Biomedical Concept Normalization by Leveraging Hypernyms

Cheng Yan[1,2], **Yuanzhe Zhang**[1,2], **Kang Liu**[1,2], **Jun Zhao**[1,2], **Yafei Shi**[3], **Shengping Liu**[3]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Unisound AI Technology Co., Ltd.
{cheng.yan, yzzhang, kliu, jzhao}@nlpr.ia.ac.cn,
{shiyafei, liushengping}@unisound.com

Abstract

Biomedical Concept Normalization (BCN) is widely used in biomedical text processing as a fundamental module. Owing to numerous surface variants of biomedical concepts, BCN still remains challenging and unsolved. In this paper, we exploit biomedical concept hypernyms to facilitate BCN. We propose Biomedical Concept Normalizer with Hypernyms (BCNH), a novel framework that adopts list-wise training to make use of both hypernyms and synonyms, and also employs norm constraint on the representation of hypernym-hyponym entity pairs. The experimental results show that BCNH outperform the previous state-of-the-art model on the NCBI dataset. Code will be available at https://github.com/yan-cheng/BCNH.

1 Introduction

Biomedical Concept Normalization (BCN) plays an important and prerequisite role in biomedical text processing. The goal of BCN is to link the entity mention in the context to its normalized CUI (Unique Concept Identifier) in the biomedical dictionaries such as UMLS (Bodenreider, 2004), SNOMED-CT (Spackman et al., 1997) and MedDRA (Brown et al., 1999). Figure 1 is an example of BCN from NCBI dataset (Doğan et al., 2014), the mention *B-cell non-Hodgkins lymphomas* should be linked to *D016393 Lymphoma, B-Cell* in the MEDIC (Davis et al., 2012) dictionary.

Recent works on BCN usually adopt encoders like CNN (Li et al., 2017), LSTM (Phan et al., 2019), ELMo (Peters et al., 2018; Schumacher et al., 2020) or BioBERT (Lee et al., 2020; Fakhraei et al., 2019; Ji et al., 2020) to embed both the mention and the concept's name entities, and then feed the representations to the following classifier or ranking network to determine the corresponding concept in the biomedical dictionary. However, biomedical dictionaries are generally sparse in nature: a concept is usually provided with only CUI, referred name (recommended concept name string), synonyms (acceptable name variants, synonyms), and related concepts (mainly hypernym concepts). Therefore, effectively using the limited information in the biomedical dictionary where the candidate entities came from is paramount for the BCN task.

For concept's synonym entities, recent BNE (Phan et al., 2019) and BIOSYN (Sung et al., 2020) tries to make full use of them by synonym marginalization to enhance biomedical entity representation and achieved consistent performance improvement. Unfortunately, previous works generally ignore concept hypernym hierarchy structure, which is exactly the initial motivation of biomedical dictionary: organization of thousands of concepts under a unified and multi-level hierarchical classification schema.

We believe that leveraging hypernym information in the biomedical dictionary can improve the BCN performance based on two intuitions. First, hard negative sampling (Fakhraei et al., 2019; Phan et al., 2019) is vital for the BCN model's discriminating ability and a hypernym is a hard negative example for its hyponym naturally. Second, injecting the hypernym hierarchy information during the training process is beneficial for encoders, since currently used encoders like BioBERT only encodes the context semantics in biomedical corpora instead of the biomedical concept structural information.

To this end, we propose Biomedical Concept Normalizer with Hypernyms (BCNH), a novel framework combining the list-wise cross entropy loss with norm constraint on hypernym-hyponym entity pairs. Concretely, we reformulate the candidate target list as a three-level relevance list to consider both synonyms and hypernyms, and apply

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3512–3517
November 7–11, 2021.

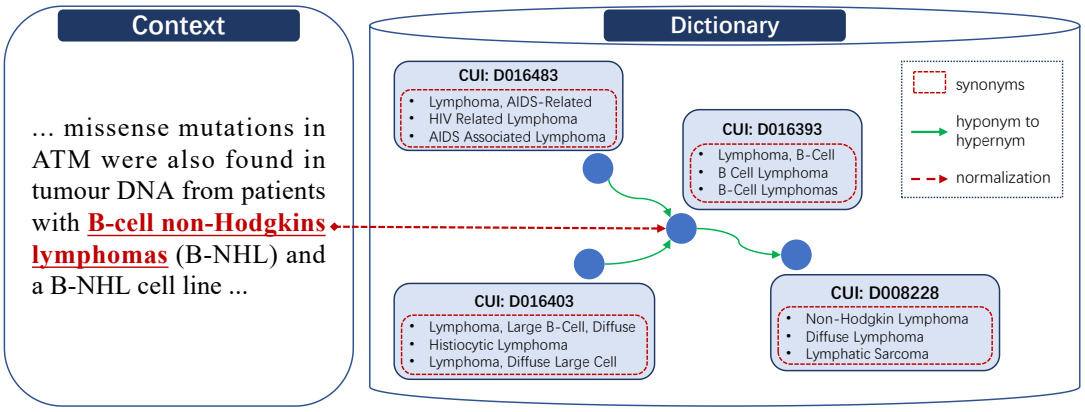

Figure 1: A biomedical concept normalization example.

the list-wise cross entropy loss. On the one hand, synonyms help to encode surface name variants, on the other hand, hypernyms help encode hierarchical structural information. We also apply the norm constraint on the embedding of hypernym-hyponym entity pairs to further preserve the principal hypernym relation. Specifically, for a hypernym-hyponym entity pair (e_{hyper}, e_{hypo}), we constraint that the norm of hypernym entity e_{hyper} is larger than that of e_{hyper} in a multi-task manner. We conduct experiments on the NCBI dataset and outperforms the previous state-of-the-art model.

To sum up, the contributions of this paper are as follows. First, for the first time, we reformulate the candidate target list as a three-level relevance list and apply the list-wise loss to attend all candidate entities. Second, we innovatively use norm constraint to model the hypernym-hyponym relation, preserving the hierarchy structure information inside the entity representation. The proposed BCNH outperforms the previous state-of-the-art model on the NCBI dataset, leading to an improvement of 0.73 % on top1 accuracy.

2 Methodology

The architecture of our framework is illustrated in Figure 2. Our model is composed of three parts: candidate generator to generate the candidate entities from the dictionary, list-wise ranker to train the encoder, hypernym normalizer to apply the hypernym-hyponym norm constraint.

2.1 Iterative Candidate Generator

We reuse the iterative candidate generator module from BIOSYN (Sung et al., 2020). Each mention m and entity e_i in dictionary $D = \{e_1, e_2, \cdots\}$ are represented first with sparse representations and dense representations. The sparse representations of m and e_i are denoted as $(v^s_m, v^s_{e_i})$ which is calculated based on the character-level n-grams statistics computed over all entities from D. The dense representations of m and e_i are denoted as $(v^d_m, v^d_{e_i})$, which are obtained from the pre-trained BioBERT.

The candidate generator then computes the similarity score between mention m and each entity e_i by combining the sparse similarity score $S_{sparse}(m, e_i)$ with dense similarity score $S_{dense}(m, e_i)$:

$$z_i = S_{dense}(m, e_i) + \lambda S_{sparse}(m, e_i) \quad (1)$$

$$S_{dense}(m, e_i) = f(v^d_m, v^d_{e_i}) \quad (2)$$

$$S_{sparse}(m, e_i) = f(v^s_m, v^s_{e_i}) \quad (3)$$

where function f is the inner product and λ is a trainable sparse score scalar weight. In the end, the top k_1 entities with the highest similarity scores are selected into candidate list $[e_1, e_2, \cdots, e_{k_1}]$, and their similarity score list is denoted as $z = [z_1, z_2, \cdots, z_{k_1}]$. The candidate list is pre-computed and iteratively updated at the beginning of every training step.

At inference time, the entity $e^* \in D$ with top similarity score is retrieved, and the CUI of entity e^* is returned as predicted CUI.

2.2 List-wise Ranker

For mention m and its top k_1 candidate list $[e_1, e_2, \cdots, e_{k_1}]$, we reformulate the targets of candidate list as a three-level relevance score list. The relevance score is defined as the degree of relevance between mention m and candidate entity e_i.

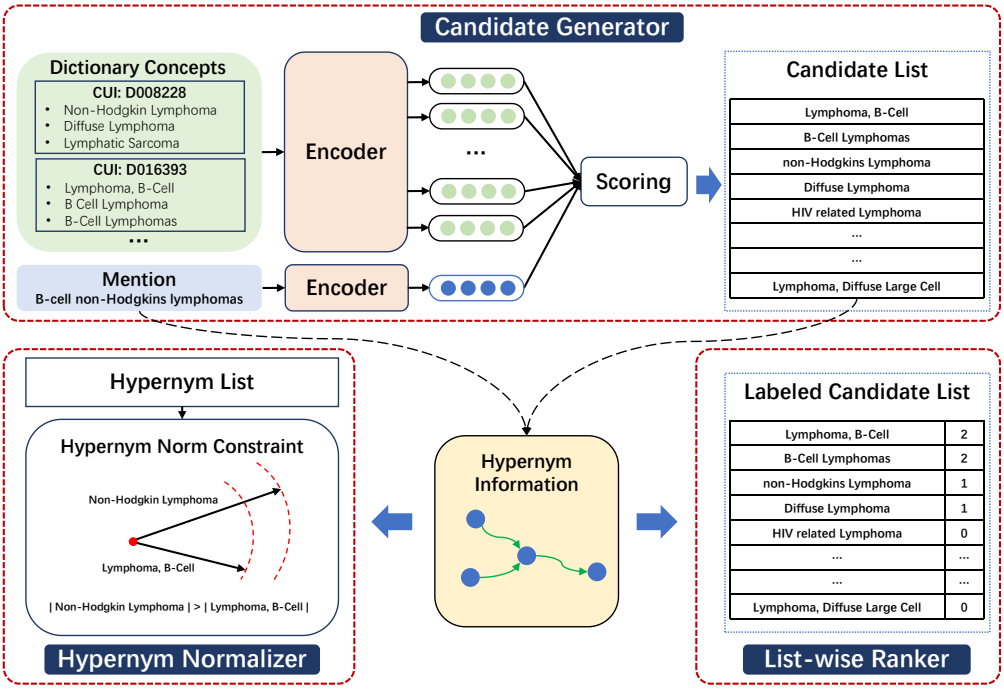

Figure 2: The architecture of BCNH.

Specifically, for a candidate entity e_i, the relevance score y_i is set to 2 if e_i is synonym of m, 1 if e_i is hypernym of m, 0 if neither synonym or hypernym. Therefore, we have a pseudo relevance score target $y = [y_1, y_2, \cdots, y_{k_1}]$ and the candidate similarity score list $z = [z_1, z_2, \cdots, z_{k_1}]$.

The list-wise cross entropy loss (LCE) (Cao et al., 2007) then is applied on the relevance score y and candidate similarity score z. The objective of learning candidate similarity is formalized as minimization of the total LCE losses on all examples:

$$ListLoss = \frac{1}{M}\sum_{j=1}^{M} LCE(y^j, z^j) \quad (4)$$

$$LCE(y^j, z^j) = -\sum_{i=1}^{k_1} P_{y_i} log(P_{z_i}) \quad (5)$$

$$P_{y_i} = \frac{e^{y_i}}{\sum_{i=1}^{k_1} e^{y_n}}, P_{z_i} = \frac{e^{z_i}}{\sum_{i=1}^{k_1} e^{z_n}} \quad (6)$$

where M is the number of mentions in the training dataset.

Leveraging hypernyms for the list-wise learning targets can be interpreted as a hard negative sampling technique (Kalantidis et al., 2020), which is crucial under the contrastive learning framework.

2.3 Hypernym Normalizer

Though we take hypernyms into account by the list-wise training, the hypernym hierarchy information inside the dictionary is still absent in concept entity representation. It has been proven in (Vulić and Mrkšić, 2018) that the asymmetric norm distance is an effective way to encode the hierarchical ordering between hypernym and hyponym entities.

During training, we prepare a k_2 length hypernym list $(e_{h_1}, e_{h_2}, \cdots)$ for mention m. We denote the norm distance between mention m and its all hyponyms as $NormLoss$:

$$NormLoss = \frac{1}{k_2}\sum_{i=1}^{k_2} \frac{|v_m^d| - |v_{h_i}^d|}{|v_m^d| + |v_{h_i}^d|} \quad (7)$$

By minimizing the $NormLoss$, we constraint that the norm of hypernym embedding vector $v_{h_i}^d$ is larger than the mention embedding vector v_m^d under the intuition that the norm constraint fine-tunes norm values in the Euclidean embedding space to reflect the hierarchical organization of biomedical concept entities.

In the end, the BCNH jointly optimizes cost:

$$Loss = ListLoss + NormLoss \quad (8)$$

3 Experiments

3.1 Experimental setup

Dataset We train and evaluate our model on the NCBI Disease corpus, a collection of 793 PubMed abstracts with disease mentions and their concepts corresponding to the MEDIC dictionary. In this work, we use the MEDIC of version February 1, 2021 that contains 13,103 CUIs, 74,215 synonyms, and 21,999 hypernyms.

Preprocessing We follow the same dataset preprocessing including lower-casing, punctuation removing, abbreviations expanding, composite mentions splitting in previous works (Leaman and Lu, 2016; Wright, 2019; Phan et al., 2019; Sung et al., 2020). We use the top k accuracy metric to evaluate the task.

Hyper-parameters We set all the parameters in the candidate generator exactly the same with BIOSYN for fair comparison. Our model only introduces a new hyper parameter $k_2 = 10$ in our experiments. When the hypernyms of mention m in the dictionary is more than k_2, we truncate it to k_2; and pad null entity if less than k_2. The Adam optimizer (Kingma and Ba, 2014) is used to minimize the final loss.

3.2 Results

The main results are shown in Table 1. Our proposed BCNH outperforms the previous state-of-the-art model BIOSYN (Sung et al., 2020) on Acc@1 and Acc@5 with an improvement of 0.73% and 1.18%, respectively. Our model also obtains a smaller confidence interval.

Models	Acc@1	Acc@5
Sieve-Based (D'Souza and Ng, 2015)	84.7	-
Taggerone (Leaman and Lu, 2016)	87.7	-
CNN Ranking (Li et al., 2017)	86.1	-
NormCo (Wright, 2019)	87.8	-
BNE (Phan et al., 2019)	87.7	-
BERT Ranking (Ji et al., 2020)	89.1	-
TripletNet (Mondal et al., 2019)	90.0	-
BIOSYN † (Sung et al., 2020)	89.88 ± 0.22	93.82 ± 0.26
BCNH (Ours)	**90.61** ± 0.17	**95.00** ± 0.14

† Since the original result is reported for a different version MEDIC dictionary, we use the author's provided code to evaluate the BIOSYN model with different seeds 10 times with 95% confidence. The results of other models are directly cited.

Table 1: Acc@1 and Acc@5 on NCBI dataset.

For comparison, we list the top 10 predictions of an example mention *B-cell non-Hodgkins lymphomas (D016393)* from the NCBI test dataset in Table 2. BIOSYN fails to rank the synonyms before the hypernyms when the mention string is closer to hypernyms than to synonyms while BCNH manages to return fine-grained results.

BIOSYN	BCNH
nonhodgkins lymphoma†	cell lymphomas*
non hodgkins lymphoma†	b cell lymphoma*
lymphoma non hodgkins†	lymphomas b cell*
lymphoma nonhodgkin†	lymphoma b cell*
lymphoma nonhodgkins†	lymphoma non hodgkin†
lymphoma nonhodgkin s†	lymphoma nonhodgkin†
nonhodgkin s lymphoma†	non hodgkins lymphoma†
lymphoma non hodgkin†	lymphoma non hodgkins†
hodgkins lymphoma	nonhodgkins lymphoma†
lymphoma non hodgkin s†	lymphoma non hodgkin s†

Table 2: Changes in the top 10 predictions given the mention from the NCBI test set. Synonyms having correct CUIs are indicated with an asterisk*, hypernyms are indicated with a dagger†.

Ablation study

We conduct the ablation study to figure out the contributions of the two proposed components. The results are presented in Table 3. The first experiment reports the results of BIOSYN and the second reports for BIOSYN with a joint hypernym norm constraint. The third experiment reports the results of BCNH with list-wise training only, and the last experiment reports for BCNH with both list-wise training and norm constraint.

Models	Acc@1	Acc@5
BIOSYN	89.88	93.82
BIOSYN (+ norm)	90.07 (+0.19)	94.18 (+0.36)
BCNH (+ list-wise)	90.55 (+0.67)	94.81 (+0.99)
BCNH (both)	**90.61** (+0.73)	**95.00** (+1.18)

Table 3: Ablation study results for each component.

The results demonstrate that norm constraint indeed endows the concept entity representation with the hypernym-hyponym hierarchy structure. It also verifies that hypernyms are beneficial for harder negative sampling and paying attention to all candidate entities including hypernyms list-wisely is more appropriate than marginalization solely on the synonyms.

4 Conclusion

In this paper, we propose BCNH to leverage hypernyms in the biomedical concept normalization task. We adopts both list-wise training and norm constraint with the help of hypernym information. The experimental results on the NCBI dataset show

that BCNH outperforms previous state-of-the-art models.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61922085, No.61976211, No.61906196) and the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006). This work is also supported by Beijing Academy of Artificial Intelligence (BAAI2019QN0301), the Open Project of Beijing Key Laboratory of Mental Disroders (2019JSJB06) and in part by the Youth Innovation Promotion Association CAS.

References

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Elliot G Brown, Louise Wood, and Sue Wood. 1999. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136.

Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. 2012. Medic: a practical disease vocabulary used at the comparative toxicogenomics database. *Database*, 2012.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. 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)*, pages 297–302, Beijing, China. Association for Computational Linguistics.

Shobeir Fakhraei, Joel Mathew, and José Luis Ambite. 2019. Nseen: Neural semantic embedding for entity normalization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 665–680. Springer.

Zongcheng Ji, Qiang Wei, and Hua Xu. 2020. Bert-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings*, 2020:269.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21798–21809. Curran Associates, Inc.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Robert Leaman and Zhiyong Lu. 2016. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32(18):2839–2846.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. 2017. Cnn-based ranking for biomedical entity normalization. *BMC bioinformatics*, 18(11):79–86.

Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, Pawan Goyal, Jitesh Pillai, Amitava Bhattacharyya, and Mahanandeeshwar Gattu. 2019. Medical entity linking using triplet network. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 95–100, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Minh C. Phan, Aixin Sun, and Yi Tay. 2019. Robust representation learning of biomedical names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3275–3285, Florence, Italy. Association for Computational Linguistics.

Elliot Schumacher, Andriy Mulyar, and Mark Dredze. 2020. Clinical concept linking with contextualized neural representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8585–8592, Online. Association for Computational Linguistics.

Kent A Spackman, Keith E Campbell, and Roger A Côté. 1997. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145, New Orleans, Louisiana. Association for Computational Linguistics.

Dustin Wright. 2019. *NormCo: Deep disease normalization for biomedical knowledge base construction*. Ph.D. thesis, UC San Diego.