes the uncertainty of model prediction caused by the data. Thus, data uncertainty of the old model should be similar to the new model as much as possible. In this work, we utilize a linear layer to compute aleatoric uncertainty followed by the encoder. Finally, we define the aleatoric distillation loss $\mathcal{L}_u$ as follows:

$$\mathcal{L}_{au} = \sum_{i=1}^{n-1} \left\| (\sigma_{o,i})^2 - (\sigma_{n,i})^2 \right\|^2 \tag{10}$$

where $\sigma_{o,i}$ and $\sigma_{n,i}$ are the aleatoric uncertainty of previous model and current model. $n-1$ represents the number of observed event classes in the old model.

**Total Loss.** Combining the losses presented above, the final objective function is formulated as:

$$\mathcal{L} = \mathcal{L}_{ed} + \mathcal{L}_{fl} + \mathcal{L}_{pl}^* + \mathcal{L}_{au}. \tag{11}$$

Based on the balanced fine-tuning module and cosine normalization module, our method effectively addresses the data imbalance problem. Meanwhile, aleatoric uncertainty loss alleviates the knowledge transfer challenge.

## 3. EXPERIMENTS

### 3.1. Datasets and Evaluation

We evaluate our model on the ACE 2005[1] and TAC KBP 2017[2] datasets. Following the previous work [8], we select the top 10 event classes according to their frequency, and split them into different sets. Meanwhile, for the TAC KBP dataset, we also utilize the top 10 event classes for the incremental ED task. We employ *Average F1* score and *Whole F1* score as evaluation metrics [8]. *Average F1* score means the average F1 scores of all classes (i.e., $\frac{1}{k}\sum_{i=1}^{k} F1_i$), and *Whole F1* represents the F1 score on the all observed event classes. We select 10 and 50 representative examples from each event class into memory on the ACE 2005 dataset and the TAC KBP dataset, respectively.

### 3.2. Baselines

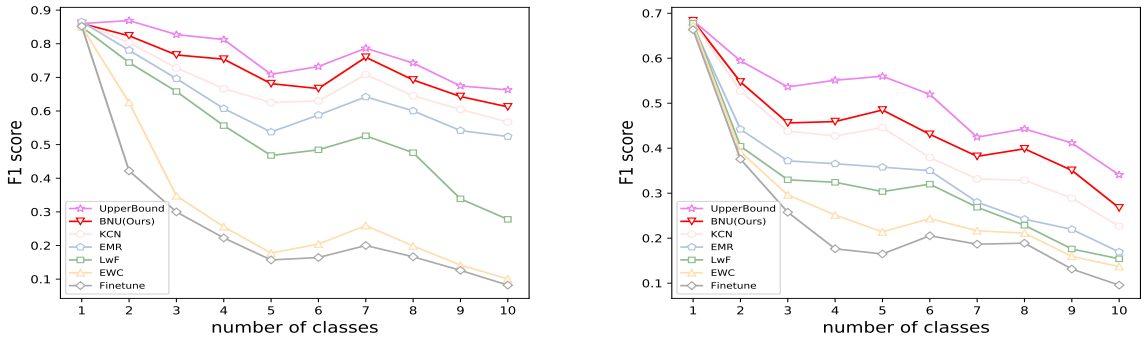
We compare our method with the following baselines: **Finetune:** The method only finetunes the pre-trained model on new data. **EWC [3]:** This is a parameter-based method, which reserves the optimal parameters of the previous model while training for new classes. **LwF [17]:** The method uses new class data to train the model while preserving the original capabilities for incremental learning. **EMR [18]:** The method reserves a few representative examples of old classes, and combines them with examples of the new class to continue training. **KCN [8]:** The method proposes a hierarchical distillation module to alleviate catastrophic forgetting. **UpperBound:** The model is trained on training examples of all observed event classes, which can achieve the best performance because of optimizing the model on all observed sets.

### 3.3. Overall Results

The F1 scores on ACE and TAC KBP datasets are shown in Figure 2. We also list the average F1 and Whole F1 in Table 1. From the results, we can observe that 1) Our method achieves significant improvement over all baselines. For example, BNU achieves 4.52% and 5.08% improvement in whole F1 on ACE and TAC KBP datasets, respectively, compared with a competitive model KCN, which indicates that our proposed method BNU is effective for incremental ED; 2) Another interesting observation is that the performance of models tends to decline

[1]https://catalog.ldc.upenn.edu/LDC2006T06

[2]https://tac.nist.gov/2017/KBP/data. html




(a) ACE dataset (b) TAC KBP dataset

**Fig. 2**: The performance on the ACE (a) and TAC KBP (b) datasets. Our method achieves better performance than other models.

since the difficulty increases as the number of event classes increases.

**Table 1**: The average F1 (%) on all obversed classes ("Ave"), and whole F1 (%) on the whole test datset ("Whole").

| Method | ACE | | TAC KBP | |
|---|---|---|---|---|
| | Ave | Whole | Ave | Whole |
| Finetune | 26.92 | 8.21 | 24.26 | 7.56 |
| EWC [3] | 31.41 | 10.06 | 27.94 | 13.68 |
| LwF [17] | 53.79 | 27.78 | 31.87 | 15.42 |
| EMR [18] | 63.83 | 52.43 | 34.71 | 16.93 |
| KCN [8] | 68.39 | 56.71 | 41.76 | 22.66 |
| BNU (Ours) | **71.81** | **61.23** | **44.35** | **27.74** |

### 3.4. Discussion

**Ablation experiment.** We remove balanced fine-tuning (BF), cosine normalization (CN), and aleatoric uncertainty (AU) respectively to evaluate the effectiveness of each module. The detailed results are reported in Table 2. We can find that removing any module brings performance degradation, which demonstrates the effectiveness of each module for incremental ED task.

**Table 2**: Ablation studies by removing BF, CN, and AU.

| Model | ACE | | KBP | |
|---|---|---|---|---|
| | Avg | Whole | Avg | Whole |
| BNU (Ours) | **71.81** | **61.23** | **44.35** | **27.74** |
| w/o BF | 70.29 | 60.30 | 43.04 | 26.47 |
| w/o CN | 71.05 | 59.74 | 42.23 | 25.38 |
| w/o AU | 70.71 | 59.62 | 43.35 | 25.76 |

**The effect of the number of reserved samples.** We compare our method BNU with method KCN on the ACE dataset. Table 3 shows the effect of the number of reserved samples in each class. We can find that with the number of reserved samples increase, the performance of KCN and our BNU become better. In each size, the performance of BNU is superior to KCN, which again indicates the effectiveness of our BNU.

**Comparison with semi-supervised scenario.** To further verify the effectiveness and practicability of our model, we compare our BNU with the semi-supervised scenario. Since only new event class is labeled in the new class set, we use a previously trained model to predict the presence of old events, and train the model with the set annotated with old and new event classes. From Figure 3, we observed that our method is significantly superior to the semi-supervised setting which demonstrates our BNU is much more practical for ED.

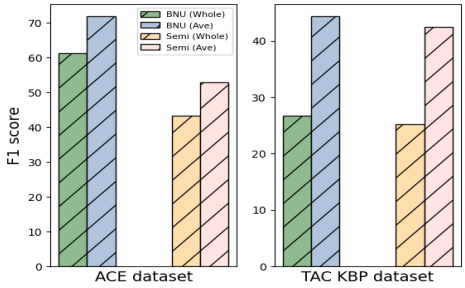



**Fig. 3**: Comparison of the proposed method with the semi-supervised setting. "Semi" is the semi-supervised setting.

**Table 3**: The effect of the number of reserved samples.

| Size | BNU (Ours) | | KCN | |
|---|---|---|---|---|
| | Avg | Whole | Avg | Whole |
| 10 | 71.81 | 61.23 | 68.39 | 56.71 |
| 20 | 73.65 | 63.86 | 69.88 | 58.43 |
| 30 | 75.04 | 65.46 | 71.21 | 59.28 |
| 40 | 75.57 | 66.07 | 72.03 | 61.24 |
| 50 | **76.14** | **66.79** | **72.87** | **61.95** |

## 4. CONCLUSION

In this paper, we propose a BNU model to address data imbalance and knowledge transfer problems in incremental ED task. To alleviate the imbalance problem, we incorporate the balanced fine-tuning stage and cosine normalization. To transfer the previous knowledge into the current model, aleatoric uncertainty loss is carefully introduced. Experimental results demonstrate that our model outperform previous state-of-the-art methods by a substantial margin.

## 5. REFERENCES


[1] Michael McCloskey and Neal J Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*, vol. 24, pp. 109–165. Elsevier, 1989.

[2] Robert M French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[3] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[4] Zixuan Ke, Hu Xu, and Bing Liu, "Adapting bert for continual learning of a sequence of aspect sentiment classification tasks," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 4746–4755.

[5] Thien Huu Nguyen and Ralph Grishman, "Event detection and domain adaptation with convolutional neural networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 365–371.

[6] Hongjoon Ahn, Donggyu Lee, Sungmin Cha, and Taesup Moon, "Uncertainty-based continual learning with adaptive regularization," *CoRR*, vol. abs/1905.11614, 2019.

[7] Natawut Monaikul, Giuseppe Castellucci, Simone Filice, and Oleg Rokhlenko, "Continual learning for named entity recognition," in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

[8] Pengfei Cao, Yubo Chen, Jun Zhao, and Taifeng Wang, "Incremental event detection via knowledge consolidation networks," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 707–717.

[9] David Lopez-Paz and Marc'Aurelio Ranzato, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, Eds., 2017, pp. 6467–6476.

[10] Amanda Rios and Laurent Itti, "Closed-loop memory GAN for continual learning," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus, Ed. 2019, pp. 3332–3338, ijcai.org.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[13] Da Yin, Tao Meng, and Kai-Wei Chang, "Sentibert: A transferable transformer-based architecture for compositional sentiment semantics," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, Eds. 2020, pp. 3695–3706, Association for Computational Linguistics.

[14] Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu, "SKEP: sentiment knowledge enhanced pre-training for sentiment analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, Eds. 2020, pp. 4067–4076, Association for Computational Linguistics.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[16] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin, "Learning a unified classifier incrementally via rebalancing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 831–839.

[17] Zhizhong Li and Derek Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[18] Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang, "Sentence embedding alignment for lifelong relation extraction," *arXiv preprint arXiv:1903.02588*, 2019.