ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

Indo-Word-Net Dictionary: A Review

[1] Ritika Sharma, [2] Charanjiv Singh
[1] M. Tech (Scholar), [2] Assistant Professor
University college of Engineering, Punjabi University Patiala

Abstract: Language is a primary component of human age. Several languages are used in India respective to their corresponding families. A dictionary can be used to get different meaning of a word in different languages. Various sorts of dictionaries like encyclopedia, thesaurus, phonetic dictionaries etc are used. In this Paper, we've discussed WordNets, its origin and several tools incorporated with it. IndoWordNet is an integrated multilingual WordNet for Indian languages. In addition we've invested details about IndoWordNet Dictionary with the requirement and features of such dictionary and existing tools of IndoWordNet dictionary and the limitations of those tools. We've surveyed previous research done by analysts. Furthermore, we've discussed API's, IndoWordNet API, its need, Design and architecture of IndoWordNet API and various commands used in it.

Keywords –Indian Languages, IndoWordNet API and Dictionary.

1. INTRODUCTION

Dialect is a constituent component of human advancement. In a nation like India, decent variety is its essential angle. This prompts shifted dialects and their lingos. There are various dialects in India which have a place with various dialect families. The Eighth Schedule of the Indian Constitution records 22 dialects, which have been alluded to as planned dialects and given acknowledgment, status and authority consolation. Recently the period of computers, is an answer to such complaints. The primary motivation to keep on-line word dictionaries is that the literal data could be available any time and read by PCs. Moreover the PC can search quicker than people and in alphabetical order [1]. WordNet is a proposition for a more compelling mix of conventional lexicographic data and present day rapid calculation.

Dictionary can be called as an asset managing the individual expressions of a dialect alongside its orthography, articulation, use, equivalent words, inference, history, historical underpinnings, etc. masterminded in a request for accommodation of referencing the words. Different models utilized for grouping this asset are - thickness of passages, number of dialects included, nature of sections, level of fixation on entirely lexical information, pivot of time, course of action of passages, etc. Common dictionaries are [2]:

• Encyclopedia: Single or multi-volume distribution that contains gathered and legitimate information regarding a matter organized one after another in order. E.g. Britannica reference book.

• Thesaurus: Thesaurus is a lexicon that rundowns words in gatherings of equivalent words and related ideas.

• Etymological Dictionary: An etymological lexicon talks about the historical underpinnings/starting point of the words recorded. It is the result of research in recorded semantics.

• Dialect Dictionary: These lexicons manage the expressions of a specific geological area or social gathering which are non standard.

• Specialized Dictionary: These word references covers generally confined arrangement of marvels.

• Bilingual or Multilingual Dictionary: These are etymological lexicons in at least two dialects.

• Reverse Dictionary: These lexicons depend on the idea/thought/definition to words.

• Learner's Dictionary: These lexicons are implied for remote understudies/voyagers to take in the utilization of the word in dialect.

• Phonetic Dictionary: These lexicons help in looking through the words by the way they sound.

• Visual Dictionary: These lexicons utilize pictures to outline the importance of words.

1.1 WordNet

WordNet is literal database for the English dialect. It bunches English words into sets of equivalent words called synsets, gives short, general definitions, and records the different semantic relations between these equivalent word sets. The reason for existing is twofold: to create a blend of word reference and thesaurus that is all the more naturally usable, and to help programmed content examination and manmade brainpower applications. The database and programming assets have

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

been discharged under a BSD style permit and can be downloaded and utilized unreservedly.

WordNet was made and further maintained at Cognitive Science Laboratory of Princeton University under the guidance of George A. Mill. Development started in 1985. The project received financing from government offices keen on machine interpretation. Starting in 2006, the database contains 155,287 words sorted out in 117,659 synsets for an aggregate of 206,941 word-sense sets; in packed shape, the approximate size is 12 megabytes [3]. In 2009, the WordNet group incorporates the accompanying individuals from Cognitive Science Laboratory. WordNet has been bolstered by grants from National Science Foundation, DARPA and REFLEX. George Miller and Christiane Fellbaum were awarded by 2006 Antonio Zampolli Prize for their contributions in WordNet. It recognizes nouns, verbs, descriptors and modifiers since they take after various linguistic guidelines—it does exclude relational words, determiners and so on. The significance of synsets is additionally cleared up with short characterizing gleams

1.1.1 Software tools for WordNet

WordNet tools are required to aid the general improvement of a working WordNet. Devices can be extensively arranged relying upon the WordNet write, reason and WordNet creation approach. Figure underneath demonstrates the arrangement of the distinctive devices for a WordNet [4].

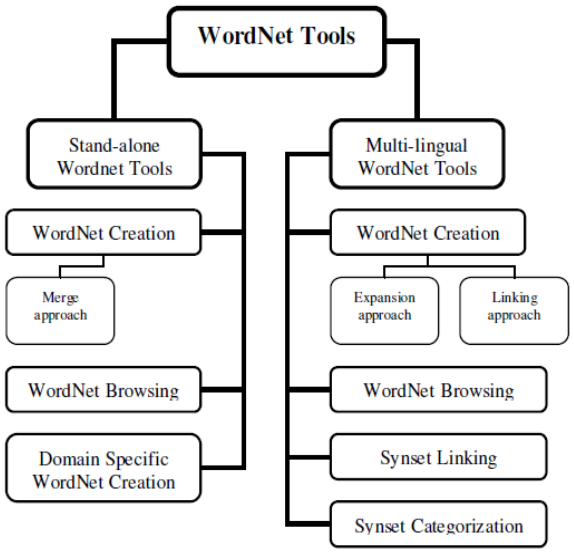

Figure 1. Categorization of software tools for wordnet. [4]

1.2 Indo Word Net Dictionary

IndoWordNet Dictionary7 or IWN Dictionary is an online interface to render multilingual IndoWord-Net data in the word reference arrange. It enables client to see the outcomes in various organizations according to the need. Additionally, client can see the outcome in various dialects at the same time. The look and feel of the IWN Dictionary is remembered same as a customary word reference keeping the client versatility. Up until this point, it renders WordNet data of 19 Indian dialects. These dialects are: Bengali, Assamese, Hindi, Bodo, Gujarati, Kannada, Kashmiri, Malayalam, Konkani, Maithili, Manipuri, Punjabi, Marathi, Nepali, Odia, Sanskrit, Tamil, Urdu and Telugu. The WordNet data is likewise rendered in English. The English WordNet data is taken from Princeton University8 site and imported into IndoWordNet database structure. The information is foreign to this database utilizing DB Import apparatus created under the Indradhanush WordNet Project9. The work is as yet going ahead to incorporate other non-Indian dialects and putting away them in the World WordNet Database Structure proposed by Redkar et al. (2015).

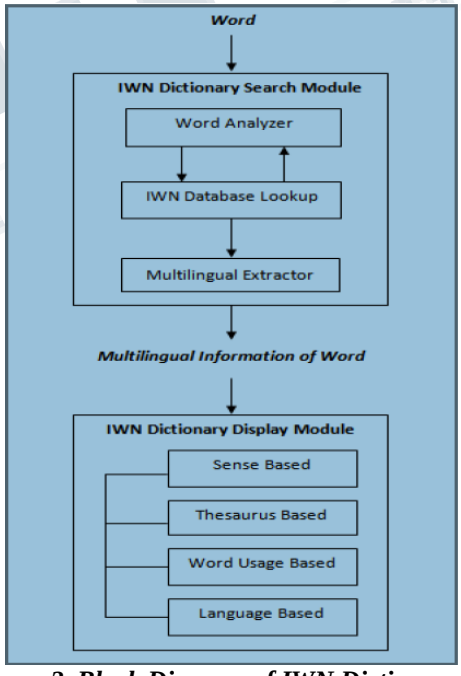

Figure 2. Block Diagram of IWN Dictionary [5]

1.2.1 Need of Indo Word Net Dictionary

The contribution to program is an inquiry string in any of the two dialects and the yield is the item for both the dialects. The essential use of this interface is to enable clients to get the semantic data of the pursuit string in

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

both Hindi and Marathi. The word reference interface permits looking between Hindi-Assamese and Hindi-Bodo dialect matches at once. Every interface can shows the meanings in mostly in 2 languages with linguistic data accessible in wordnet. Several justifications for the requirement of multi-lingual dictionaries are: [5]

• To understand different languages.
• To locate the significance of a word in various dialects on a solitary stage.
• To comprehend different dialects with the assistance of a rotate dialect.
• To comprehend synonymous words in input dialect and in another dialects.
• To comprehend extra data like grammatical form, gender, etc. in input and in other distinctive dialects.
• To comprehend the semantic varieties in various dialects.
• To satisfy the social need of connecting the correspondence hole.
• To comprehend the content of different dialects.
• To comprehend a neighbourhood dialect when moved to that territory.
• To enhance one's own dialect vocabulary.
• To address social and instructive needs.

1.2.2 Characteristics of Indo Word Net Dictionary

The striking highlights of IndoWordNet Dictionary are as per the following: [5]
• Renders data in different dialects at once.
• Different perspectives to show the data: sense based, thesaurus based, word utilization based, dialect based.
• Transliteration include accessible in various dialects which helps in perusing successfully.
• Can help dialect interpretation assignment where multi-linguality is essential concern.
• Can be alluded for instructive and social needs.
• Automatic word recommendation is helped.
• Similar/nearest word recommendation is helped.
• Usage examination and insights accessible.
• Provides measurements, for example, most as often as possible sought word regarding the chose dialect, input dialect looked, and so on.

1.2.3 Existing Tools of Indo Word Net Dictionary

Amid the improvement of IndoWordNet we felt the need of different devices created which drove us to reevaluate the capacity procedure used to store the IndoWordNet: [6]

• An independent interface enables clients to see the Hindi synsets, ideas, illustration sentences on one side and all the while keying the objective dialect synsets, ideas and illustration sentence. The device additionally has the Princeton WordNet English synsets interlinked. The produced target dialect synset files which are put away in level content records have expansion as .syns.
• English-Hindi Linkage Tool is a heuristic based device to connect Hindi-English synsets.
• Synset Categorization Tool is utilized to picked basic linkable synsets over all dialects by arranging them as Universal, Pan Indian, In family, dialect particular, Rare, Synthesized and Core.
• The Sense Marking Tool is utilized to discover the synset scope of a WordNet.

1.2.4 Limitations of Tools of Indo Word Net Dictionary

The devices utilized for the advancement of WordNets utilizing the Expansion approach were for the most part in light of level records. Level records have their own preferences however there are a few impediments as well. A portion of the issues confronted while taking a shot at the above devices were as per the following:
• Synset tallying as for various criteria, for example, getting the synset tally having a place with a particular syntactic class or range.
• Combining synset documents, discovering basic arrangement of synsets.
• Security and Data honesty
• Status of synsets – to know whether the synset is approved or not.
• Extra information about synsets (meta information – source, helpful connections, video, sound, space, pictures which gives extra data about the synset, and so forth).
• Likewise it is hard to store lexical and semantic relations amongst words and synsets in a level document.

2. LITERATURE REVIEW

Hanumant Redkar, et al., (2015) [5] focussed around giving an online interface – IndoWordNet Dictionary to non-researchers and specialists. India is a nation with different culture, dialect and fluctuated legacy. Because of this, it is exceptionally rich in dialects and their lingos. Being a multilingual society, a word reference in numerous dialects turns into its need and one of the real assets to help a dialect. There are word references for some Indian dialects, however not very many are accessible in different dialects. WordNet is a standout amongst the most conspicuous lexical assets in the field

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

of Natural Language Processing. IndoWordNet is an incorporated multilingual WordNet for Indian dialects. These WordNet assets are utilized by analysts to examination and resolve the issues in multi-linguality through calculation. Be that as it may, there are few situations where WordNet is utilized by the non-analysts or overall population. It is created to render multilingual WordNet data of 19 Indian dialects in a lexicon arrange. The WordNet data is rendered in numerous perspectives, for example, sense based, thesaurus based, word use based and dialect based. English WordNet data is additionally rendered utilizing this interface. The IndoWordNet lexicon will help clients to know implications of a word in numerous Indian dialects.

Sudha Bhingardive, et al., (2017) [7] presented the utilization of different highlights of IndoWordNet in performing WSD.Word Sense Disambiguation (WSD) is considered as one of the hardest issue in the field of Natural Language Processing. IndoWordNet is a connected structure of WordNets of real Indian dialects. As of late, a few IndoWordNet based WSD approaches have been proposed and executed for Indian dialects. Here, we have utilized highlights like connected WordNets, semantic and lexical relations, and so forth. We have taken after two unsupervised methodologies, viz., (1) utilization of IndoWordNet in bilingual WSD for finding the sense dispersion with the assistance of Expectation Maximization calculation, (2) utilization of IndoWordNet in WSD for finding the most successive sense utilizing word and sense embeddings. Both these methodologies legitimize the significance of IndoWordNet for word sense disambiguation for Indian dialects, as the outcomes are observed to guarantee and can beat the baselines.

Hanumant Redkar, et al., (2016) [8] presented an online device - Samāsa-Kartā for creating compound words. Samāsa or mixes are a general element of Indian Languages. They are likewise found in different dialects like German, Italian, French, Russian, Spanish, and so on. Compound word is developed from at least two words to shape a solitary word. The importance of this word is gotten from every one of the individual expressions of the compound. To build up a framework to produce, recognize and decipher mixes, is an essential undertaking in Natural Language Processing. Here, the attention is on Sanskrit dialect because of its abundance in utilization of mixes; be that as it may, this approach can be connected to any Indian dialect and in addition different dialects. IndoWordNet is utilized as an asset for words to be intensified. The inspiration driving making compound words is to make, to enhance the vocabulary, to decrease detect uncertainty, and so forth keeping in mind the end goal to improve the WordNet. The Samāsa-Kartā can be utilized for different applications viz., compound classification, sandhi creation, morphological investigation, summarizing, synset creation, and so forth.

Dhirendra Singh, et al., (2016) [9] focussed around two classes of MWEs - Compound Nouns and Light Verb Constructions. Discovery of Multi-Word Expressions (MWEs) is one of the principal issues in Natural Language Processing. These two classes can be handled utilizing learning bases, as opposed to unadulterated insights. We research ease of use of IndoWordNet for the location of MWEs. Our IndoWordNet based approach utilizes semantic and ontological highlights of words that can be separated from IndoWordNet. This approach has been tried on Indian dialects viz., Assamese, Bengali, Hindi, Konkani, Marathi, Odia and Punjabi. Results demonstrate that ontological highlights are observed to be exceptionally helpful for the location of light verb developments, while utilization of semantic properties for the discovery of compound things is observed to be tasteful. The proposed technique can be effortlessly adjusted by other Indian dialects. Recognized MWEs can be interjected into WordNets as they help in speaking to semantic learning.

Brijesh Bhatt, et al., (2011) [10] explained the connecting of WordNets of Indian dialects with an upper metaphysics SUMO (Suggested Upper Merged Ontology). Thinking about characteristic dialect requires joining semantically rich lexical assets with world learning, gave by ontologies. This makes multilingual asset for Indian dialects which can be utilized as a part of different characteristic dialect preparing applications. They also display the engineering of IndoWordNet-Linking of WordNets of seventeen diverse Indian dialects and furnish a technique to interface it with upper metaphysics SUMO. Two distinct frameworks: IndoWordNet guide and SIGMAKEE interface for Indian dialects are produced to get to this asset.

3. APPLICATION PROGRAMMING INTERFACE

An API is characterized as an arrangement of charges, capacities and conventions which designer can utilize when building programming. It enables the designer to utilize predefined capacities to associate with frameworks, rather than keeping in touch with them without any preparation. The attributes of good API are as per the following: [11]

• Simple to learn and utilize, hard to abuse.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

• Simple to peruse and keep up code that utilization it.
• It is customizing dialect nonpartisan.
• Adequately capable to help every single computational necessity.

3.1 IndoWordNet API

The IndoWordNet API gives a straightforward and simple approach to get to and control the WordNet asset free of the fundamental stockpiling innovation. The usefulness is uncovered through an arrangement of all around characterized objects that engineer can make and control according to his/her preparing necessity. The IndoWordNet API enables parallel access and updates to single or numerous dialect Word-Nets. Another outline utilizing social database has been actualized for this reason. This database outline (IndoWordNet database) bolsters stockpiling of numerous dialect WordNets. An exertion has been made to upgrade the outline to diminish repetition.

3.2 Functional Requirement

Clients regularly need to depend on others to perform capacities that he/she may not be capable or allowed to do without anyone else. Thus, practically all product needs to ask for other programming to do a few things for it. To achieve this, the soliciting program utilizes a set from institutionalized solicitations, called application programming interfaces that have been characterized for the program being called upon. Designer can make asks for by incorporating brings in the code of their applications. The language structure is depicted in the documentation of the application being called. By giving a way to asking for program benefits, an API is said to allow access to or open an application [1].

3.3 Architecture and Design

The IndoWord API has 2 layered structures. The upper layer is Application layer and lower layer is Data layer. The class graph of IndoWordNet API (Application layer) is demonstrated as follows. The Application layer uncovered the arrangement of classes and strategies which the designer will use to get to and control the WordNets. The Application layer does not straightforwardly get to the information put away on the plate however utilizes Data layer for this reason. The Data layer gives this administration through an arrangement of information classes and strategies which it opens to the Application layer. The Data layer comprehends the outline and capacity innovation used to store the information i.e. social database, level content records, filed documents, XML and so on. The Data layer is in charge of real access and control of information put away in documents/database and is relied upon to redesign the information in memory with the goal that it can be presented to the Application layer utilizing the Data objects. This shields the Application layer from changes away innovation or capacity outline.

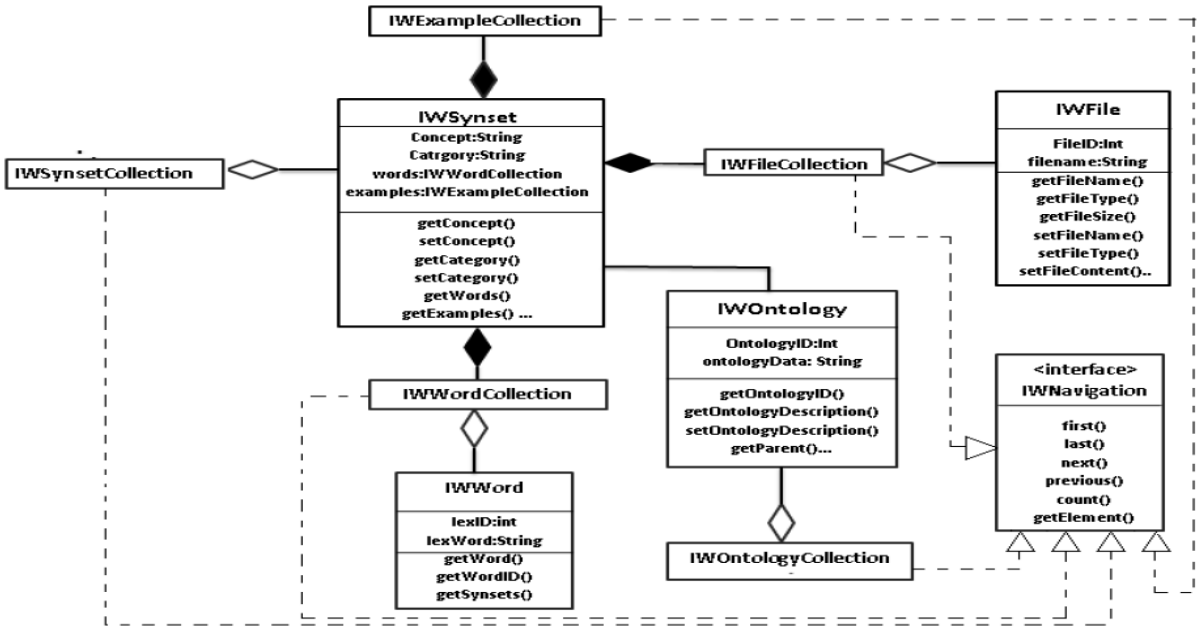

Figure 3. Simplified Class Diagram of IndoWordNet (Application Layer) [12]

Several significant classes of Application Layer are:

• ***IWAPI:*** A static class that permits initialising the IndoWordNet API library for utilize. To utilize the IndoWordNet API the principal thing you have to do is confirm the client i.e.

• ***IWLanguage:*** A class that gives association with dialect WordNets by utilizing IWLanguage langObj = IWAPI.getLanguageObject(IWLanguageConstants.KON KANI).

• ***IWSynset:*** A class that indicates a synset. It allows programmer to get/set the category, domain etc. and add/remove files and relations of synset.

• ***IWSynsetCollection:*** A class which displays set of synsets and allows programmer to get the collection size and iterate via collection.

• ***IWWord:*** a class shows word object and allows to get word Id and several lexical relations.

• ***IWWordCollection:*** A class represents words collection and developer is able to get the size and iterate the collection.

• ***IWFile:*** A class represents files and programmer can get/set the file content.

• ***IWOntology:*** represents ontology node, where each node in tree is mapped with synset.

• ***IWException***: A class explains the expectations of error or failure case.

4. CONCLUSION

Dialect is an essential part of human age. A few dialects are utilized as a part of India individual to their relating families. A lexicon can be utilized to get diverse significance of a word in various dialects. Different sorts of word references like reference book, thesaurus, phonetic lexicons and so forth are utilized. In this Paper, we've examined WordNets, its birthplace and a few instruments consolidated with it. IndoWordNet is a coordinated multilingual WordNet for Indian dialects. Moreover we've contributed insights about IndoWordNet Dictionary with the prerequisite and highlights of such word reference and existing instruments of IndoWordNet lexicon and the confinements of those apparatuses. We've overviewed past research done by experts. Besides, we've examined API's, IndoWordNet API, its need, Design and engineering of IndoWordNet API and different charges utilized as a part of it.

REFERENCES

1. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. International journal of lexicography, 3(4), 235-244.

2. Kunchukuttan, A., Puduppully, R., & Bhattacharyya, P. (2015). Brahmi-Net: A transliteration and script conversion system for languages of the Indian subcontinent. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations(pp. 81-85).

3. Promethee.philo.ulg.ac.be. (2018). [online] Available at: http:// promethee .philo. ulg.ac.be/ engdep1/ download/bacIII/wordnet_wikipedia.pdf [Accessed 7 May 2018].

4. Desai, S., Karmali, R., Naik, S., Walawalikar, S., & Ghanekar, D. Tools for IndoWordNet Development.

5. Redkar, H., Singh, S., Joshi, N., Ghosh, A., & Bhattacharyya, P. (2015). IndoWordNet Dictionary: An Online Multilingual Dictionary using IndoWordNet. In Proceedings of the 12th International Conference on Natural Language Processing(pp. 71-78).

6. Prabhu, V., Desai, S., Redkar, H., Prabhugaonkar, N. R., Nagvenkar, A., & Karmali, R. (2012). An effcient database design for IndoWordNet development using hybrid approach.

7. Bhingardive, S., & Bhattacharyya, P. (2017). Word sense disambiguation using IndoWordNet. In The WordNet in Indian Languages (pp. 243-260). Springer, Singapore.

8. Redkar, H., Joshi, N., Singh, S., Kulkarni, I., Kulkarni, M., & Bhattacharyya, P. (2016, January). Samāsa-Kartā: An Online Tool for Producing Compound Words using IndoWordNet. In 8th Global WordNet Conference.

9. Bhattacharyya, D. S. S. B. P. (2016, January). Detection of Compound Nouns and Light Verb Constructions using IndoWordNet. In Global WordNet Conference (p. 399).

10. Bhatt, B., & Bhattacharyya, P. (2011, December). IndoWordNet and its linking with ontology. In Proceedings of the 9th International Conference on Natural Language Processing (ICON-2011).

11. Bloch, J. (2006, October). How to design a good API and why it matters. In Companion to the 21st ACM SIGPLAN symposium on Object-oriented programming systems, languages, and applications (pp. 506-507). ACM.

12. Prabhugaonkar, N. R., Nagvenkar, A., & Karmali, R. (2012). IndoWordNet Application Programming Interfaces.