Enriching WordNet with Subject Specific Out of Vocabulary Terms Using Existing Ontology

Conference Paper · January 2021

CITATIONS
0

READS
34

5 authors, including:

Kanika Kanika
Netaji Subhas Institute of Technology
10 PUBLICATIONS **42** CITATIONS

SEE PROFILE

Enriching WordNet with Subject Specific Out of Vocabulary Terms using Existing Ontology

Kanika[1], Shampa Chakraverty[1], Pinaki Chakraborty[1], Aditya Agarwal[1], Manan Madan[1] and Gaurav Gupta[1]

[1]Netaji Subhas Institute of Technology, Delhi, India
kanikatehlan@gmail.com

Abstract. WordNet is a huge repository being used as a tool in various fields. With an increasing number of applications referring to WordNet as a dictionary, several attempts have been made to update it. The paper proposes to extend the huge repository by adding words and relationships derived from students' class notes through wikidata. These terms can be phrases, technical terms or any subject specific terminology appearing in students' notes of a specific subject. Although various WordNet enriching techniques are available, it is for the first time that subject specific terminology is being added. The resulting version of WordNet has some very common phrases and technical terms along with the generic terms. Making subject specific and generic terms available in a hierarchy can improve the accuracy of various applications like text summarization and clustering for text belonging to a specific domain.

Keywords: English WordNet, Hyponym enrichment, Wikidata.

1 Introduction

WordNet is a recognized source of conceptual information for all kinds of linguistic processing in English[1]. It is a lexical resource extensively used as a research tool in Natural language processing (NLP). In the database, nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms called synsets. Each synset represents a unique concept [2]. The words are linked with respect to a particular sense in which they are used. Hence, we can say that the database also labels the semantic relations among words [3]. The huge repository is a useful tool for computational linguistics and NLP because of its structure. Improved general word sense disambiguation [4] and domain specific word sense disambiguation [5] are two most common applications where WordNet is directly used. It is a great resource to facilitate text categorization [1, 6], text summarization [7], and document clustering [8] too.

The use of WordNet is not just confined to text and documents, but the repository is also used to improve the online searching and learning experiences. Search engines utilize the information stored in hierarchy to improve the precision of search query

results [9]. It also helps in building dynamic learner profiles by extracting learners' interest [10]. The online dictionary is also used in extracting aspect terms from online reviews efficiently [11]. Computing semantic similarity between concepts [12]and learning a well-founded domain ontology [13] are a few other areas where WordNet plays a crucial role.

For almost two decades now, the Artificial Intelligence community working on computational linguistics and other fields has been using WordNet. While the resource has proved to be an excellent knowledge base, since it is a dictionary, it needs to evolve just like human language [14]. The enrichment and expansion of the online dictionary is also required to make it the best research tool for the newly emerged domain specific applications using it. In short, to prevent WordNet from failing to catch up with new applications and relations of various concepts and techniques [14], enriching it with "out of vocabulary" terms is an effective solution. Utilizing wiktionary [15] Wikipedia [16],Topic signatures [17], and using the meta properties of concepts [13] are some of the ways one can improve the lexical database.

In this paper, we try to utilize the enormous knowledge base available through wikidata to incorporate subject related terms into WordNet. The idea is to collect terms specific to a subject from notes, find more related terms present at wikidata and place them at an appropriate place in the hierarchy. The proposed inclusion of subject specific terms will lead to an enriched version of WordNet that may efficiently assist in several e learning applications. The paper proposes a novel hyponym enrichment in WordNet by enriching the database with subject related technical terms. This is important since the database has not been updated for a long time now [18]. To the best of our knowledge, we are the first to use wikidata for adding subject specific terms and relationships to WordNet.

The remaining paper is structured as follows. In Section 2, we discuss the various approaches used for WordNet enhancement and expansion in the past. Section 3 expounds the methodology adopted to add the new terms to the database. Section 4 presents the results and analysis of our attempt to enrich WordNet enrichment. . We conclude and give future possibilities in the next section.

2 Related Work

With an enormous increase in the applications relying on the huge repository, a lot of attempts have been made to enrich, extend and improve the contents of WordNet. Many automatic and semi- automatic approaches that improve WordNet in one way or the other exist. One way proposed almost two decades ago was the use of immense information already present on the World Wide Web (WWW). Since WordNet presents hierarchical information, researchers believe that spotting and fixing errors and

ambiguities at different levels of taxonomy can result in varying betterments. [19] attempted to detect existing anomalies in the WordNet. The authors target to improve the quality of the database by using an automatic method to propagate domain information. The study revolved around assigning labels to unlabelled synsets in WordNet 3.0 and new domain labeling for synsets with variations. On comparison through a word sense disambiguation task, clear improvements were observed with the new labeling mechanism.

Verdezoto and Vieu [20] aimed to improve the top level of the taxonomy. For this, they proposed a semi automatic error detection approach to spot errors in the lower levels of taxonomy. The system proposed detected errors automatically, however, there was a need for intervention from lexicographers and domain experts to decide on the part of solving the error. Slightly different approach was adopted by [13]. Rather than fixing the lower level errors and expecting its impact on top levels of taxonomy, the authors mapped the noun synsets to top level constructs of Unified Foundational Ontology (UFO). The semantically enriched WordNet thus generated had a wide scope of domain specific improvements with the philosophical meta properties of concepts available [13]. Apart from improving the quality, spotting errors and dealing with ambiguities, several proposals for extending the database also exist.

CROWN (Community enRiched Open WordNet) is one such example that improves quality of WordNet with technical words and idioms. The authors try to grow the size of WordNet by adding hypernym and antonym relations. For adding information at appropriate places, they used wiktionary. Wiktionary is a collaboratively constructed online dictionary. With CROWN, one can end up growing WordNet to double its size [15]. Not only by increasing its size by adding a variety of lemmas or establishing new relationships, but researchers have also attempted to improve the dense and complicated structure by enhancing the visualization. There are arguments that by improving the visualization, the understanding of WordNet connections can improve. The authors embed the concept of tag clouds into the synonyms rings present in WordNet. The results were improved human recognition of different senses in which words are used [21].

Considering the fact that WordNet for English is with us for more than two decades now, and the same version with slight variations was used by various applications relying on it, setting up new guidelines to update English WordNet is going to give the much required upgrade to the lexical database [22]. The introduction of new synsets and senses are underway. Developing standard guidelines for this addition and the integration of contributions from other projects are some recent developments in the enrichment of WordNet [18].Some studies have tried to exploit the structure of wikipedia to reach correct synsets. However, the study maps those synsets to 14 languages other than English [16]. The aim of this paper is to improve the WordNet in English language. Since there are evidences that we can develop the existing WordNet using wiki resources such as wikipedia [23], and using Wikipedia to build and improve WordNet is not yet explored in depth [16], we try to analyze the potential of

aligning the subject specific terms from wiki resources with WordNet. However, instead of using wikipedia, we used wikidata- an ever evolving resource for structured data, relationships and taxonomies [24].

3 Methodology

In order to add some more subject specific terms in a meaningful manner to WordNet, we try to exploit the data arranged in hierarchical manner in wikidata. Wikidata is an open source for structured data utilized by other resources like wikipedia [24]. Being one of the fastest growing wiki resources, one can rely on wikidata for updating crucial online dictionaries like WordNet. For demonstration purposes, we add terms related to Artificial Intelligence a computer science subject to WordNet. In order to collect relevant terms, we use classroom notes. We took notes on the introduction to Artificial Intelligence (AI) from undergraduate students of Netaji Subhas Institute of Technology studying computer engineering. These notes are processed and keywords are extracted using the following steps.

3.1 Preprocessing the data

In order to break down the entire text into meaningful sentences and syllables, we first remove/ replace the pronouns first. During this process, it is made sure that the meaning the text conveys remains unchanged. For this, the data is given as input to a coreferencing tool. After co-referencing the entire text, it is now broken into sentences using an NLTK sentence tokenizer. To the words of these sentences, the system checks for any spelling mistakes. In case a particular word is not matching any word in the dictionary, it is converted to the closest match available.

3.2 Extract informative words

To the corrected set of words, part of speech tagging is applied and this assigns each word to a lexical category. This is done to give each word a tag that helps in identification of the word in categories such as noun or adjective. Since it is believed that nouns carry most of the valuable information in any text, from the collection of tagged words, the system selects all nouns and nouns followed by adjectives as keywords. These informative words are then treated as nodes/ vertices and a graph is created using these nodes.

3.3 Generate graph using wikidata

We have a set of vertices represented by important keywords. The next step is to establish linkages between these vertices. For this, for every informative word/ phrase, a

closest equivalent called generic name existing on wikidata is used. Let us assume this generic name represents a wiki entity. Along with every generic name extracted, the corresponding wikidata id is also stored. For example, the informative word "maths" is converted into mathematics. The system ends up storing mathematics and its corresponding wikidata id which is Q395. Using SPARQL, a query language for accessing the data available on wikidata, the system stores the ancestor and descendant of every wiki entity. For illustration, the process of associating ancestors and descendants is repeated up to 3 levels. We now have a directed cyclic graph, G with each node representing a wiki entity related to a particular subject somehow. So, G={T, E} where, T is the set of subject specific terms present as wiki entities on wikidata and E is the set of edges between these terms. Each edge in G represents one of the two types of relations (i)"is a part of" and (ii) "is a subclass of". The graph thus generated represents the collection of wiki entities in a hierarchy. Due to overlap in topics each entity can belong to, as you go up in hierarchy, G can contain more than just one subject/domain.

3.4 Identify out of vocabulary (OOV) terms in WordNet

The system then traverses the graph G to reach every node present in T as concept. On reaching a particular vertex representing a term T_i in the graph, it searches for T_i in the WordNet corpus. The search can result in following outcomes:(i) T_i from graph G is available on WordNet and (ii) It is unavailable. The system searches for the child nodes of T_i in the WordNet database. This search can result in one of the three scenarios. The first case is when no child node of Ti is present in WordNet: This is the case when T_i but none of its children are present in the database. The term T_i alone is present in the existing database. Another scenario is when a node is present but not connected. The two nodes under consideration are present in the database, but are not directly related to each other through a hyponym/hypernym relation. Present and connected is the third situation. In such a case both T_i and the corresponding child node, T_j are present and are connected through any relation. This means both- the terms as well as relationship exist. A slightly different case is when the current node, T_i, is not available on WordNet. This is possible only for the root node of G as all other nodes will be treated as child nodes and the cases are listed above. This is the case when the root node of the Graph G is not present in the WordNet database.

3.5 Assign a place in WordNet hierarchy

All the terms present in G fall into one of the five categories listed above. According to the scenario, following actions are taken:

T_i present but not the child T_j from G: If T_j is not present in the database, the same is appended to the hierarchy in such a way that it is a hyponym T_i.

Present but not connected: In such cases, we propose a relation between the vertex- T_i and its child node- T_j in accordance to the wiki graph G.
Present and connected: If both T_i and T_j are present and are connected through any relation, we move on to the next vertex.
Root of the graph is not found: If the root of the graph is not found, the system searches for any of its child nodes in WordNet. If a child node is present, the root node is added as a hypernym to the child node. Else, add both the root node and its child to WordNet connected with a hyponym-hypernym relation.

4 Results

The proposed method ends up improving the WordNet by adding several subject specific terms. We take Artificial intelligence as a subject for demonstration purpose. From the notes of one class of artificial intelligence, a total of 3.76% of the terms were either not present on WordNet or an appropriate relation according to wikidata was missing. In Fig. 1, the key term artificial intelligence was present on the online dictionary. However, “expert systems” is not available on the huge dictionary. According to the hierarchy provided by wikidata, expert system is a subclass of Artificial intelligence.

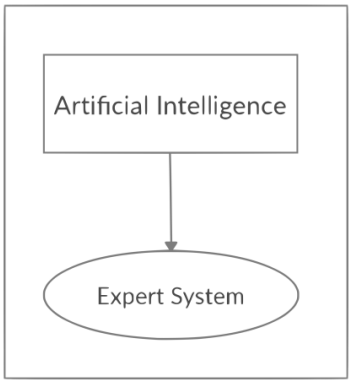

Fig. 1: Addition of a new node under Artificial intelligence

Hence it is added as a node following the relation determined by wikidata. Although there may be many other connected terms for the parent term, Fig. 1 shows only the newly added term. Similarly, chess theory is added as a new node under the term “chess” (see Fig. 2). It was observed that terms as common as tree data structure and binary tree were not available on WordNet. So, adding such terms can increase the vocabulary.

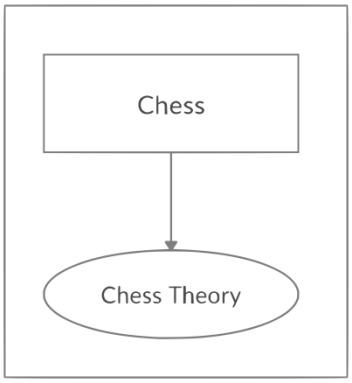

Fig. 2: Addition of new term related to the term chess

Conclusion

The paper proposes an extended version of WordNet that incorporates technical terms and subject specific words/ phrases. These phrases are extremely common in texts belonging to certain subjects. However, it was observed that not many of them were present on the existing WordNet database. WordNet is an online dictionary and just like human language, there is a need to update it from time to time. Apart from adding new out of vocabulary terms, one way to increase the accuracy of applications revolving around a domain is to populate the database with terms belonging to that domain. We aimed at enriching WordNet with subject specific terms. For experimental purposes, we extracted terms from notes on a topic of Artificial Intelligence. Out of 186 subject specific extracted from the notes, we ended up adding 3.6% of the terms to WordNet.

The experiment is based on a small data of 186 words. It can be scaled up by taking into account notes of various other topics. In future, we will gather data from notes of an entire course. This will give more wiki entities covering almost the entire subject along with the relations. Even though we have considered a technical subject, the same method can be employed to incorporate terms related to any subject like social science or biology into WordNet.

References

1. Elberrichi, Z., Rahmoun A, and Bentaalah M. A.: Using WordNet for Text Categorization. International Arab Journal of Information Technology 5(1), 16-24 (2008).
2. Miller, G. A.: WordNet: a lexical database for English. Communications of the ACM 38 (11), 39-41 (1995).

3. Miller, G. A.: WordNet: An electronic lexical database. MIT press (1998).
4. Wang, Y., Ming W., and Fujita H.: Word sense disambiguation: A comprehensive knowledge exploitation framework. Knowledge-Based Systems, 190 (2020).
5. Lopez-Arevalo, I., Sosa-Sosa V. J., Rojas-Lopez F., and Tello-Leal, E.: Improving selection of synsets from WordNet for domain-specific word sense disambiguation. Computer Speech & Language 41, 128-145 (2017).
6. Rodriguez, M., Hidalgo, J. and Agudo B. Using WordNet to complement training information in text categorization. In: Proceedings of 2nd International Conference on Recent Advances in Natural Language Processing II, pp. 353-364 Bulgaria (2000).
7. Pal, A. R. and Saha, D. An approach to automatic text summarization using WordNet. In: IEEE International Advance Computing Conference, pp. 1169-1173, India (2014).
8. Shehata, S. A WordNet-based semantic model for enhancing text clustering. In: IEEE International Conference on Data Mining Workshops, pp. 477-482, USA (2009).
9. Moldovan, D. I. and Mihalcea R.:Using WordNet and lexical operators to improve internet searches.IEEE Internet Computing 4(1), 34-43 (2000).
10. Sheeba, T. and Krishnan R. A Semantic Approach of Building Dynamic Learner Profile Model Using WordNet. In: Advanced Computing and Intelligent Engineering, pp. 263-272. Springer, Singapore (2020).
11. Tao, J. and Zhou L.: A Weakly Supervised WordNet-Guided Deep Learning Approach to Extracting Aspect Terms from Online Reviews.ACM Transactions on Management Information Systems 11 (3), 1-22 (2020).
12. Zhang, X, Sun, S. and Zhang, K.: A New Hybrid Improved Method for Measuring Concept Semantic Similarity in WordNet. International arab journal of information technology 17(4), 433-439 (2020).
13. Leão, F., Revoredo, K. and Baião, F.: Extending WordNet with UFO foundational ontology.Journal of Web Semantics 57, 100499 (2019).
14. Rusert, J. Language Evolves, so should WordNet-Automatically Extending WordNet with the Senses of Out of Vocabulary Lemmas (2017).
15. Jurgens, D., and Pilehvar M. T. Reserating the awesometastic: An automatic extension of the WordNet taxonomy for novel terms. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1459-1465, ACL, Colorado (2015).
16. Haziyev, F. Automatic WordNet Construction Using Wikipedia Data. Master's thesis, Fen Bilimleri Enstitüsü, 2019.
17. Agirre, E., Ansa, O., Hovy, E., and Martinez, D. Enriching WordNet concepts with topic signatures.In: Proceedings of the SIGLEX workshop on "WordNet and Other Lexical Resources: Applications, Extensions and Customizations (2001).
18. McCrae, J.P., Rademaker, A. Rudnicka, E. and Bond, F. English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. In: Proceedings of the LREC 2020 Workshop on Multimodal WordNets (MMW2020), pp. 14-19 ELRA, France (2020).
19. González, A., Rigau G., and Castillo M. A graph-based method to improve WordNet domains. In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 17-28, Springer, Berlin (2012).
20. Verdezoto, N. and Vieu L. Towards semi-automatic methods for improving WordNet. In: Proceedings of the Ninth International Conference on Computational Semantics, pp. 275-284, United Kingdom (2011).

21. Caldarola, E. G., and Rinaldi, A. M. Improving the visualization of WordNet large lexical database through semantic tag clouds. In: IEEE International Congress on Big Data (BigData Congress), pp. 34-41. IEEE, Washington (2016).
22. Khodak M., Risteski A., Fellbaum C., Arora, S.: Extending and Improving WordNet via Unsupervised Word Embeddings. Linguistic Issues in Language Technology – LiLT 10 (4), 1-17(2017).
23. Oliver, A. Aligning Wikipedia with WordNet: a Review and Evaluation of Different Techniques. In:Proceedings of the 12th Language Resources and Evaluation Conference, pp. 4851-4858 Marseille(2020).
24. Vrandečić, D. and Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78-85 (2014).