Inducing Taxonomic Knowledge from Pretrained Transformers

Catherine Chen*, **Kevin Lin***, **Dan Klein**
University of California, Berkeley
{cathychen,k-lin,klein}@berkeley.edu

Abstract

We present a method for inducing taxonomic trees from pretrained transformers. Given a set of input terms, we assign a score for the likelihood that each pair of terms forms a parent-child relation. To produce a tree from pairwise parent-child edge scores, we treat this as a graph optimization problem and output the maximum spanning tree. We train the model by finetuning it on parent-child relations from subtrees of WordNet and test on non-overlapping subtrees. In addition, we incorporate semi-structured definitions from the web to further improve performance. On the task of inducing subtrees of WordNet, the model achieves 66.0 ancestor F_1, a 10.4 point absolute increase over the previous best published result on this task.

1 Introduction

Taxonomic knowledge is useful for a variety of NLP tasks, including question answering (Miller, 1998) and information retrieval (Yang and Wu, 2012), and as a resource for understanding and building systematicity into neural models (Geiger et al., 2020; Talmor et al., 2020). These systems generally retrieve taxonomic information from lexical databases such as WordNet (Miller, 1998), which contain manually created taxonomies that are incomplete and expensive to maintain (Hovy et al., 2009).

Traditionally, models for automatically creating taxonomies have relied on statistics of large web-scale corpora. These models generally apply lexico-syntactic patterns (Hearst, 1992) to corpora, and derive taxonomic trees based on these statistics (e.g., Snow et al., 2005; Kozareva and Hovy, 2010; Bansal et al., 2014; Shang et al., 2020).

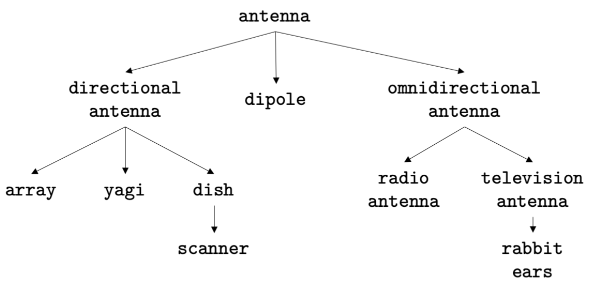

Figure 1: *An example subtree from the WordNet hierarchy.*

Recent work has shown that language models pretrained on large web-scale corpora contain relational knowledge about the world (Bouraoui et al., 2019; Bosselut et al., 2019). In this work, we propose a two-step procedure for taxonomy induction. We finetune pretrained transformers as parenthood predictors and then induce a taxonomy that maximizes parenthood scores.

We focus on the setting proposed by Bansal et al. (2014), where the task is to organize a set of input terms into a taxonomic tree. We compare to previous work, in which models were allowed access to web-based corpora.

Even without access to external data after pretraining, the model achieves a 8.5% absolute improvement over previous methods which were given access to web-based data. Our model obtains an additional 1.9% absolute improvement on the mean F_1 of reconstructed WordNet sub-hierarchies when allowed access to external web-based information.

2 Taxonomy Induction

We define taxonomy induction as the task of creating a tree-structured hierarchy $T = (V, E)$, where V is a set of terms and E is a set of edges representing hypernym relationships. In this task, the model must induce the tree T given a set of terms

* indicates equal contribution

V, where each term can be a single word or a short phrase. Our model performs taxonomy induction in two steps: parenthood prediction followed by tree reconciliation.

2.1 Parenthood Prediction

For a set of terms $V = \{v_1, ..., v_n\}$, we use a pretrained model to predict $\mathbb{P}[parent(v_i, v_j)]$ for all pairs $(v_i, v_j) \in V$. We provide each pair of terms as input into a contextual language model using standard encoding tokens (Devlin et al., 2018). We fine-tune the pretrained model to score each pair of terms using a sentence-level classification task on the sequence "[CLS] hypothesis", where the hypothesis is based on a template expressing hypernymy, e.g., "A dog is a mammal."

2.2 Tree Reconciliation

We then infer a tree from the model's pairwise parenthood predictions. We use the Chu-Liu-Edmonds algorithm for MST, which finds the highest scoring tree in $O(n^2)$ time (Chu, 1965).

3 Experiments

3.1 Experimental Settings

We test the model in two settings: closed-book and open-book. In the *closed-book* setting, the system performs inference using only the set of terms V. In the *open-book* setting, the system additionally uses contexts retrieved from the web. (See Section 3.1.1 for details.) Previous work has primarily tested models in the open-book setting, where models have access to web-based corpora. We use pretrained models from the Huggingface library (Wolf et al., 2019).

3.1.1 Web Definitions

In the open-book setting, we allow the model to access contexts from the web. [1] After retrieving these contexts, each term $v \in V$ is associated with a list of definitions $d_v^1, ..., d_v^n$. We reorder the definitions by their relevance to the current tree. We define relevance to the current tree as the cosine similarity between the average GloVe embedding of the words in the defintions with stopwords removed, to the average GloVe embedding of all terms in the subtree (Pennington et al., 2014). We then finetune the pretrained model on pairs of terms (v_i, v_j) using the sequence "[CLS] v_i $d_{v_i}^1, ..., d_{v_j}^n$. [SEP] v_j $d_{v_j}^1, ..., d_{v_j}^n$."

3.2 Datasets

3.2.1 English Dataset

We test our model using the dataset of medium-sized WordNet trees created by Bansal et al. (2014) as the main setup. This dataset consists of bottomed-out full subtrees of height 3 (4 nodes from the root to leaf) that contain between 10 and 50 terms. This comprises 761 trees, with 533/114/114 train/dev/test trees respectively.

3.2.2 Finnish Dataset

WordNet was originally constructed in English, and it has been extended to many different languages such as Finnish, Italian, and Greek (e.g., Magnini et al., 1994; Lindén and Niemi, 2014; Stamou et al., 2004). These have been linked to English WordNet in the Open Multilingual WordNet project (Bond and Paik, 2012). These alignments have been created using a mix of automatic and manual methods (e.g., Magnini et al., 1994; Lindén and Niemi, 2014). The coverage between different languages varies widely, covering between 14% and 100% of the terms in the trees in our English dataset. The Finnish WordNet (Lindén and Niemi, 2014) contains alignments to all of the terms, so we use Finnish WordNet as a comparison for the results on our English dataset.

3.3 Evaluation Metrics

As with previous work (Bansal et al., 2014; Mao et al., 2018), we report the ancestor F_1 score $\frac{2PR}{P+R}$, where

$$P = \frac{|is\_a_{predicted} \cap is\_a_{gold}|}{|is\_a_{predicted}|}$$

$$R = \frac{|is\_a_{predicted} \cap is\_a_{gold}|}{|is\_a_{gold}|}$$

We report these scores averaged over the subtrees in the test set.

4 Results

Our results for the English wordnet subtree reconstruction task are shown in Table 1, which contains the ancestor precision, recall, and F_1 scores on the test set. The model, which uses RoBERTa-large in the parenthood prediction stage, outperforms existing state-of-the-art models on subtree reconstruction. It does so even without using any external

[1] We scrape definitions from wiktionary.com, merriam-webster.com, and wikipedia.org. For Wikipedia we use the first paragraph of the page associated with the term.

web data, and it achieves additional improvements when given the web definitions described in Section 3.1.1.

We show comparisons to different model sizes and to BERT in Section 4.2.

	P	**R**	**F1**
Bansal et al. (2014)	48.0	55.2	51.4
Mao et al. (2018)	52.9	58.6	55.6
This work (closed-book)	68.1	62.4	64.1
This work (open-book)	**68.5**	**65.7**	**66.0**

Table 1: *English Results, Comparison to Previous Work.* Our approach outperforms previous approaches on reconstructing WordNet subtrees, even when the model is tested in the closed-book setting in which it cannot access external web data. In this table, the reported results for our work are averaged over three random restarts.

4.1 Web Definitions

Models perform better in the open-book setting, which involves retrieving definitions from the web.

Many of the terms in our dataset are polysemous, and the naively retrieved web definitions often include definitions for the incorrect sense of a term. For example, the term ‘dish’ appears in the sample subtree we show in Figure 1. The retrieval system obtains multiple definitions for ‘dish’, including “(telecommunications) A type of antenna with a similar shape to a plate or bowl.”, “(metonymically) A specific type of prepared food.”, and “(mining) A trough in which ore is measured.”

Re-ranking definitions based on relevance to the current subtree addresses this problem.

As we further show in Table 2, providing the models with definitions obtained directly from WordNet glosses further improves subtree reconstruction, showing that there is still room for improvement in definition retrieval.

4.2 Comparison of Pretrained Models

For both the open-book and closed-book settings, RoBERTa-large attains the highest F_1 score out of the different pretrianed models. As we show in Table 3, subtree reconstruction improves with both increased model size and improved pretraining.

4.3 Finnish

We tested our model on constructing subtrees from the Finnish alignment to the Bansal et al. (2014) dataset. We compared subtree reconstruction performance, using both multilingual BERT and FinBERT in the parenthood prediction stage of our model.

As we show in Table 4, the model's performance on reconstructing Finnish trees is noticeably worse than its performance on English trees.

	P	**R**	**F1**
BERT-base	61.3	55.1	56.7
+ web definitions	60.3	63.1	60.6
+ WordNet definitions	79.6	79.1	78.6

Table 2: *English Results, Definition Comparison.* Adding web definitions improves performance over only using input terms. Models achieve additional improvements in subtree reconstruction when given WordNet definitions, showing room for improvement in retrieving web definitions.

	P	**R**	**F1**
BERT-base	61.3	55.1	56.7
BERT-large	66.2	62.1	62.8
RoBERTa-large	68.5	65.7	66.0

Table 3: *English Results, Comparison of Pretrained Models.* Larger models perform better and RoBERTa outperforms BERT.

Language	Model	**P**	**R**	**F1**
English	English BERT	61.3	55.1	56.7
	mBERT	58.1	48.1	51.8
Finnish	FinBERT	41.0	28.7	32.7
	mBERT	31.6	21.0	24.5

Table 4: *Finnish Results, Comparison to English Subtrees.*

This is likely due to multiple factors. First, English pretrained language models generally perform better than models in other languages because of the additional resources devoted to the development of English models (See e.g., Bender, 2011; J., 2016; Joshi et al., 2020). Second, Open Multilingual Wordnet aligns wordnets to English WordNet, but the subtrees contained in English WordNet might not be directly applicable to terms in other languages. These results highlight the importance of evaluating on non-English languages and the difference in available lexical resources.

5 Related Work

Taxonomy Induction Taxonomy induction has been studied extensively, with both pattern-based and distributional approaches. Many systems for taxonomy induction have used pattern-based features such as Hearst patterns to infer hypernym releations from large corpora (e.g. Hearst, 1992; Snow et al., 2005; Kozareva and Hovy, 2010). For example, Snow et al. (2005) propose a system that extracts pattern-based features from a corpus to predict hypernymy relations between terms. Kozareva and Hovy (2010) propose a system that similarly uses pattern-based features to predict hypernymy relations, in addition to harvesting relevant terms and using a graph-based longest-path approach to induce a legal taxonomic tree.

Later work suggested that, for hypernymy detection tasks, pattern-based approaches outperform those based on distributional models (Roller et al., 2018). Subsequent work pointed out sparsity in pattern-based features derived from corpora, and showed that combining distributional and pattern-based approaches can improve hypernymy detection by addressing this problem (Yu et al., 2020).

In this work we consider the task of organizing a set of terms into a medium-sized taxonomic tree. Bansal et al. (2014) treat this as a structured learning problem and use belief propagation to incorporate siblinghood information. Mao et al. (2018) propose a reinforcement learning based approach that combines the stages of hypernymy detection and hypernym organization. In addition to the task of constructing medium-sized WordNet subtrees, they show that their approach can leverage global structure to construct much larger taxonomies from the TExEval-2 benchmark dataset, which contain hundreds of terms (Bordea et al., 2016). Likewise, Shang et al. (2020) apply graph neural networks and show that they improve performance in constructing large taxonomies in the TExEval-2 dataset.

Explicit Knowledge from Pretrained Language Models Another relevant line of work involves extracting structured declarative knowledge from pretrained language models. For instance, Bouraoui et al. (2019) showed that a wide range of relations can be extracted from pretrained language models such as BERT. Our work differs in that we consider tree structures and incorporate web definitions. Bosselut et al. (2019) use transformers to generate explicit open-text descriptions of commonsense knowledge. Other work has focused on extracting knowledge of relations between entities (Petroni et al., 2019; Jiang et al., 2020).

6 Discussion

Our experiments show that pretrained transformers can be used to induce taxonomic trees over sets of terms. Importantly, the knowledge encoded in these pretrained language models can be used to induce subtrees without accessing additional web-based information. These models produce subtrees with higher mean F_1 than those produced in previous work, even though previous approaches have used information from web queries.

When given access to additional semi-structured contexts such as web definitions, pretrained language models can produce improved taxonomic trees. The gain from accessing web definitions shows that incorporating both implicit knowledge of input terms and explicit textual descriptions of knowledge are promising ways to extract relational knowledge from pretrained models.

Our experiments on the aligned Finnish WordNet dataset emphasize that more work is needed in investigating the differences between taxonomic relations in different languages and in improving pretrained language models in non-English languages.

References

Mohit Bansal, David Burkett, Gerard De Melo, and Dan Klein. 2014. Structured learning for taxonomy induction with belief propagation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1041–1051.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, A. Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2019. Inducing relational knowledge from bert. *arXiv preprint arXiv:1911.12753*.

Yoeng-Jin Chu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

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

A. Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of BlackBoxNLP 2020*.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Coling 1992 volume 2: The 15th international conference on computational linguistics*.

E. Hovy, Zornitsa Kozareva, and E. Riloff. 2009. Toward completeness in concept extraction and classification. In *EMNLP*.

Sabrina J. 2016. Language diversity in acl 2004 - 2016.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1110–1118.

Krister Lindén and Jyrki Niemi. 2014. Is it possible to create a very large wordnet in 100 days? an evaluation. *Language Resources and Evaluation*, 48:191–201.

B. Magnini, C. Strapparava, F. Ciravegna, and E. Pianta. 1994. A project for the construction of an italian lexical knowledge base in the framework of wordnet.

Yuning Mao, Xiang Ren, Jiaming Shen, Xiaotao Gu, and Jiawei Han. 2018. End-to-end reinforcement learning for automatic taxonomy induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2462–2472, Melbourne, Australia. Association for Computational Linguistics.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Stephen Roller, Douwe Kiela, and Maximilian Nickel. 2018. Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–363, Melbourne, Australia. Association for Computational Linguistics.

Chao Shang, Sarthak Dash, Md Faisal Mahbub Chowdhury, Nandana Mihindukulasooriya, and Alfio Gliozzo. 2020. Taxonomy construction of unseen domains via graph-based cross-domain knowledge transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2198–2208.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Advances in neural information processing systems*, pages 1297–1304.

Sofia Stamou, Goran Nenadic, and Dimitris Christodoulakis. 2004. Exploring Balkanet shared ontology for multilingual conceptual indexing. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, page 781–784, Lisbon.

Alon Talmor, Oyvind Tafjord, P. Clark, Y. Goldberg, and Jonathan Berant. 2020. Teaching pre-trained models to systematically reason over implicit knowledge. *ArXiv*, abs/2006.06609.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

CheYu Yang and Shih-Jung Wu. 2012. Semantic web information retrieval based on the wordnet. *International Journal of Digital Content Technology and Its Applications*, 6:294–302.

Changlong Yu, Jialong Han, Peifeng Wang, Yangqiu Song, Hongming Zhang, Wilfred Ng, and Shuming Shi. 2020. When hearst is not enough: Improving hypernymy detection from corpus with distributional models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*.