KR2: Revisiting Pre-trained Language Models as Knowledge Resource

Yimin Fan[1]*, Nan Duan[2], Houqiang Li[1] and Ming Zhou[3]
[1]University of Science and Technology of China, [2]Microsoft Research Asia,
[4]Sinovation Ventures
fym0503@mail.ustc.edu.cn, lihq@ustc.edu.cn
{nanduan}@microsoft.com,
mingzhou926@hotmail.com

Abstract

There are growing interests in integrating external knowledge into language models among NLP researchers. While some works try to use explicit knowledge resources including knowledge graph and knowledgeable text, we propose **KR2** (**K**nowledge **R**esource and **R**eader) to utilize implicit information from pre-trained language models as external knowledge. When facing a specific NLP task, the intermediate representations of a frozen in-domain pre-trained language model are extracted and serve as **K**(nowledge) **Resource**, which can help improve the performance of the **K**nowledge **Reader** (task model). We conduct experiments on multilingual fine-tuning and physical commonsense reasoning tasks. Consistent gains are obtained compared with strong baselines. We also empirically compare our approach with knowledge distillation, a well-recognized method to transfer implicit knowledge between models, to illustrate the effectiveness of our approach.

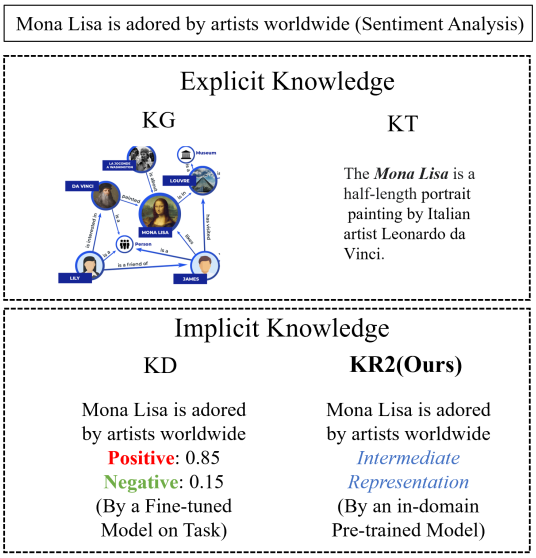

Figure 1: Comparison between our KR2 model with other types of knowledge.

1 Introduction

With the rapid development of computing devices and the increasing amount of available data, it is now much easier for NLP researchers to train large-scale language models and boost task performance. Recently, there has been growing interest in integrating *explicit and implicit* external knowledge into language models. For example, *explicit knowledge* including entities and relations in knowledge graphs are converted into contextualized embeddings (KG) or knowledgeable text (KT), and they are fused with the input text or intermediate representation of language models. This approach not only leads to significant gains on knowledge intensive tasks (Liu et al., 2019a; Zhang et al., 2019), but also helps with general Natural Language Understanding (NLU) (Xu et al., 2021). Meanwhile, researchers also find that the output of a higher-capacity or a better performing trained model can be used as *implicit knowledge* to benefit the target task. This approach is known as Knowledge Distillation (KD)(Hinton et al., 2015), which generalizes well among different model architectures and tasks.

In this paper, we explore new possibilities of incorporating implicit knowledge from existing language models when fine-tuning our models on downstream tasks. The differences between our approach and previous works are illustrated in Figure 1. As deep pre-trained models have "seen" massive amount of data and can be used for a wide range of tasks, the models are considered to contain diverse implicit knowledge inside. We call these models **KResource** (**K**(nowledge) **Resource**)). The task model, which is trained for combining knowledge from KResource and solving the downstream tasks, is referred to as

* Work is done during internship at Microsoft Research Asia.

KReader(**K**(nowledge) **Reader**)), so we name our method **KR2** (**K**nowledge **R**esource and **R**eader).

When facing a specific NLP task, we first extract embeddings from the frozen KResource models. The parameters of KResource models are frozen during the fine-tuning to better keep the knowledge inside the models. Though the size, architecture and pre-training purpose of KResource models may be different, the embeddings of KResource models show how these models perceive and represent the input sentences, so the embeddings can be considered a kind of implicit knowledge in the KResource models. We transform the embeddings of the KResource models and concat the embeddings from KResource with the input of KReader. The KReader model takes the input and is fine-tuned on the target task.

We conduct experiments on diverse settings and tasks, including multilingual fine-tuning and physical commonsense reasoning. Experiments show that implicit knowledge from KResource models can significantly boost the KReader performance on downstream tasks. We also compare our method with knowledge distillation, as they both have the purpose of using the *implicit knowledge* in one model (KResource/Teacher) to improve another model (KReader/Student). We find that when the Teacher/KResouce model is weaker than the KReader/Student model, our method performs significantly better than knowledge distillation, which demonstrates the usefulness of our method.

2 Approach

Given the input text with length T, the length of input tokens to KReader and KResource models are l_1 and l_2 respectively. Assuming KResource is a L layers Transformer encoder, the dimension of the hidden states of KResource model is h_2. The intermediate representation of KResource model is $\mathbf{H}^{0:L}_{KResource} \in R^{l_2 \times (L+1) \times h_2}$ size tensor (the word embeddings are also included).

The dimension of hidden states of the KReader model is h_1. As h_1 may be not equal to h_2, we need to project the intermediate representation of KResource model into $\mathbf{G}^0_{KReader} \in R^{l_2 \times h_1}$ dimension. There are several ways for the transformation. One natural way is to directly project the $(L+1) * h_2$ dimension tensor into h_1 dimensions. To avoid high computation complexity, we consider the alignment between positions in different layers.

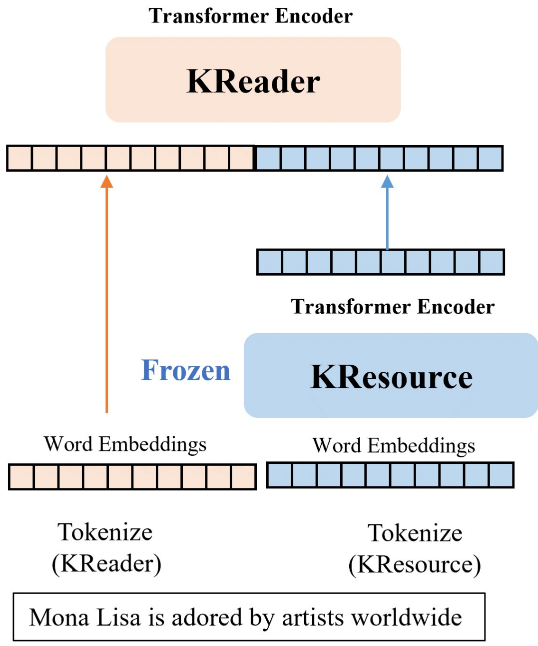

Figure 2: An illustration of our approach.

We define

$$\hat{\mathbf{H}}_{KResource} = \sum_{i=0}^{L} \omega_i \mathbf{H}^i_{KResource}$$

The fusion weight $\{\omega_i\}_{i=0}^{L}$ can either be trainable parameters, uniform distribution over some or all the layers, or one-hot vector selecting one specific layer. We will empirically study the effects of different strategies in Section 3. We denote the projection from dimension m to n (implemented by several layers of MLP) as $f_{m:n}$. The projection process can be written as

$$\mathbf{G}^0_{KReader} = f_{h_2:h_1}(\hat{\mathbf{H}}_{KResource})$$

In this paper, we only consider cases where $h_2 = h_1$, so we omit this projection process.

The input to the KReader model can be written as follows

$$\hat{\mathbf{H}}^0_{KReader} = \texttt{Concat}(\mathbf{H}^0_{KReader}, \mathbf{G}^0_{KReader})$$

The KReader model is then fine-tuned on the target task. The parameters of the KResource are kept frozen during training.

3 Experiments

We conduct experiments on diverse settings and tasks to verify the effectiveness of our methods. With appropriate choice of KResource, our method achieves significant gains compared with strong baselines.

Domain	Multilingual POS tagging						Physical Commonsense Reasoning			
Tasks	De	Ko	Ja	Zh	Ar	Es	SNLI	SWAG	PIQA	HellaSwag
Model Config	XLM-R/Monolingual LMs						BERT/OSCAR			
KReader	84.3	45.6	48.9	61.5	62.2	81.8	89.3	78.1	61.7	**38.7**
KResource	83.8	25.8	46.4	48.8	58.5	81.6	88.6	65.5	61.6	35.5
KReader2	84.6	46.2	49.5	61.5	62.0	82.1	89.2	77.8	62.7	38.0
KR2	**84.7**	**46.2**	**50.0**	**61.9**	**62.8**	**82.4**	**89.4**	**78.4**	**63.3**	38.0

Table 1: Performance of our method on multilingual POS tagging and physical commonsense reasoning tasks. Results are averaged over three seeds. Model Config A/B means A is KReader and B is KResource. KReader/KResource denotes directly fine-tuning KReader/KResource. KReader2 denotes using the same KResource as KReader. Hyperparameters of fine-tuning KReader and KResource are kept the same.

3.1 Settings

Multilingual Fine-tuning We use the XLM-Roberta (Conneau et al., 2020) base model as the KReader model because XLM-Roberta (Conneau et al., 2020) is a well-recognized strong and general multilingual encoder. Monolingual pre-trained BERT (Devlin et al., 2019a) and RoBERTa (Liu et al., 2019b) models are used as KResource. Though monolingual pre-trained models may have inferior performance compared with XLM-Roberta as a result of lack of computing resource, they have better monolingual tokenizers and are trained on monolingual corpora only, so we believe they contain more monolingual implicit knowledge. The details of these models are shown in Appendix A. We conduct experiments on multilingual Part of Speech tagging (Zeman et al., 2019) with translation data in that language. We train 10 epochs with batch size 32 and learning rate 2e-5.

Physical Commonsense Reasoning The data of Physical Commonsense Reasoning datasets mainly comes from image/video captioning datasets or descriptions of a physical process. We choose four datasets in this setting. SNLI (Bowman et al., 2015) is a Natural Language Inference (NLI) dataset that includes content from the Flickr 30k corpus (Plummer et al., 2016) and the VisualGenome corpus (Krishna et al., 2016). The PIQA (Bisk et al., 2019) dataset introduces the task of multiple choice physical commonsense reasoning. The SWAG dataset (Zellers et al., 2018) is also a multiple choice questions answer dataset. Each question is a video caption from LSMDC (Rohrbach et al., 2017) or ActivityNet Captions (Krishna et al., 2017), with four answer choices about what might happen next in the scene. The HellaSwag (Zellers et al., 2019) dataset is a more challenging and realistic verison of SWAG. All questions in HellaSwag are from ActivityNet Captions. We use BERT(Devlin et al., 2019a) as the KReader model and OSCAR encoder (Li et al., 2020) as the KResource model. OSCAR is initialized with BERT and primarily trained as a vision-language model. Though training on vision-language tasks may harm the performance of OSCAR on text-only tasks (catastrophic forgetting), OSCAR gets more implicit knowledge related to the visual and physical world which is likely to benefit physical commonsense reasoning tasks.

3.2 Results

In Table 1, we compare our method KR2 against several strong baselines, including directly fine-tuning the KReader model (KReader), directly fine-tuning the KResource model (KResource) and using the same KResource model as KReader model (KReader2). We find that even directly fine-tuning the KResource model doesn't outperform the KReader model, its intermediate representations can help to boost the performance of our method KR2. As KReader2 brings gains against the KReader baseline but still perform worse than KR2 model, we can see that the improvement of KR2 not only comes from integrating more features into model input, but also is a result of diverse implicit knowledge in KResource model.

4 Analysis

4.1 Choice of Layer Fusion Weight

In this part, we discuss different strategies for selecting the layer fusion weight. There are several potential strategies, including trainable weighted fusion, average over some or all the layers and selecting one specific layer. We conduct experiments in multilingual Part-of-Speech tagging fine-tuning tasks. We choose Germany, Chinese, Arabic and Spainish languages. The results are shown in Table

POS	De	Zh	Ar	Es	AVG
XLM-R/Monolingual LMs					
KReader	84.3	61.5	62.2	81.8	72.5
Layer 0	84.6	61.4	63.0	82.5	72.9
Layer 1	84.7	61.5	62.4	82.4	72.8
Layer 3	84.7	62.4	61.6	82.4	72.8
Layer 5	84.4	61.9	62.6	82.5	72.9
Layer 7	84.6	61.5	63.1	82.3	72.9
Layer 9	84.8	61.5	62.7	82.5	72.9
Layer 11	84.6	62.0	62.7	82.1	72.9
AVG(All)	84.5	61.9	62.7	82.3	72.9
AVG(Last 4)	84.7	61.9	62.8	82.4	**73.0**
Fuse	84.7	62.2	63.0	82.5	**73.1**

Table 2: Comparison between different layer fusion strategies. Results are averaged over three seeds. Trainable weights yield the best results. Layer 0 denotes word embeddings.

Domain	Multilingual POS tagging					
Tasks	De	Ko	Ja	Zh	Ar	Es
KReader(A)	84.3	45.6	48.9	61.5	62.2	81.8
KR2(B-A)	84.7	46.2	50.0	61.9	62.8	82.4
KD(B-A)	84.6	46.1	48.6	61.3	62.3	82.0
KResource(B)	83.8	25.8	46.4	48.8	58.5	81.6
KR2(A-B)	84.4	25.9	46.8	48.7	59.2	82.6
KD(A-B)	84.4	26.3	46.6	49.2	59.4	82.4

Table 3: Comparison between knowledge distillation (KD) and KR2. Results are averaged over three seeds. A-B means that A is the KResource/Teacher and B is the KReader/Student.

2. We find that embeddings from later layers contain more beneficial knowledge than front layers. To achieve stable improvements across different languages and simplify the training process, we choose the average of last four layers as the fusion strategy.

4.2 Comparison with Knowledge Distillation

Knowledge Distillation (Hinton et al., 2015) has been considered a very effective way to transfer implicit knowledge from one model to another, while our method has the same purpose. We empirically compare knowledge distillation with our approach. Details of knowledge distillation is shown in Appendix B. As shown in Table 3, we find that though knowledge distillation can improve the performance when the KResource/Teacher is weak, our method brings more stable and significant gains. When the KResource/Teacher is stronger than KReader/Student model, both KR2 and KD can help improve the model performance. Note that in knowledge distillation we use a KResource/Teacher model fine-tuned on the target task, in KR2 we use a frozen pre-trained KResource/Teacher model. When there exists many in-domain but weaker pre-trained models or fine-tuned teacher models are unavailable, our method is a simpler and more effective way to utilize the implicit knowledge inside these models.

5 Related Works

Explicit Knowledge Enhanced NLP Many works have tried to incorporate explicit knowledge from Knowledge Graph since the emergence of BERT(Devlin et al., 2019b). There are several directions in this field. Several works (Sun et al., 2019; Lauscher et al., 2020; Rosset et al., 2021; Xiong et al., 2019) utilize entity representation in the knowledge graph or use the entity as pretraining tasks for language models. Some other works (Sun et al., 2019; He et al., 2020; Yu et al., 2020; Wang et al., 2020a) perform joint training between knowledge graph neural networks and language models. These works directly use knowledge from knowledge graph training. We call these methods **KG**. There are several other works called **KT** that automatically convert entities and relations in knowledge graph into human-readable text and combine these text with the task input (Joshi et al., 2021; Liu et al., 2019a; Agarwal et al., 2021).

Knowledge Distillation Enhanced NLP Knowledge Distillation (Hinton et al., 2015) has proven to be a very effective method to transfer knowledge in one model to another. With knowledge distilled from high-capacity and well-performing teacher models, supreme performance can be obtained across a wide range of NLP tasks (Sanh et al., 2020; Jiao et al., 2020; Yin et al., 2021).

6 Conclusion

In this paper, we present a new way to integrate implicit knowledge from frozen pre-trained models into language model fine-tuning. Extensive empirical results validate the effectiveness of our approach. Comparison with knowledge distillation illustrates that our method is an effective way to transfer implicit knowledge between models.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bin He, Di Zhou, Jing Xie, Jinghui Xiao, Xin Jiang, and Qun Liu. 2020. Ppke: Knowledge representation learning by path-based pre-training.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding.

Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2021. Contextualized representations using textual encyclopedic knowledge.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing unsupervised pretraining models for word-level semantic similarity.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. K-bert: Enabling language representation with knowledge graph.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul Bennett, and Saurabh Tiwary. 2021. Knowledge-aware language model pretraining.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2020a. Kepler: A unified model for knowledge embedding and pre-trained language representation.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, and Kewei Tu. 2020b. Structure-level knowledge distillation for multilingual sequence labeling.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2019. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model.

Ruochen Xu, Yuwei Fang, Chenguang Zhu, and Michael Zeng. 2021. Does knowledge help general nlu? an empirical study.

Yichun Yin, Cheng Chen, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2021. Autotinybert: Automatic hyper-parameter optimization for efficient pre-trained language models.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2020. Jaket: Joint pre-training of knowledge graph and language understanding.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskait, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lee Hong, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubei, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareek, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu

Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Taksum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities.

A Choice of Monolingual Models

`https://huggingface.co/uklfr/gottbert-base` (German)
`https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased` (Spanish)
`https://huggingface.co/nghuyong/ernie-1.0` (Chinese)
`https://huggingface.co/asafaya/bert-base-Arabic` (Arabic)
`https://huggingface.co/monologg/kobert` (Korean)
`https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking` (Japanese)

B Details of Knowledge Distillation

Knowledge distillation on sequence labeling tasks like POS tagging has been studied (Wang et al., 2020b) and many non-trivial tricks have been proposed. In this paper, to keep the comparability between general KD and our method, we choose to distil knowledge in token level rather than structure level. The KD loss is set to $L_{hard} + 0.5L_{soft}$ to avoid noise in training.