Hypernymy Detection for Low-Resource Languages via Meta Learning

Changlong Yu[1*] **Jialong Han**[2*] **Haisong Zhang**[3] **Wilfred Ng**[1]
[1]HKUST, Hong Kong, China [2]Amazon, USA [3]Tencent AI Lab, China
{cyuaq, wilfred}@cse.ust.hk
jialonghan@gmail.com, hansonzhang@tencent.com

Abstract

Hypernymy detection, *a.k.a.* lexical entailment, is a fundamental sub-task of many natural language understanding tasks. Previous explorations mostly focus on monolingual hypernymy detection on high-resource languages, *e.g.,* English, but few investigate the low-resource scenarios. This paper addresses the problem of low-resource hypernymy detection by combining high-resource languages. We extensively compare three joint training paradigms and for the first time propose applying meta learning to relieve the low-resource issue. Experiments demonstrate the superiority of our method among the three settings, which substantially improves the performance of extremely low-resource languages by preventing over-fitting on small datasets.

1 Introduction

Hypernymy is a fundamental asymmetric lexico-semantic relation. It expresses `is-a` relationship between concepts and is widely used to build taxonomies (Miller, 1995) or large-scale knowledge bases (Wu et al., 2012; Seitner et al., 2016). Lexico-semantic patterns (*e.g.,* X such as Y) are generally employed to harvest benchmark datasets or resources from large English corpus due to their high precision (Hearst, 1992). However, Hearst-like patterns of English can not be easily transferred to other languages such as Chinese. Creating high-quality hypernymy benchmarks for other languages requires much more human-annotation efforts and hypernymy detection in those languages remains low-resource tasks (Vulić et al., 2019). In this paper, we focus on the question: how could we make full use of hypernymy pairs of high-resource languages such as English for other low-resource languages, *e.g.,* Japanese and Thai?

To investigate this question, we firstly assume a strong feasibility of semantic relation transfer across languages, which is in line with existing findings on human cognition. Youn et al. (2016) uncovered the universal conceptual structure of human lexical semantics among cross-lingual dictionaries and revealed the language-independent attribute for semantic similarity of concepts. Wang et al. (2019) studied cross-lingual training by simply merging high-resource language pairs and low-resource language ones, which is prone to over-fitting to low-resource ones. Based on the above interesting findings and the datasets in Wang et al. (2019), we study three training paradigms of combining training data from multiple different languages, *i.e., cross-lingual training*, *multilingual training*, as well as *meta learning*.

To the best of our knowledge, meta learning algorithms have not been previously applied to hypernymy detection. We propose applying meta learning algorithms in low-resource hypernymy detection and perform extensive comparisons with multilingual training. Meta-learning algorithms aim at learning language-independent models and then fine-tuning on multiple languages with minimal training instances. In our experiments, we further explore the two following questions:

- Considering the language-agnostic lexical semantics, would multilingual training improve the performance by employing additional regularization?
- Regarding the effectiveness of meta learning in low-resource scenarios (Dou et al., 2019), can we leverage meta learning to help multilingual training?

The results for question 1 are surprising. Obvious improvement is observed from neither bilingual cross-training nor multilingual training. The perfor-

[*] Work done when C. Yu and J. Han were with Tencent AI Lab.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3651–3656
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

mance even drops on extremely low-resource languages as the models easily over-fit low-resource language datasets. Meta learning algorithms, on the other hand, significantly relieve these cases by learning good model initialization for all languages. In the end, meta learning achieves the best performance among three training paradigms, which answers the main questions of this work.

2 Training Settings

In this section, we first introduce the base supervised model for hypernymy detection, and then illustrate three joint training paradigms.

2.1 Base Model

As discussed in Section 1, pattern-based models are highly language-dependant and can not generalize to arbitrary languages. We resort to supervised distributional models as base models, which take the distributional representation of terms as input features to train hypernymy relation classifiers (Roller et al., 2014; Yu et al., 2015; Rei et al., 2018). Luckily, pre-trained distributional vectors (*e.g.,* fastText word embedding (Bojanowski et al., 2017)) are widely available for most languages.

Formally, given a pair of terms (x, y) in one language, we denote the corresponding word vectors by $\mathbf{x}$ and $\mathbf{y}$. The hypernymy detection models learn a classifier f_θ to make binary prediction, where the input features could be the concatenation, difference, or other complex combinations of $\mathbf{x}$ and $\mathbf{y}$. To keep the base model simple and effective, we directly concatenate the two vectors and train a two-layer MLP, *i.e.,* $f_\theta(\mathbf{x} \oplus \mathbf{y}) = \mathbf{MLP}(\mathbf{x} \oplus \mathbf{y})$. Note the performances of base model are comparable with the ones in Wang et al. (2019) without feature extractors and self training.

2.2 Joint Models

Cross-lingual Training. Following the setting of Wang et al. (2019), cross-lingual hypernymy detection aims to predict low-resource language pairs combining large training data from high-resource languages. Specially, in our case, English is the only high-resource language. Therefore, we train a joint model on the mixture of our large English dataset and the small dataset of another language such as Japanese. Due to the different representation spaces of languages, word translation techniques are required to transfer knowledge and align the feature space across languages. We adopt the technique of Conneau et al. (2017) to learn a mapping matrix $\mathbf{W_{l-en}}$ to project the word embedding space of language l to that of English. The input feature to the classifier f_θ for language l is then $(\mathbf{W_{l-en}x}, \mathbf{W_{l-en}y})$. The quality of translation matrix $\mathbf{W_{l-en}}$ highly affects the transfer performance and we carefully choose the best mapping according to the evaluation on bilingual word translation benchmarks[1]. Detailed results are omitted due to the limited space.

Algorithm 1: Meta Learning procedure

```
Initialize base model $f$ with parameter $\theta$
for $i$ in {1,2, ... nsteps} do
    Randomly draw $L$ tasks $\{T_1, T_2, ...T_L\}$
    for $l$ in {1,2 ... L} do
        Update $k$ steps $\theta_l^k$ with Equation 1
    end
    Update $\theta$ using Equation 2
end
Fine-tune $\theta$ on each low-resource language.
```

Multilingual Training. Instead of training on a pair of languages, multilingual training combines all available pairs in any language. Glavaš and Vulić (2018) has showed that semantic relation classification tasks benefit from the additional regularization resulted from multilingual training. We also investigate whether multilingual training for low-resource hypernymy detection could learn a model that has better generalization ability on all languages. Due to the language-independent structure of semantic relation, the interaction among datasets of all languages imports more external knowledge than cross-lingual training. However the characteristic of limited training instances for low-resource languages may make the model easily over-fit and hurt the generalization. In the following experiments, we would answer and analyze the question thoroughly.

Meta Learning. Inspired by low-resource machine translation in Gu et al. (2018) and general language representation in Dou et al. (2019), we propose applying meta learning algorithms to hypernymy detection. We firstly learn language-independent models based on multiple high-resource languages and then adapt to low-resource language pairs. Here we adopt the most representative model-agnostic meta-learning (MAML)

[1] https://github.com/facebookresearch/MUSE

algorithm (Finn et al., 2017). Formally, given the base model f_θ with parameters θ, we denote training on each language l as task T_l. For each task (language) T_l, we sample a batch of data as the support set $T_l(S)$ and another batch of data as the query set $T_l(Q)$. During the meta training stage, we randomly sample L tasks $\{T_1, T_2, ...T_L\}$, and then update the model parameters by k gradient steps for each task T_l:

$$\theta_l^k = \theta_l^{k-1} - \alpha \nabla_{\theta_l^{k-1}} \mathcal{L}_{T_l(S)}(f_{\theta_l^{k-1}}). \tag{1}$$

Here $\mathcal{L}$ is the loss function for task T_l and α is the learning rate. The overall objective function for meta learning is $\min_\theta \sum_l \mathcal{L}_{T_l(Q)}(f_{\theta_l^k})$. Hence the model parameters are updated by:

$$\theta = \theta - \beta \nabla_\theta \sum_{l=1}^{L} \mathcal{L}_{T_l(Q)}(f_{\theta_l^k}), \tag{2}$$

where β is the learning rate for meta learning. The overall meta learning procedure is formulated in Algorithm 1. After $nsteps$ of meta learning iterations, we use several small-batch data from each language to fine-tune the model parameter θ.

Compared with multilingual training in Section 2.2, meta learning algorithms have the same input but different learning procedures or parameter updating strategies. Instead of simply merging all the high-resource and low-resource datasets to learn a joint model, meta learning algorithms learn a good initialization for all languages that can be adapted to one specific language. An obvious advantage of universal initialization is that it avoids the case where the model may favor high-resource languages in multilingual training (Dou et al., 2019).

3 Experiments

3.1 Experimental Setup

We conduct experiments on the hypernymy detection datasets of several languages in Wang et al. (2019)[2]. The languages are French (FR), Chinese (ZH), Finnish (FI), Italian (IT), Thai (TH), Japanese (JA), and Greek (EL). True hypernymy pairs are extracted from Open Multilingual WordNet (Bond and Foster, 2013) while false pairs are a mixture of synonymy and other relation pairs. For the English dataset, it combines five commonly-used benchmarks and we refer to Wang et al. (2019) for the description of data construction. We further categorize the seven low-resource datasets as moderately low-resource ones *e.g.,* FR, ZH, FI, IT and extremely low-resource ones *e.g.,* TH, JA, EL according to relative dataset sizes. The statistics of all datasets are shown in Table 1.

	Lang.	#True	#False	#Vocab
High-Resource	EN	17,591	57,164	47,305
Moderately Low-Resource	FR	4,035	8,947	12,979
	ZH	2,962	6,382	7,372
	FI	7,157	9,433	16,082
	IT	3,034	6,081	11,572
Extremely Low-Resource	TH	1,156	1,977	2,715
	JA	1,448	3,203	7,301
	EL	2,612	1,454	4,303

Table 1: Statistics for all languages' hypernymy detection datasets. #True and #False are the number of data with true/false labels. #Vocab stands for the vocabulary size.

For all three low-resource joint training paradigms, we randomly split the non-English language datasets with 20% for training, 20% for development, and 60% for testing, following Wang et al. (2019). For English we also take out the 20% development set for model selection. Word embeddings for each language are from pre-trained fastText word vectors[3] whose dimensions are set to 300. We report averaged accuracy of 5-fold cross-validation for low-resource languages. For the three joint models, we uniformly run 5,000 steps and select the best model for each language based on its development set. The hidden layer size for the base models is set to 400. We use vanilla SGD to optimize the meta learner with batch size 32 and learning rate $\beta = 0.5$. We set the sampled task number L in each step to 8, update step k to 5, and inner learning rate α to 0.001. Our code is available at `https://github.com/ccclyu/metaHypernymy`.

3.2 Experimental Results

In Table 2, we demonstrate the main results of all training paradigms. Empirically we answer the two questions raised in the Section 1.

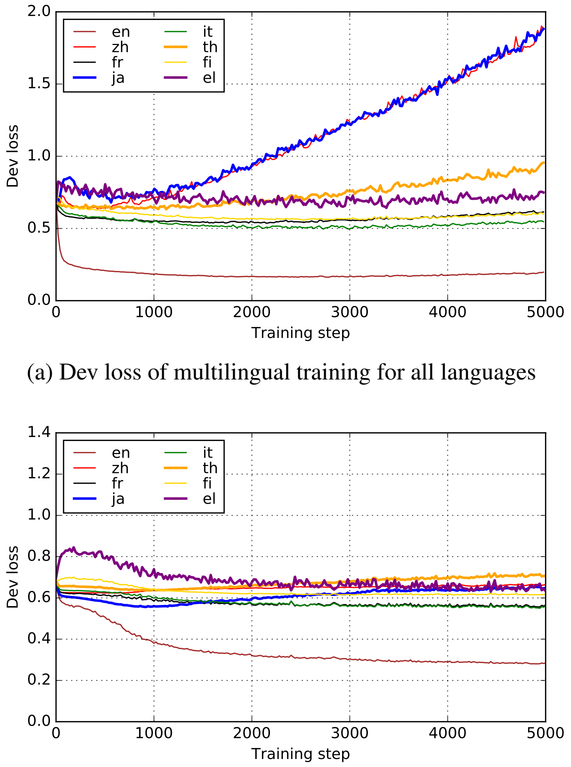
Do simple joint multilingual models work? In the first row, we report performances of the base

[2] The reason why we do not evaluate on Bordea et al. (2016); Vulić et al. (2019) is either no false hypernymy pair or unfit setting.

[3] `https://fasttext.cc/docs/en/crawl-vectors.html`

	FR	ZH	FI	IT	TH	JA	EL
Mono	.744	.697	.692	.744	.659	.740	.702
Cross	.748	.679	.711	.752	.693	.711	.684
Multi	.756	.690	.713	.760	.657	.711	.653
ZeroMeta	.765	.700	.713	.762	.702	.747	.712
Finetune	**.769**	**.713**	**.714**	**.764**	**.709**	**.756**	**.734**

Table 2: Experimental comparison of training paradigms for low-resource hypernymy detection.



(a) Dev loss of multilingual training for all languages

(b) Dev loss of meta learning for all languages

Figure 1: Comparison of training curve of two settings. Bold lines are extremely low-resource ones (TH, JA, EL).

monolingual model on all seven low-resource languages, denoted by "Mono". On top of it, Cross-lingual training (or bilingual, denoted by "Cross") obtains marginal improvements for moderately low-resource languages. However, the performance drops dramatically for two extremely low-resource languages, *i.e.,* JA from 0.740 to 0.711 and EL from 0.702 to 0.684. We note that data sparsity leads to the over-fitting issue and thus bad generalization. Similar observations could be drawn from multilingual training ("Multi" for short). In summary, for extremely low-resource datasets, effective and advanced joint training is needed.

Is meta learning better than multilingual training? As discussed in Section 2.2, simple multilingual training and meta learning have the same input. But our experiments indicate that even the model initialized by meta learning (not fine-tuned, denoted by "zeroMeta" in Table 2) achieves superior performances. For example, on Thai, the accuracy jumps from 0.657 to 0.702 without fine-tuning. After fine-tuning with several batches of data, meta learning (denoted by "Finetune") achieves the best performance for all low-resource languages. To fully understand the difference of the two training paradigms, we use the same batch size and run the two joint training models for 5,000 steps. Figure 1 shows the loss curve of the development set for each low-resource language as well as English. We have two major observations: **1)** Both the two joint training paradigms could well fit English, the high-resource dataset, but multilingual training converges quickly then over-fits severely on extremely low-resource datasets (indicated by bold lines in Figure 1a), which results in dropping performances. Instead, meta learning has a relatively stable trend on the descending loss. For EL (the purple bold line in Figure 1b), though the loss first increases, it finally decreases and reaches a lower level. **2)** The converging dev losses of meta learning reach to lower numbers and have lower variances among all languages. This demonstrates that meta learning aims at learning a language-independent model/initialization that is helpful for fine-tuning rather than over-fitting on some languages.

3.3 Discussion

Experiments are based on good word representations and bilingual lexicon induction methods. However, the quality of them would impact results considerably, which we briefly discuss below.

Transferability of Word Vector Space. One of the limitation of training paradigms in our work might be non-isomorphic embedding spaces, which are largely caused by the intrinsic property of dissimilar languages. The projection matrix $\mathbf{W}_{\mathbf{l-en}}$ is learned unsupervisedly based on strong assumption that the embedding spaces for two languages are isometric, *i.e.,* similar in terms of structures (Vulić et al., 2020). However when generalizing to more low-resource languages, it does not always hold. It would be necessary in practice to carefully quantify isomorphism between two word vector spaces and adopt the approaches that relax the isomorphic assumption (Patra et al., 2019).

Contextualized Word Representation (CWR). Replacing static word vectors with CWRs such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) has achieved dominant performances on almost every NLP task. Ethayarajh (2019) show that principal component embeddings of CWR in lower layers of BERT outperform GloVe and fastText on many static embedding benchmarks such as word similarity and analogy. However it remains unclear how to use CWR to fully help lexical semantic tasks. We are also interested in whether zero-shot multilingual CWR pre-training such as Multilingual BERT (Pires et al., 2019) would benefit this task. Another promising direction is to devise the lexical knowledge from large pre-training language models (Bosselut et al., 2019; Petroni et al., 2019). We left them for the future work.

4 Related Work

Cross-Lingual Hypernymy Detection. Wang et al. (2019) firstly studies hypernymy detection in multilingual joint settings, Other similar tasks intend to predict whether a pair of words from two different languages exhibit hypernymy relationship (Vyas and Carpuat, 2016; Upadhyay et al., 2018; Glavaš and Vulić, 2019) or to what extent the relationship (Vulić et al., 2019) is. In this work, we focus on the former task.

Meta Learning. Also known as *learn to learn*, it aims at developing models that could learn new tasks or adopt to new tasks with a few training examples. Recently it has attracted more attention due to the simple yet effective models such as MAML (Finn et al., 2017) and Reptile (Nichol et al., 2018).

There are emerging investigations of applying meta learning in NLP tasks such as machine translation (Gu et al., 2018), semantic parsing (Huang et al., 2018), personalized dialogue system (Madotto et al., 2019), relation classification (Obamuyide and Vlachos, 2019) and code-switched speech recognition (Winata et al., 2020). Our work is inspired by Dou et al. (2019) that compares multi-task learning and meta learning for general language representations.

5 Conclusion

Transferring lexical knowledge across languages are important especially for low-resource cases. In this paper, we investigate three joint training paradigms for detecting hypernymy in low-resource languages. We show that simple multilingual training is not helpful for all tasks and we significantly improve the performance using meta learning. Our study demonstrates the feasibility and effectiveness to combine high- and low-resource data to jointly train hypernymy detection models.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. We would also like to thank Guanlin Li, Zi-Yi Dou, Yong Wang and Shuo Wang for their insightful discussions and invaluable help.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of ACL*, pages 1352–1362, Sofia, Bulgaria.

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 (SemEval-2016)*, pages 1081–1091, San Diego, California.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of ACL*, pages 4762–4779.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. In *Proceedings of EMNLP-IJCNLP*, Hong Kong, China. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of EMNLP-IJCNLP*, pages 55–65.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML*, pages 1126–1135.

Goran Glavaš and Ivan Vulić. 2018. Discriminating between lexico-semantic relations with the specialization tensor model. In *Proceedings of NAACL*, pages 181–187.

Goran Glavaš and Ivan Vulić. 2019. Generalized tuning of distributional word vectors for monolingual and cross-lingual lexical entailment. In *Proceedings of ACL*, pages 4824–4830.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of EMNLP*, pages 3622–3631.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545.

Po-Sen Huang, Chenglong Wang, Rishabh Singh, Wentau Yih, and Xiaodong He. 2018. Natural language to structured query generation via meta-learning. In *Proceedings of NAACL*, pages 732–738.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of ACL*, pages 5454–5459.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Abiola Obamuyide and Andreas Vlachos. 2019. Model-agnostic meta-learning for relation classification with limited supervision. In *Proceedings of ACL*, pages 5873–5879.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of ACL*, pages 184–193.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP*, pages 2463–2473.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Marek Rei, Daniela Gerz, and Ivan Vulić. 2018. Scoring lexical entailment with a supervised directional similarity network. In *Proceedings of ACL*, pages 638–643.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING*, pages 1025–1036.

Julian Seitner, Christian Bizer, Kai Eckert, Stefano Faralli, Robert Meusel, Heiko Paulheim, and Simone Paolo Ponzetto. 2016. A large database of hypernymy relations extracted from the web. In *LREC*.

Shyam Upadhyay, Yogarshi Vyas, Marine Carpuat, and Dan Roth. 2018. Robust cross-lingual hypernymy detection using dependency context. In *Proceedings of NAACL*, pages 607–618.

Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2019. Multilingual and cross-lingual graded lexical entailment. In *Proceedings of ACL*, pages 4963–4974.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? *arXiv preprint arXiv:2004.04070*.

Yogarshi Vyas and Marine Carpuat. 2016. Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of NAACL*, pages 1187–1197.

Chengyu Wang, Yan Fan, Xiaofeng He, and Aoying Zhou. 2019. A family of fuzzy orthogonal projection models for monolingual and cross-lingual hypernymy prediction. In *The World Wide Web Conference*, pages 1965–1976. ACM.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, , Zihan Liu, Peng Xu, and Pascale Fung. 2020. Meta-transfer learning for code-switched speech recognition. *arXiv preprint arXiv:2004.14228*.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of SIGMOD*, pages 481–492. ACM.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.

Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.