Proceedings of the EURALI Workshop @LREC2020, pages 75–80
Marseille, 20 June 2022

Towards Bengali WordNet Enrichment using Knowledge Graph Completion Techniques

Sree Bhattacharyya [α], **Abhik Jana** [β]
[α]Indian Institute of Engineering Science and Technology Shibpur, India,
[β]Universität Hamburg, Germany
sreebhattacharyya.ug2018@cs.iiests.ac.in, abhik.jana@uni-hamburg.de

Abstract

WordNet serves as a very essential knowledge source for various downstream Natural Language Processing (NLP) tasks. Since this is a human-curated resource, building such a resource is very cumbersome and time-consuming. Even though for languages like English, the existing WordNet is reasonably rich in terms of coverage, for resource-poor languages like Bengali, the WordNet is far from being reasonably sufficient in terms of coverage of vocabulary and relations between them. In this paper, we investigate the usefulness of some of the existing knowledge graph completion algorithms to enrich Bengali WordNet automatically. We explore three such techniques namely DistMult, ComplEx, and HolE, and analyze their effectiveness for adding more relations between existing nodes in the WordNet. We achieve maximum Hits@1 of 0.412 and Hits@10 of 0.703, which look very promising for low resource languages like Bengali.

Keywords: Bengali WordNet, Knowledge graph, Automatic enrichment

1. Introduction

Several NLP applications use WordNet which was first introduced by Miller (1995) and is one of the important human-curated resources. WordNet has several NLP applications (Morato et al., 2004), including information retrieval (Mandala et al., 1998), query expansion (Smeaton et al., 1995; Gong et al., 2005; Pal et al., 2014), improvement of text retrieval responses (Gonzalo et al., 1998) and natural language generation (Jing, 1998), to name a few. WordNet in English is sufficiently rich in terms of vocabulary and semantic relation coverage. On the other hand, even though Bengali is one of the most widely spoken language[1], the Bengali WordNet (Bhattacharyya, 2010) is far from reaching the coverage of English WordNet. Therefore, automatic expansion of such a lexical resource for low-resource languages could be really useful. With this goal, we investigate whether it is possible to apply the existing knowledge graph completion techniques on the WordNet to accurately predict relations between different concepts. This direction could be leveraged in further enriching the WordNet by automatically predicting new relational links.

In this work, we attempt to enrich Bengali WordNet with the use of existing knowledge graph completion techniques. Towards this goal, we modify the original structure of the existing Bengali WordNet to make it suitable to be used as input to those algorithms. As a part of our investigation, we explore the applicability of three knowledge graph completion techniques namely DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016) and HolE (Nickel et al., 2016). Note that, in this work, we pose this WordNet enrichment problem as a closed world problem where no new node gets added to the graph, only edges (signifying semantic relations) get added. We achieve three-fold cross-validation MRR of 0.504 (maximum) and Hits@1 of 0.412 (maximum) for Bengali WordNet which is really promising for such low resource languages. This investigation presents a clear view of whether these models are able to capture the semantic intricacies of this WordNet. Note that, the main challenge for these models to work accurately on the WordNet is the existence of hierarchical semantic relations. This analysis could also be significantly useful in further establishing a methodology of using larger unsupervised lexical resources to automatically increase the coverage of the WordNet, by applying knowledge graph completion techniques. We make all our code and data publicly available[2]. The rest of the paper is organised as follows: Section 2 provides a brief account of Related Work, Section 3 describes about the Bengali WordNet and the methodology of how the graph structure of Bengali WordNet is altered, Section 4 describes the experimental analysis, and Section 5 draws the conclusion.

2. Related Work

A lot of efforts have been made to deal with the problem of automatic knowledge graph completion. Additive models include TransE (Bordes et

[1]At present, there are roughly 7000 languages in the world, among which Bengali is the 7th most widely spoken [1]

[2]`https://github.com/uhh-lt/bengali-wordnet-extension`

al., 2013), TransH (Wang et al., 2014), TransM (Fan et al., 2014), TransR (Lin et al., 2015), where the relations in the knowledge graph are regarded as translation vectors, translating a head entity to a tail entity. Multiplicative models like DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), embed entities and relations into a unified continuous vector space, followed by defining a scoring function to measure the authenticity of the triples. There are several other neural network based models like ConvKB (Nguyen et al., 2018), ConvE (Dettmers et al., 2018), HypER (Balažević et al., 2019), CompGCN (Vashishth et al., 2020) and SACN (Shang et al., 2019) among others. Models like HAKE (Zhang et al., 2020) are also able to accurately model the hierarchical semantic relationships of knowledge graphs. Further, several different approaches for WordNet enrichment or completion have been developed in the past. Biemann et al. (2004) describes a language-independent approach for semiautomatic extension of WordNets using a bootstrapping method. Biemann et al. (2018) presents a framework for combining information from distributional semantic models with manually constructed lexical resources. The framework is applied to produce a novel hybrid resource obtained by linking a disambiguated distributional lexical network to WordNet and BabelNet. In language-specific examples of previous related work, (Lee et al., 2001), introduce a semi-automatic method to construct a Korean noun semantic hierarchy by using a monolingual (Japanese) thesaurus and Korean MRD and uses the advantage of the similarity between the two languages. Rahit et al. (2018) introduce a baseline for a Bengali WordNet (BanglaNet). It uses an approach of making semantic relations between Bengali WordNet and Princeton WordNet (Miller, 1995), which is used to derive relations between concepts. Chakravarthi et al. (2018) present an expand approach for generating and improving WordNets, which uses machine translation and applies to the Dravidian languages of Tamil, Telugu and Kannada.

3. Methodology

Since our objective is to enrich Bengali WordNet, we first discuss the WordNet itself. Then, we discuss the approach to process the WordNet such that it could be fed to knowledge graph completion techniques. Finally, we discuss the three techniques we follow for our analysis.

3.1. Bengali WordNet

For our experiment we use the Bengali WordNet from IndoWordNet (Bhattacharyya, 2010). It consists of 36346 synsets (categorized as 27281 nouns, 2804 verbs, 5815 adjectives, 445 adverbs), which

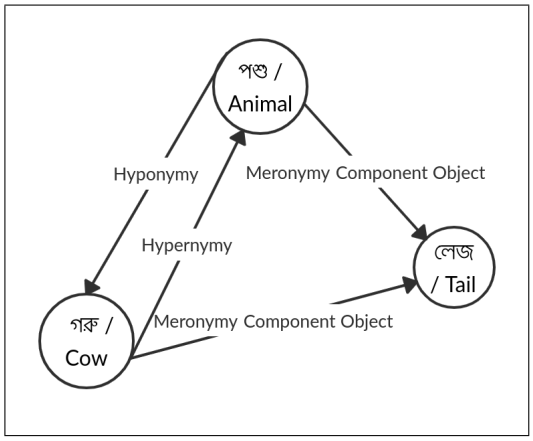

Figure 1: A snapshot of synsets as nodes and relations as edges in Bengali WordNet

are connected to other synsets through 30 different relations. The relations can be put into the following broad categories: hypernymy, hyponymy, meronymy, holonymy, ability verbs, noun attributes, verb causatives, verb entailments, function verb, and troponymy. The total number of unique words present in the Bengali WordNet is 45497. Each synset can have multiple example words (lemma names), and one head word. One word can appear in more than one synset. To obtain the data in a graphical format, the application programming interface (API) introduced by Panjwani et al. (2018) is used. As per this API, A list of edges is created - where both the nodes are synsets and the relation connecting them is the edge attribute. 72703 such edges are obtained to create a graph. In that, 36039 synsets are connected out of the total 36346 synsets of the WordNet. A snapshot is shown in Figure 1.

3.2. WordNet Structure Modification

The WordNet graph is present in the following form - synset represents a node, and different synsets are connected to each other through several relations which are represented by labeled edges. We modify the original structure to obtain a graph where each constituent word present under a Synset, becomes a node, and existing edges are replicated to connect such nodes, having relations as their attributes. To achieve this modification two things are done - First, A relation is introduced -'Synonymy' in addition to the existing 30 relations. This relation associates the words present under the same synset with each other. Edges are created connecting all possible pairs of words within the same synset, with edge attribute as the 'Synonymy' relation. Next, edges are created between every possible pair of words, which are present in two different synsets, and the edges are labelled with the same relation which is shared by the corresponding synsets that the words belong to. For any two synsets connected by an edge with

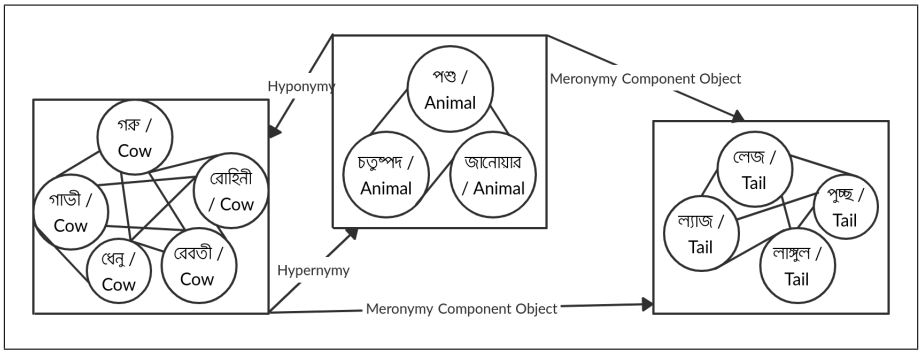

Figure 2: A snapshot of the modified WordNet graph structure. Each box is one synset. Each pair of nodes within the same box is connected to each other with the Synonymy relationship. Each pair of nodes present within two different boxes shares the relation shared between the respective boxes.

attribute 'r' (relation between synsets is 'r'), edges are created for all possible word pair combinations of the words present under both the synsets and the edge attribute assigned to the edges is also 'r'. The directionality of edges between synsets is also considered and maintained when creating edges with words as nodes. The modified structure of WordNet is presented in Figure 2. With the modified graph structure, each edge is now regarded as a triple, having subject and object entities (words), and a predicate (relation) connecting the two. The synonymy triples contain only one example of the form (head, Synonymy, tail) for two unique head and tail words. In other words, Synonymy is treated as an undirected relationship. For all the directed relationships, triples for that relationship and its inverse (if that exists in the WordNet), occur separately. For example, for two unique head and tail words, (head, Hypernymy, tail) and (tail, Hyponymy, head) both occur separately if the parent synsets of head and tail are also similarly related in the original graph.

Model	MRR	Hits@1	Hits@3	Hits@10
DistMult	0.438	0.342	0.472	0.643
ComplEx	0.492	0.382	**0.558**	0.696
HolE	**0.504**	**0.412**	0.538	**0.703**

Table 1: Three-fold cross-validated model performances.

3.3. Knowledge Graph Completion Approaches:

The three techniques used for this task are DistMult (Yang et al., 2015), HolE (Nickel et al., 2016) and ComplEx (Trouillon et al., 2016). In a nutshell, all of these models use the learned embedding vectors of the entities and relationsand use unique scoring functions to score triples (head, relation, tail). The details of these methods are described below.

DistMult: DistMult is a multiplicative model, and uses a bi-linear scoring function (Lin et al., 2018) for a triple (h, r, t) which is defined as:

$$f_r(h,t) = h^T M_r t \tag{1}$$

M_r is a 2-D matrix operator instead of a tensor operator, and is a diagonal matrix.

ComplEx: ComplEx employs eigenvalue decomposition model to take complex valued embeddings into consideration. It uses Hermitian dot product, the complex counterpart of standard dot product between real vectors. The scoring function for a triple (h, r, t) of ComplEx is defined as:

$$f_r(h,t) = sigmoid(X_{hrt}) \tag{2}$$

and $f_r(h,t)$ is expected to be 1 when (h,r,t) holds, otherwise -1. Here, X_{hrt} is further calculated as follows: $X_{hrt} =< Re(w_r), Re(h), Re(t) > + < Re(w_r), Im(h), Im(t) > - < Im(w_r), Re(h), Im(t) > - < Im(w_r), Im(h), Re(t) >$ where $M_r \in R^{d \times d}$ is a weight matrix, $< a,b,c >= \sum_k a_k b_k c_k$, Im(x) indicates the the imaginary part of x and Re(x) indicates the the real part of x (Lin et al., 2018).

HolE: Holographic Embeddings (HolE) uses circular correlation to create compositional representations of entire knowledge graphs, which is related to holographic models of associative memory. The circular correlation is denoted by: (Lin et al., 2018)

$$[a * b]_k = \sum_{i=0}^{d-1} a_i b_{(i+k) mod d}$$

The score function for a triple (h, r, t) is given as (Lin et al., 2018):

$$f_r(h,t) = sigmoid(r^T(h * t))$$

For the implementation of these three algorithms, the Ampligraph framework (Costabello et al., 2019) is used[3].

[3] https://github.com/Accenture/AmpliGraph

	Head	Translation	Relation	Tail	Translation
DistMult	মুনি	Spiritual mentor	Hyponymy	কাশ্যপ_ঋষি	An Indian spiritual mentor
	অসুর	Demon	Synonymy	রাক্ষস	Evil Spirit
ComplEx	ব্যক্তি	Human being/person	Synonymy	বান্দা	Person
	কঠিন	Difficult	Modifies_noun	কাজ	Work
HolE	ঘরোয়া	Relating to the home	Modifies_noun	জিনিস	Tangible Object
	দমন_করা	Oppress	Hypernymy	করা	To do

Table 2: Some of the top predicted triples by each knowledge graph completion approaches

4. Experiments and Analysis

For our experiment, we fix embedding dimensions to 100, epochs to 10, negatives generated per positive to 50, and optimizer to Adagrad (Duchi et al., 2011). The search space for the learning rate is set to [0.0001, 0.001, 0.01, 0.1], the losses chosen are Pairwise Loss (Bordes et al., 2013), Absolute Margin Loss (Hamaguchi et al., 2017) and Negative Log Likelihood Loss (Trouillon et al., 2016). The search space for the margin parameter is set to [0.5, 2, 10]. The optimization search is carried out both without and with L2 regularization, and the search space for λ is set to [1e-3, 1e-4, 1e-5]. After performing the grid search, the best hyper-parameter set-up for which the results are obtained are described as follows: DistMult with a learning rate of 0.1, L2 regularization with λ = 0.001, and Negative Log Likelihood Loss. ComplEx is trained with a learning rate of 0.01, L2 regularization with $\lambda = 0.0001$, and Pairwise Loss with margin = 0.1. HolE uses a learning rate of 0.1, no regularization, and Pairwise Loss with margin = 0.5. The model performances are obtained by three-fold cross-validation, with test size being equal to 40000.

4.1. Metrics Used

The following rank-based metrics are used for evaluation:

Mean Reciprocal Rank: The Reciprocal Rank (RR) information retrieval measure calculates the reciprocal of the rank at which the first relevant document was retrieved. For a single query, the reciprocal rank is 1 where the rank is the position of the highest-ranked answer (1,2,3,...,N for N answers returned in a query). If no correct answer is returned in the query, then the reciprocal rank is 0. When averaged across queries, the measure is called the Mean Reciprocal Rank (MRR) (Craswell, 2009). It is formulated as follows:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_{(s,p,o)_i}} \quad (3)$$

where Q is a set of triples and (s,p,o) is a triple $\in$ Q.

Hits@N Score: Intuitively, hits@N refers to the count of positive triples which are present in the top-N positions. Link prediction models generate a score for each of the triples, which is used to rank all the triples present. It is formally defined as:

$$Hits@N = \sum_{i=1}^{|Q|} 1\, if rank_{(s,p,o)_i} \leq N \quad (4)$$

where Q is a set of triples and (s,p,o) is a triple $\in$ Q.

4.2. Performance Analysis

The performances of the three models are presented in Table 1. HolE produces the best MRR, Hits@1, and Hits@10, whereas ComplEx produces the best Hits@3. DistMult proves to be the weakest among these three approaches. Given that, Bengali WordNet size is not that big the maximum obtained MRR of 0.504 and maximum obtained Hits@1 of 0.412 is really promising. Some of the top predicted triples by each of the three link prediction techniques are shown in Table 2. The results show that semantically close words are predicted to have almost accurate relationships between them. As a first ever attempt to enrich Bengali WordNet using knowledge graph completion techniques, we believe the results are encouraging for investigating further in this direction.

5. Conclusion

In this study, we show that off-the-shelf knowledge graph completion approaches like DistMult, ComplEx, and HolE produce promising results for predicting Bengali WordNet relations as well. This work could help largely in further enriching the WordNet using an unsupervised resource like a thesaurus. Inclusion of predicted links in the WordNet should be preceded by manual correction to ensure overall accuracy of WordNet. This work could ultimately be useful for tasks like word sense disambiguation, machine translation, etc. The immediate future work could be exploring other categories of such algorithms to be applied on WordNet for the same purpose. The broad plan is to create a framework to enrich WordNet of a range of low resource languages automatically without human intervention.

6. Bibliographical References

Balažević, I., Allen, C., and Hospedales, T. M. (2019). Hypernetwork knowledge graph embeddings. In *International Conference on Artificial Neural Networks*, pages 553–565. Springer.

Bhattacharyya, P. (2010). Indowordnet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Biemann, C., Shin, S., and Choi, K.-S. (2004). Semiautomatic extension of corenet using a bootstrapping mechanism on corpus-based cooccurrences. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1227–1232.

Biemann, C., Faralli, S., Panchenko, A., and Ponzetto, S. P. (2018). A framework for enriching lexical semantic resources with distributional semantics. *Natural Language Engineering*, 24(2):265–312.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2018). Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86, Nanyang Technological University (NTU), Singapore, January. Global Wordnet Association.

Costabello, L., Pai, S., Van, C., McGrath, R., McCarthy, N., and Tabacof, P. (2019). Ampligraph: a library for representation learning on knowledge graphs. *Retrieved October*, 10:2019.

Craswell, N. (2009). Mean reciprocal rank. *Encyclopedia of database systems*, 1703.

Dettmers, T., Minervini, P., Stenetorp, P., and Riedel, S. (2018). Convolutional 2d knowledge graph embeddings. In *Thirty-second AAAI conference on artificial intelligence*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Fan, M., Zhou, Q., Chang, E., and Zheng, F. (2014). Transition-based knowledge graph embedding with relational mapping properties. In *Proceedings of the 28th Pacific Asia conference on language, information and computing*, pages 328–337.

Gong, Z., Cheang, C. W., and Hou, U. L. (2005). Web query expansion by wordnet. In *International Conference on Database and Expert Systems Applications*, pages 166–175. Springer.

Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with wordnet synsets can improve text retrieval. *arXiv preprint cmp-lg/9808002*.

Hamaguchi, T., Oiwa, H., Shimbo, M., and Matsumoto, Y. (2017). Knowledge transfer for out-of-knowledge-base entities : A graph neural network approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1802–1808.

Jing, H. (1998). Usage of wordnet in natural language generation. In *Usage of WordNet in Natural Language Processing Systems*.

Lee, J., Un, K., Bae, H.-S., and Choi, K.-S. (2001). A korean noun semantic hierarchy (wordnet) construction. In *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*, pages 290–295.

Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*.

Lin, Y., Han, X., Xie, R., Liu, Z., and Sun, M. (2018). Knowledge representation learning: A quantitative review. *arXiv preprint arXiv:1812.10901*.

Mandala, R., Tokunaga, T., and Tanaka, H. (1998). The use of wordnet in information retrieval. In *Usage of WordNet in Natural Language Processing Systems*.

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

Morato, J., Marzal, M. A., Lloréns, J., and Moreiro, J. (2004). Wordnet applications. In *Proceedings of GWC*, pages 20–23.

Nguyen, D. Q., Nguyen, T. D., Nguyen, D. Q., and Phung, D. (2018). A novel embedding model for knowledge base completion based on convolutional neural network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 327–333, New Orleans, Louisiana, June. Association for Computational Linguistics.

Nickel, M., Rosasco, L., and Poggio, T. (2016). Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Pal, D., Mitra, M., and Datta, K. (2014). Improving query expansion using wordnet. *Journal of the Association for Information Science and Technology*, 65(12):2469–2478.

Panjwani, R., Kanojia, D., and Bhattacharyya, P. (2018). pyiwn: A python based api to access indian language wordnets. In *Proceedings of the 9th Global Wordnet Conference*, pages 378–383.

Rahit, K. T. H., Hasan, K. T., Al-Amin, M., and Ahmed, Z. (2018). Banglanet: Towards a word-

net for bengali language. In *Proceedings of the 9th Global Wordnet Conference*, pages 1–9.

Shang, C., Tang, Y., Huang, J., Bi, J., He, X., and Zhou, B. (2019). End-to-end structure-aware convolutional networks for knowledge base completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3060–3067.

Smeaton, A. F., Kelledy, F., and O'Donnell, R. (1995). Trec-4 experiments at dublin city university: Thresholding posting lists, query expansion with wordnet and pos tagging of spanish. *Harman [6]*, pages 373–389.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. (2016). Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR.

Vashishth, S., Sanyal, S., Nitin, V., and Talukdar, P. (2020). Composition-based multi-relational graph convolutional networks. In *International Conference on Learning Representations*.

Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014). Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Yang, B., Yih, W., He, X., Gao, J., and Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Zhang, Z., Cai, J., Zhang, Y., and Wang, J. (2020). Learning hierarchy-aware knowledge graph embeddings for link prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3065–3072.