Supplementary Material: Mining Knowledge for Natural Language Inference from Wikipedia Categories

Mingda Chen[3*] **Zewei Chu**[2*] **Karl Stratos**[1] **Kevin Gimpel**[3]
[1]Rutgers University, NJ, USA
[2]University of Chicago, IL, USA
[3]Toyota Technological Institute at Chicago, IL, USA
{mchen,kgimpel}@ttic.edu,zeweichu@gmail.com,stratos@cs.rutgers.edu

	MNLI	RTE	PPDB	Break	avg.
BERT	74.4	71.8	68.6	80.2	73.3
BERT & WIKINLI	**75.6**	**74.4**	**71.2**	85.7	**76.7**
− one cat. layer	74.6	73.3	70.9	**87.0**	76.5
− two cat. layers	75.4	72.9	**71.2**	82.5	75.5
+ page titles	74.2	73.6	70.6	80.7	74.8

Table 1: Comparing pruning levels for hierarchies available in Wikipedia. The highest numbers in each column are boldfaced.

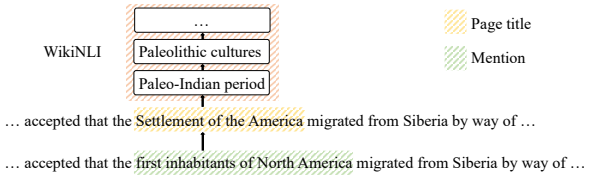

Figure 1: An example of WIKISENTNLI and higher-level categories that are used to construct WIKINLI.

A Hyperparameter and Model Size

As our models are based on BERT Large and RoBERTa, they share similar model sizes: 334 million. We mostly follow the hyperparameters recommended in the original papers (Devlin et al., 2019; Liu et al., 2019). For RoBERTa, we find that finetuning without using learning rate annealing will lead to training divergence, and thus we follow the recommendation in Liu et al. (2019) to use 10% of the training steps for learning rate annealing. For other hyperparameters, we only make changes to batch sizes when the finetuning is limited by the computational resources. When doing so, we try to use the maximum batch sizes that can fit into the GPU memory, without performing any tunning.

B Runtime and Computing Infrastructures

We always train our models on single GPU, including NVIDIA TITAN X or NVIDIA 2080 Ti. On average, pretraining on 100k WIKINLI, WordNet, or Wikidata takes 3.5 hours, and finetuning them takes approximately 1.0 hours.

C Wikipedia Pages, Mentions, and Layer Pruning

The variants of WIKINLI we considered so far have used categories as the lowest level of hierarchies. We are interested in whether adding Wikipedia page titles would bring in additional knowledge for inference tasks.

We experiment with including Wikipedia page titles that belong to Wikipedia categories to WIKINLI. We treat these page titles as the leaf nodes of the WIKINLI dataset. Their parents are the categories that the pages belong to.

Although Wikipedia page titles are additional source of information, they are more specific compared to Wikipedia categories. A majority of Wikipedia page titles are person names, locations, or historical events. They are not general summaries of concepts. To explore the effect of more general concepts, we try pruning leaf nodes from the WIKINLI category hierarchies. As higher-level nodes are more general and abstract concepts compared to lower-level nodes, we hypothesize that pruning leaf nodes would make the model learn higher-level concepts. We experiment with pruning one layer and two layers of leaf nodes in WIKINLI category hierarchies.

Table 1 compares the results of adding page titles and pruning different numbers of layers. Adding page titles mostly gives relatively small improvements to the model performance on downstream tasks, which shows that the page title is not a useful addition to WIKINLI. Pruning layers also slightly hurts the model performance. One exception is Break, which shows that solving it requires knowledge of higher-level concepts.

*Equal contribution. Listed in alphabetical order.

Sentence 1	Sentence 2	Rel.
He then moved to **Scottish society** as an actuary for Standard Life Assurance Company. However, he transferred back to London with the company.	He then moved to **Edinburgh** as an actuary for Standard Life Assurance Company. However, he transferred back to London with the company.	P
Dobroselo () is a village in **Croatia** . It is connected by the D218 highway. According to the 2011 census, Dobroselo had 117 inhabitants.	Dobroselo () is a village in **Southern European countries** . It is connected by the D218 highway. Accordingto the 2011 census, Dobroselo had 117 inhabitants.	C
His oldest brother Charuhasan, like Kamal, is a National Film Award-winning actor who appeared in the **ladino-language** film "Tabarana Kathe".	His oldest brother Charuhasan, like Kamal, is a National Film Award-winning actor who appeared in the **Kannada** film "Tabarana Kathe".	N

Table 2: Examples from WIKISENTNLI. C = child; P = parent; N = neutral. Boldfaced are page titles, WIKINLI categories, or their mentions in the context.

	MNLI	RTE	PPDB	Break	avg.
BERT	74.4	71.8	68.6	80.2	73.3
WIKINLI	**76.4**	**74.4**	**71.2**	**85.7**	**76.7**
+ page & mention	72.2	69.0	70.8	58.5	67.6
WIKISENTNLI	67.0	62.8	69.1	56.9	64.0
WIKISENTNLI cat.	71.8	67.1	70.6	84.0	73.4

Table 3: Comparison using WIKISENTNLI. The highest numbers in each column are boldfaced.

	MNLI	RTE	PPDB	Break	avg.
WikiNLI	**75.6**	**74.4**	**71.2**	**85.7**	**76.7**
length 200	75.3	71.5	70.5	81.4	74.7
w/o keyword filtering	74.8	74.0	71.1	85.2	76.3

Table 4: Effect of length and keyword filtering.

1968 in Hungary	Years of the 20th century in Hungary
The NHL Network (1975–79) affiliates	American television network affiliates
1616 establishments in the Ottoman Empire	1616 establishments by country
Russian mezzo-sopranos	Russian singers by voice type
Places of worship in South Dublin (county)	Buildings and structures in South Dublin (county)
National Register of Historic Places in Rockland County, New York	National Register of Historic Places in New York (state) by county
Military facilities on the National Register of Historic Places in Michigan	Buildings and structures on the National Register of Historic Places in Michigan
Populated places established in 2001	2001 establishments

Table 5: Examples of filtered pairs based on our length and key words filtering criteria.

D WIKISENTNLI

To investigate the effect of sentential context, we construct another dataset, which we call WIKISENTNLI, that is made up of full sentences. The general idea is to create sentence pairs that only differ by several words by using the hyperlinks in the Wikipedia sentences. More specifically, for a sentence with a hyperlink (if there are multiple hyperlinks, we will consider them as different instances), we form new sentences by replacing the text mention (marked by the hyperlink) with the page title as well as the categories describing that page. We consider these two sentences forming candidate child-parent relationship pairs. An example is shown in Figure 1. As some page titles or category names do not fit into the context of the sentence, we score them by BERT-large, averaging over the loss spanning that page title or category name. We pick the candidate with the lowest loss. To generate neutral pairs, we randomly sample 20 categories for a particular page mention in the text and pick the candidate with the lowest loss by BERT-large. WIKISENTNLI is also balanced among three relations (child, parent and neutral), and we experiment with 100k training instances and 5k development instances. Table 2 are some examples from WIKISENTNLI.

Table 3 shows the results. In comparing WIKINLI to WIKISENTNLI, we observe that adding extra context to WIKINLI does not help on the downstream tasks. It is worth noting that the differences between WIKINLI and WIKISENTNLI are more than sentential context. The categories we considered in WIKISENTNLI are always immediately after Wikipedia pages, limiting the exposure of higher-level categories.

To look into the importance of those categories, we construct another version of WIKISENTNLI by treating the mentions and page title layer as the same level ("WIKISENTNLI cat."). This effectively gives models pretrained on this version of WIKISENTNLI access to higher-level categories. Practically, when creating child sentences, we randomly choose between keeping the original sentences or replacing the text mention with its linked

WIKINLI	Chinese mWIKINLI
albums	中国(China)
songs	中华人民共和国(P. R. C.)
players	行政区划(administrative division)
male	人(man)
people	政治(politics)
American	人物(people)
British	各国(countries)
writers	组织(organization)
(band)	各省(provinces)
female	建筑物(building)
templates	美国(American)
music	历史(history)
articles	作品(work)
women	文化(culture)
films	属(category)
French	校友(alumnus)
artists	官员(offial)
German	地理(geography)
rock	公司(company)
musicians	城市(city)
culture	单位(unit)

Table 6: Top 20 most frequent words in WIKINLI, and mWIKINLI in Chinese. Each Chinese word is followed by a translation in parenthesis.

page title. When creating parent sentences, we replace the text mention with the parent categories of the linked page. Then, we perform the same steps as described in the previous paragraph. Pretraining on WIKISENTNLI cat. gives a sizable improvement compared to pretraining on WIKISENTNLI.

Additionally, we try to add mentions to WIKINLI, which seems to impair the model performance greatly. This also validates our claim that specific knowledge tends to be noisy and less likely to be helpful for downstream tasks. More interestingly, these variants seem to affect Break the most, which is in line with our previous finding that Break favors higher-level knowledge. While most of our findings with sentential context are negative, the WIKISENTNLI cat. variant shows promising improvements over BERT in some of the downstream tasks, demonstrating that a more appropriate way of incorporating higher-level categories can be essential to benefit from WIKISENTNLI in practice.

E Effect of Filtering

In Table 6, we list the top 20 most frequent words in WIKINLI, and mWIKINLI in Chinese. In Table 7, we list the top 20 most frequent words in WIKINLI, Wikidata, and WordNet.

We report results for the models trained on non-filtered WIKINLI datasets, showing that our filtering criteria give the best results. Examples of filtered pairs are listed in Table 5.

WIKINLI	Wikidata	WordNet
albums	protein	genus
songs	gene	dicot
players	putative	family
male	protein-coding	unit
people	conserved	fish
American	hypothetical	tree
British	languages	bird
writers	disease	person
(band)	RNA	fern
female	language	plant
templates	function	mammal
music	process	monetary
articles	group	animal
women	unknown	process
films	syndrome	arthropod
French	activity	order
artists	cell	acid
German	binding	disease
rock	food	vein
musicians	mineral	system
culture	transport	herb

Table 7: Top 20 most frequent words in WIKINLI, Wikidata, and WordNet.

F Full list of most frequent words

We show the complete set of top 20 most frequent words in Table 6 and Table 7.

G Evaluating on IMPPRESS

IMPPRES (Jeretic et al., 2020) evaluates the ability of models in performing pragmatic inferences. When evaluating on IMPPRES, we use the models finetuned on the 3k MNLI training instances. We report results in Table 8 and Table 9.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,

	conn.		grad. adj.		grad. verb		mod.		num._10_100		num._2_3		quant.			
	log.	prag.	log.	prag.	log.	prag.	log.	prag.	log.	prag.	log.	prag.	log.	prag.	log. avg	prag. avg
BERT	60.0	49.8	66.1	50.4	65.8	51.7	63.9	48.6	55.1	62.9	55.8	52.6	65.0	59.7	61.7	53.7
+ WordNet	61.8	49.3	68.1	49.9	66.0	49.8	63.7	47.3	52.2	63.1	54.3	56.0	64.7	56.6	61.5	53.1
+ Wikidata	61.3	51.8	65.3	51.3	66.6	50.0	65.2	48.8	54.2	64.7	59.3	54.7	65.9	54.3	62.5	53.7
+ WIKINLI	63.6	53.4	64.9	50.3	66.6	50.1	63.3	51.1	50.3	65.9	48.8	64.4	64.5	59.8	60.3	56.4
RoBERTa	58.7	46.3	67.5	49.8	66.6	51.8	63.2	51.8	55.3	71.2	56.2	71.8	61.2	64.8	61.2	58.2
+ WordNet	61.3	59.1	66.9	51.9	65.6	51.4	59.5	55.3	47.8	67.2	51.4	67.3	59.1	82.9	58.8	62.2
+ Wikidata	60.2	52.3	67.0	50.6	67.3	50.9	57.9	59.7	47.8	68.5	49.8	67.9	60.8	78.6	58.7	61.2
+ WIKINLI	59.0	45.3	66.5	51.8	64.8	50.9	56.7	62.6	53.1	65.9	49.9	65.3	61.9	76.5	58.8	59.8

Table 8: Implicature results for IMPPRES.

	ANP	COS	CU	PDE	QP	BP	CE	OP	PDU	avg
BERT	52.6	31.9	33.9	61.0	61.7	49.8	56.9	53.4	41.9	49.3
+ WordNet	45.8	28.5	18.8	56.5	57.2	49.5	55.3	58.2	30.6	44.5
+ Wikidata	29.3	25.8	16.3	56.8	58.5	27.6	56.8	57.9	28.9	39.8
+ WIKINLI	41.7	26.8	27.8	55.7	62.6	44.8	55.1	62.0	29.6	45.1
RoBERTa	48.4	33.4	33.1	48.6	50.6	46.2	49.3	46.9	43.3	44.4
+ WordNet	48.9	31.4	39.5	49.9	50.6	47.0	48.3	44.4	49.0	45.5
+ Wikidata	45.5	33.5	45.9	46.1	47.1	43.3	49.1	44.3	40.9	44.0
+ WIKINLI	40.9	30.0	35.5	51.4	50.2	36.5	51.3	45.7	39.3	42.3

Table 9: Presupposition results for IMPPRES. ANP=all n presupposition. COS=change of state. CN=cleft uniqueness. PDE=possessed definites existence. QP=question presupposition. BP=both presupposition. CE=cleft existence. OP=only presupposition. PDU=possessed definites uniqueness.

Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.