ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

Building Wordnet: a Survey

Laxmi Nadageri, Y.V. Haribhakta
College of Engineering, Pune

Abstract-Development of first Wordnet (Princeton Wordnet, aka PWN) was started in 1985 by the Psychology professor George A. Miller and his team in Cognitive Science Laboratory of Princeton University. The first wordnet has gained lot of attention from linguists all over the world. Wordnet has become popular due to its importance in Word Sense Disambiguation, Language learning, Natural Language processing, Machine learning etc. Most of languages have their own wordnet built using different methods. The available resources in respective languages such as bilingual and monolingual dictionaries, corpora etc. have a significant influence on the method used for creation of wordnet. This survey discusses and compares different methods of wordnet building such as Merge approach, Expansion approach, Common Base Concepts, methods based on word sense disambiguation, Intersection method, Assign procedure, etc. Objective of this survey is to discuss and find out efficient method to build wordnet for Pali language. Pali (language of Buddha) is still far away from being processed computationally. The digitalization of Pali language will encourage future research in it as there are many precious scriptures available in this language that hasn't processed yet.

Index terms- **Wordnet, Pali, Natural Language Processing, Base Concepts etc**.

I. INTRODUCTION

Wordnet [20] is a lexical knowledge base. Each entry of this knowledge base consists of group of semantically similar words (called Synset), gloss, and usage example. Each synset represents distinct concept. Different concepts are linked with each other based on their semantic relationship, like *hypernym* (generic relationship), *hyponym* (specific relationship), *meronym* (part-of relationship) etc.

For example, Consider the synset (animal, animate_being, beast, brute, creature, fauna), which refer to the concept *Animal*. Semantic relationships of concept *Animal* with other concepts are as follows.

Organism is a hypernym of *Animal.*

Dog is a hyponym of *Animal*.

Animal-tissue is a meronym of *Animal.*

Miller and his team started development of the first wordnet (aka, Princeton wordnet) in 1985 in cognitive science laboratory of Princeton University. Princeton wordnet (PWN) gained popularity because of its large coverage of applications such as Machine learning, Word Sense Disambiguation, Language translation, Language learning etc. Wordnet projects in other languages (started after PWN) used two main approaches of wordnet building: Merge approach and Expansion approach [30]. Subsequent section discusses these main approaches.

The rest of the paper is structured as follows. Section 2 describes Merge approach. In section 3, we discuss about Expansion approach and some methodologies used to create wordnet using less number of available resources. In section 4, we discuss proposed approach of wordnet development for Pali language. In section 5, we compare all methodologies discussed in this survey. In section 6, we discuss implementations of different wordnet projects in brief. Finally, section 7 concludes the paper.

II. BUILDING WORDNET: MERGE APPROACH

Merge approach is one of the two basic methodologies of wordnet development. In a nutshell, the process of Wordnet development using Merge approach contains collecting words, building synsets (i.e. synonyms of collected words) using monolingual dictionaries, corpus etc., interlinking of synsets based on their semantic relationship (discussed above) and later on, linking to already existing lexical resource, such as Princeton wordnet. The wordnet created by Merge approach is the most accurate than Expansion approach because it is less influenced by the design decisions of a wordnet for another language [27]. Wordnets created by merge approach follow different wordnet database structure than PWN (Hindi Wordnet has different database structure than PWN). The approach requires rich resources to build the wordnet, such as monolingual dictionaries with all possible senses of word, its POS, usage examples, gloss, some semantic structuring etc. However, Merge approach is very costly and a tedious approach. It is like building a wordnet from scratch. In spite of these disadvantages, many projects like Hindi wordnet, Urdu wordnet, Kannada wordnet etc. built using Merge approach.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

III. BUILDING WORDNET: EXPANSION APPROACH

The trend of imitating Princeton wordnet structure and translating English synsets into required language has become popular. This translation-based method is popular because, it is less time consuming and it permits alignment/linking of target language wordnet with Princeton wordnet. Wordnets built using same principle are useful in multilingual NLP tasks because they share some common attributes between them like, synset ID, wordnet database structure etc. However, language specific synsets are difficult to create using this method. Language-specific synsets may create problems, such as dangling nodes (isolated synsets without any relationship with other synsets), leaving gap synsets [26] (translation of some PWN synset could not be present in target language, can create gap in Wordnet hierarchy). Following section discusses some of the methodologies of expansion approach used to create wordnet.

a. Cross-Lingual Word Sense Disambiguation

Apidianaki [1] has described this method. There are two steps in this method: Word Sense Induction (WSI) and Word Sense Disambiguation (WSD). This methodology used to extend WOLF (French) wordnet, which was a sparse wordnet previously [2]. First step of this methodology creates clusters of semantically similar words and in the second step; clusters are placed at a proper place in wordnet database by disambiguating their sense.

Word Sense Induction

In this step, Apdianaki used English–Greek corpus and it has been word aligned to create the following form of lexicon.

English word Variation is represented by following three Greek equivalents (EQVs) in the corpus: (words in bracket represents transliteration of Greek word followed by English equivalent).

διακύμανση (diakýmansi, "fluctuation"),

μεταβολή (metavolí, "alteration"),

τροποποίηση (tropopoíisi, " modification")

Every EQV represents different sense of the given word. To distinguish between each sense and to place semantically similar EQVs in same cluster, Semantic Similarity of every pair is calculated.

Semantic Similarity Calculation: The sense of word can be determined with the help of context surrounding that word. The similarity between context features can reveal similarity between EQVs. Three weights (Local weight, Global weight, and Total weight) are assigned to each context feature based on their frequency of co-occurrences in order to estimate the similarity.

After assigning total weight (Total weight = Global weight X Local weight) to each feature, similarity between two EQVs, EQVm and EQVn is calculated using Weighted Jaccard (WJ) Coefficient.

$$WJ = \frac{\sum_{1}^{nbrj} \min(w(EQV_m, f_j), w(EQV_n, f_j))}{\sum_{1}^{nbrj} \max(w(EQV_m, f_j), w(EQV_n, f_j))} \qquad (1)$$

where, fj is jth feature. w (EQVm,fj) is a total weight received by fj for equivalent EQVm.

Greek EQVs of English word Variation are clustered based on their similarity coefficient (WJ) as below.

First cluster {διακύμανση} represents fluctuation sense.

Second cluster {μεταβολή, τροποποίηση} represents alteration sense.

Word Sense Disambiguation

The inventory created in the first step is used to find the sense of words in problem. In case, we need to find out sense of word *variation* in following instance:

"Various continents got affected because of Environmental pollution resulting in Global warming; some of the countries are experiencing in overall temperature, as well as significant variation in environmental season cycles".

As per the above discussion, context features found in English – Greek corpora have been assigned weights with respect to different EQVs. Suppose, below two clusters have been created in first step along with weighted context features. First cluster represents context features of fluctuation sense and second cluster represents context features of alteration sense.

Cluster1= {significant (2.04), range (0.76), individual (1.89), woman (1.89), increase (0.76), group (0.76), pressure (0.76), external (0.76), Ireland (0.76), year (1.49)}

Cluster2= {minor (2.25/1.83), human (2.01/1.13), number (0.73/1.16)}

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

In above instance of word *variation*, two words {increase, significant} out of surrounding context features found in cluster1 representing Greek word διακύμανση with sense fluctuation. Using this approach, Greek equivalents replaced PWN synsets to create Greek wordnet. Performance of Cross lingual WSD has found to be quite promising with approximately 72% nouns, 62% verbs, 81% adjectives, and 86% adverbs are correctly distinguished. Out of these correctly distinguished words, in second step, 64% noun clusters, 53% verb clusters, 75% adjective clusters, and 73% adverb clusters are correctly assigned to their respective synsets. Total number of empty synsets of WOLF wordnet filled by this method is 3904.

b. Google similarity Distance

Google Similarity Distance [9] is a method that uses Word Sense Disambiguation as a basis for linking target word to the English synsets. This method was used to create Macedonian wordnet [29]. To link appropriate target synset to PWN synset, this method finds similarity distance between translated synset and translated PWN definition of synset into target language.

Google Similarity Distance (GSD) uses a search engine, Google, to extract semantic similarity between words/phrases. The formula for normalized Google Similarity Distance between words/phrases x and y is:

$$GSD = \frac{max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - min\{\log f(x), \log f(y)\}} \qquad (2)$$

where f(x), f(y) are the result counts returned by Google for query containing x and query containing y, respectively, and f(x, y) denotes the result count for the query with both x and y included in it.

In below example, this method finds an appropriate Hindi synset to link to PWN synset (name, epithet);

PWN synset: name, epithet

Synset definition for "name, epithet": "a defamatory or abusive word or phrase"

Translation of above definition into Hindi (Translation field also contains Transliteration in bracket): "एक अपमान सूचक या अपमान जनक शब्द या वाक्यांश" (Ek Apmaan suchak ya Apmaan janak shabd ya vakyansh, translation is same as above point).

Translation of synset (name, epithet) into Hindi: अपमान (Apmaan,"Insult"), प्रतिष्ठा (Pratishtha,"Prestige"), नाम (Naam," Name"), संज्ञा (Sangya,"Noun")

GSD calculation of translated synset with translated definition (we have calculated GSD only for अपमान and प्रतिष्ठा):

GSD (अपमान, एक अपमान सूचक या अपमान जनक शब्द या वाक्यांश) = 0.493733

GSD (प्रतिष्ठा, एक अपमान सूचक या अपमान जनक शब्द या वाक्यांश) = 0.9051055.

Two words/phrases would be similar or related if their GSD comes near zero. Moreover, they are not related if GSD is near one. Hence, from above example,"अपमान" (Apmaan, "Insult") would be equivalent target language synset for PWN original synset (name, epithet) as it's GSD is less than GSD of प्रतिष्ठा (Pratishtha, "Prestige").

The result as per the discussion in [29] shows that Google Similarity Distance method has 87% accuracy in assignment of appropriate synsets. It correctly translates 14,335 English synsets into Macedonian synsets.

c. Intersection method

In Intersection method [29], synonymy enforces equivalence classes on word senses. Romanian wordnet [4] and Macedonian wordnet [29] were created using this method. There are two rules in this method.

First one is, if the original synset contains at least one monosemous word, translation of that single monosemous word is sufficient to translate other words in the synset. Second rule is, if the original synset contains more than one polysemous word, then the intersection of the translations of each word in the synset forms translation of original synset.

To explain second rule, consider PWN synset,

EnSyn = {w1, w2… wn},

where w1, w2… wn are the words in the synset, where all words are polysemous. Next, the translations of the words w1, w2… wn in target language are defined as;

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

T (w1) = {cw11, cw12… cw1m}
T (w2) = {cw21, cw22… cw1k}
…
T (wn) = {cwn1, cwn2… cwnj}
Where, cw11, cw12… cw1m are the translations of English word w1 into target language.

If translated word belongs to more than 65% of all translation sets, then it can be included in intersection set. This resultant set becomes equivalent synset in target language for PWN synset.

This methodology proved very successful against Gold standard during Romanian wordnet building. The result of automatic construction of Romanian wordnet showed 91% accuracy with 9,610 synsets and 11,969 relations among synsets. This method was combined with Google Similarity Distance method (explained in above subsection) to show more accuracy during Macedonian wordnet development process. The method translates extra 3,391 synsets, which were not translated by Google Similarity Distance method.

d. Multiple Heuristics method

Different Set of automatic WSD heuristic methods [18] are used to link target language synsets to Princeton wordnet. Score of six heuristics define eligibility of synset for linking to Princeton wordnet. Eligible synset with sufficient score called Candidate Synset.

Heuristic1 (Maximum Similarity)

To explain this heuristic, take an example; Target language word T has many translations in English i.e. T= {e1, e2, e3,…, en}. Suppose, e2 has most similar sense to the senses of e3, e4, e5; and other senses are different to each other. Number of similar senses to e2 is more than other words in translation set; hence, this heuristic provides maximum score to e2 than other translations.

Heuristic 2 (Prior Probability)

This heuristic provides Probability to every single translation using the formula: p=1/n, where n is the number of senses of a translation. Translation of Monosemous words gives only one sense. Hence this heuristic provides maximum score (i.e. 1/1=1) to monosemous translation.

Heuristic3 (Sense Ordering)

This heuristic assumes that, assigning most frequently occurring sense would result in Sense Disambiguation to be at least 75% correct. In case of most frequently occurring polysemous words, this heuristic gives 58% correct assignment of senses.

Heuristic 4 (IS-A relationship)

If two target words have IS-A relations, then their translations in English should also have an IS-A relation.

Heuristic 5 (Word Match)

Concepts that are related are expressed with same content words. This heuristic tries to find out the total number of shared content words in the definition of the target word given in dictionary and the content words in the gloss part of corresponding English synset.

Heuristic 6 (Co-occurrence)

This heuristic exploits co-occurrence measure of words collected from definition sentences of bilingual dictionary.

Combining heuristics with decision tree learning: Decision tree (DT) is provided with candidate synsets and 6 heuristics individual scores, it has been trained to link candidate synsets to English synset or to discard. The trained decision tree classifies new candidate synset as linking or discarding using 6 heuristics scores. Korean wordnet developed using this method.

The result contains 21,654 senses of 17,696 Korean nouns with approximately 94% accuracy. Every heuristic (H1 to H6) is evaluated separately and results are given in table [1] in the form of Precision (P) (Correctly linked Korean senses / All linked senses) and Coverage (C) (Linked senses / All senses).

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

Table 1: Individual Heuristics Score

	H1	H2	H3	H4	H5	H6	DT
P %	75.21	74.66	71.87	55.49	56.48	67.24	93.59
C %	59.51	100	100	29.36	63.01	64.14	77.12

e. Combining multiple methods

Homogenous Bilingual (HBil) dictionary is used in this method [3] to link target word synset to English wordnet. HBil is the combination of both directions (i.e. Spanish – English and English – Spanish) of bilingual dictionary. Different methods like Class methods, Structural methods, and Conceptual methods are used to link target language synset to English synset.

Class methods

In this method, Homogeneous dictionary is processed and following criterion used to form word groups.
Monosemic Criterion: 1:1criterion (1 English word has 1 target word translation), 1:N (1 English word has many target word translation, but many target words have only one English translation), and N:1(Same as 1:N but here one target word and many English words are mapped to each other). The target word belonging to any one of these criteria is linked to English word.
Polysemic Criterion: (Same as Monosemic criteria (discussed above) but applied to Polysemic words).
Hybrid Criterion: Variant Criterion: Small part of a given synset is a lexical variant. If more than one variant has same target word then a target word is linked to respective PWN synset.
Field Criterion: Some dictionaries contain field identifier for entries. In addition, English entry in wordnet is bearing the field identifier, if both occur in the same synset then target word is linked to English wordnet.

Structural methods

In this method, the whole PWN structure is used to link target word to PWN synset. From HBil, all combinations of English words for each target word entry are extracted. On the extracted information, following experiments are performed.
Intersection Criterion: This experiment extracts common synsets shared by all English words in PWN. Target word is linked to these common synsets.
Parent Criterion: If a synset of an English Word is a parent of rest of the synsets then the target word is linked to all hyponym of that English word.
Brother Criterion: If synsets have common parent then the target word is linked to all co-hyponym synsets.
Distant hyperonym Criterion: If a synset of an English Word is a distant hyperonym of the rest of the English Words then the target word is linked to the lower-level synsets.
The result of all these criteria follows following structure:
Target-word <list-of-EW> <list of synsets>

Conceptual Distance methods

Conceptual distance determines the closeness of meaning among the words. Conceptual Distance is calculated to find out closer words. Two words are co-occurring in a dictionary if they appear in the same definition. Such co-occurring words are collected from monolingual dictionary and conceptual distance is computed on such pairs.

Combining methods: Synsets having accuracy of more than 85% derived from above methods are selected and they are linked to PWN.
Spanish wordnet is built using combination of these methods. Results of these methods are quite encouraging. In Spanish wordnet v.0.0, all the synsets with Confidence Score (CS) more than 85% are selected and 10,982 connections are obtained. Combining discarded synsets having CS less than but near 85% that could be acceptable as new connections have increased the number of connections by 7,244. Finally, Spanish wordnet v. 0.1 with greater accuracy of 86.4% is obtained.

f. Assign Procedure

The Assign-procedure [26] uses bilingual dictionary to build the target language synsets corresponding to synsets already present in Princeton wordnet. The input to an assign procedure is one of the senses of bilingual dictionary and the output is a pair in the form of <PWN synset, confidence score>, Candidates (aka, CandSet) with greater

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

CS than some threshold value are considered for linking by lexicographer. For ordering based on CS, some of linking rules are considered:

Generic Probability

This rule provides probability to each candidate in CandSet to be a right candidate for linking to PWN synset.

Back translation

If we link target word to a correct English word, then this rule assumes that some of the synonyms of correct English word have to be translated into same target word.

Gloss matching

Gloss matching of target word in bilingual dictionary and PWN synset.

Synset Intersection

In case of ambiguous polysemous word, intersection of all synsets of all translations gives resultant synset to link to PWN synset.

The Assign Procedure was used by MultiWordnet project to create an Italian wordnet. This procedure was followed by Lexical–Gap procedure to add gap synsets (those synsets of PWN that are represented by free combination of words in Italian language instead of lexical units).

Below are the total statistics of the first version of Italian wordnet after applying both approaches. The Statistic shows number of synsets for each POS created in Italian wordnet**:**

No. of synsets for nouns: 20,571
No. of synsets for Verbs: 4,130
No. of synsets for Adjectives: 2,413
No. of synsets for Adverbs: 1,006
Total synsets: 28,120

g. Base Concepts

Wordnet created in EuroWordnet and Balkanet projects used this method. The Base concepts play crucial role in development of wordnets. According to [31], Base concepts are important (because they are widely used) concepts in hierarchy of PWN 1.5. Base concepts have many relations with other concepts i.e. they play the role of an anchor to attach other concepts in hierarchy.

There are three types of Base concepts:

Common Base Concepts (CBCs) are common concepts in at least two languages.
Local Base Concepts (LBCs) are Base concepts in single language.
Global Base Concepts (GBCs) are Common in all languages.

In EuroWordNet, 1,024 CBCs were selected (aka, first set of CBC) and defined as Princeton WordNet1.5 synsets. To select these initial CBCs EuroWordNet languages (English, Dutch, Spanish, and Italian) were used out of English, Dutch, German, French, Spanish, Italian, Czech, and Estonian.

The BalkaNet project has also followed same approach but it had applied on different set of languages: Greek, Romanian, Serbian, Turkish, and Bulgarian. BalkaNet extended the CBC set to 4,689 synsets.

Two phases of building wordnet using Base concepts: First is, developing a core wordnet using Common Base Concepts. Second is, extending the core wordnet top-down.

Core wordnet forms around 5,000 to 10,000 synsets. After extending core wordnet with Local base concepts, it can have more than 20,000 synsets. Table [2] shows overall statistics of both projects [35], [36].

Table 2: Overall Statistics of the Wordnet projects using Base concepts

Language	No. of Synsets	Words	Lang. internal relations
English (Added to PWN)	16361	40588	42140
Dutch	44015	70201	111639
Spanish	23370	50526	55163
Italian	48529	48499	117068
German	15132	20453	34818
French	22745	32809	49494
Czech	12824	19949	26259
Estonian	9317	13839	16318

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

Bulgarian	21441	44956	28599
Greek	18461	24366	24368
Romanian	19839	33690	25885
Turkish	14626	20310	19834
Serbian	8059	13295	12787

h. MultiDic tool

Indian Institute of Technology, Bombay (IITB) has developed a MultiDic tool [8], to facilitate a creation of Indian language wordnets based on Hindi wordnet. IITB created Hindi wordnet using first principle of wordnet development.

MultiDic tool has two main panels: left panel is Source Panel, which displays Hindi wordnet records. Right panel is editable target language panel. Both panels have same structure. Right panel allows linguist to enter target language data equivalent to Hindi data by referring left side Hindi record. Hindi records can be searched by entering Synset ID or word in left panel search box. The corresponding target language record gets stored in taget_lang.syn file and used for further processing. Below table shows statistics of Indowordnet [34].

Table 3: Indowordnet Statistics: (Language (Number of Synsets))

Kashmiri (29469)	Hindi (39991)	Odiya (35284)	Telugu (21091)
Assamese (14958)	Konkani (32370)	Marathi (32226)	Gujarati (35599)
Malayalam (30140)	Bengali (36346)	Sanskrit (33078)	Kannada (22042)
Bodo (15785)	Manipuri (16351)	Tamil (25419)	Nepali (11713)

IV. PROPOSED APPROACH FOR PALI WORDNET CREATION

Pali language or more precisely language of Buddha is still, far away from being processed computationally. It was the dialect of Magadha and was the language in which Gautam Buddha preached. Pali hosts the extant literature of Buddhism. The Buddhist scriptures are called Tripitaka (Sutta, Vinaya, and Abhidhamma). Tripitaka form a canon of the holy writings. Tripitakas are estimated to be about eleven times that of the Bible, and three times of the size of the Mahabharata. Pali literature being huge in size is of great importance; it is valuable alike to the philologists, the historians, and the students of folklore. Due to development of many popular regional languages, the use of Pali has decreased over time. Thus, there is a need of digitalization of the Pali literature for the use by common people. In the present scenario; Pali language has limited resources available in electronic form. The methodologies discussed above show efficient ways of wordnet building even if resources are low in numbers.

To develop initial base model for Pali wordnet, Common Base Concepts from Hindi language (as Hindi wordnet is one of the big wordnet available in India, all other Indian languages have followed structure of Hindi wordnet) can be extracted and translated into Pali using MultiDic tool of IITB.

Word Sense Disambiguation (WSD) is one of the most important applications of Wordnet. Pali Wordnet has been created to make WSD task easier. Domain field [33] for Pali synset has been added in each record. Domain field specifies particular area of the word. E.g. mouse word belongs to two domains (Computer_Science domain and Animal domain). Such domain fields serve extra information during WSD process.

Lex/ Extended Lex algorithm is widely used for WSD purpose. As per our experiments, Lex algorithm ends with more number of comparisons of wordnet synsets and input context words. On the other hand, Domain based WSD approach finish its task with very less number of comparisons; it has to find out domain of polysemous word and domains of surrounding context area. So the number of comparisons of domain based WSD approach would be the number of words present in the input sentence.

Different domains added in Pali wordnet:

Engineering, Medicine, Astrology, Astronomy, Literature, Linguistics, History, Arts, Fauna, Flora, Mathematics, Natural Objects, Group-of, Part-of etc.

V. COMPARISON OF VARIOUS METHODS

Table 4: Comparison

Method	Advantages	Disadvantages
Merge approach	More accurate synsets and semantic relationship. No influence of other	Time consuming approach

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

	languages.	
Expansion approach	Wordnet development becomes faster as gloss, synsets, and semantic relations are already available in source language.	Target language wordnet could be influenced by source language. And it won't reflect its language specific concepts.
Following methods are used to create wordnets using Expansion approach		
Word Sense Induction and Disambiguation	The sense inventory created in this method can be used for other research along with wordnet creation.	Restrictive cluster quality criteria can penalize interesting although noisy cluster due to single wrong word. This method requires large enough parallel corpus.
Google Similarity Distance	Alternative method to use in case of small corpus or lack of corpus.	Results of this method depend on the accuracy of the Google result counts. Google index changes rapidly over time.
Intersection method	This method has been proved as the most successful against the Gold standard during the construction of Romanian Wordnet.	If the original synset contains only one polysemous word or the translations of the words in the synsets have empty intersection set, then the method is not applicable for translating the synset.
Multiple Heuristic method	Results are better than Random mapping.	The result of individual heuristic is poor. Only maximum similarity heuristic gives better performance.
Combining multiple methods	These methods are highly suitable if there is conceptual similarity between English and target language.	It follows the same structure as PWN. Hence, most of the drawbacks present in PWN also apply to target wordnet.
Assign Method	This procedure speeds up the construction of synsets and detects the divergences between PWN and the target wordnet.	Though automatic procedure, lexicographer need to confirm right synsets every time. In worst case, this method gives only wrong candidate synset.
Wordnet creation Tool (Multidic Development tool by IITB)	Can add, delete, modify synsets in target language easily	Only Hindi records can be used as source
Common Base Concepts	Base concepts are independent to any language, hence faster and accurate translation of Base concepts	Concepts that are language specific need to be processed in different way in order to add them to wordnet.

VI. DIFFERENT WORDNET PROJECT STRATEGIES ON COMMON BASELINE

We have briefly discussed some of the wordnet development methodologies in previous section. Most of the wordnet (WN) projects have followed those methods with slight modification. Not all but some of wordnet projects are discussed in table [5] to study the different aspects of implementation of above given methodologies.

Table 5: Different wordnet projects discussion in brief

Common Base Concepts (CBC): *Advantages:* Concepts are independent, hence faster and accurate translation. *Disadvantages*: Difficult for language specific concepts. Below wordnets (Turkish, Arabic, Hungarian, Greek, and Serbian) are created using CBC method.	
Turkish WN [5]	First 1310 CBC were translated and automatically extracted synonyms, antonyms and hyponyms from machine readable dictionary. In second phase, 240 gap synsets (hypernym of the first phase synsets that are not members of first phase) & 1228 additional synsets with high frequency were selected.
Arabic WN [10]	Development of Arabic Wordnet (AWN) consists of manual construction of Base concept set from Eurowordnet and BalkaNet's Common Base concepts. Using this technique 1000 nominal and 500 verbal synsets have been obtained. In next phase, vertical extension of Base concepts has been obtained.
Hungarian WN [19]	Hungarian Wordnet follows same techniques of extracting base concepts and translating them. However, it has used VisDic [12] tool (which has been developed during EuroWordnet Project) for correction and editing the synsets.
Greek WN [6]	Greek wordnet used many tools to extract linguistic information from dictionaries and corpora. Greek wordnet development started with CBCs across all involved languages of Balkanet.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

Serbian WN [16]	Translation of base concepts was done manually as any form of electronic English/Serbian dictionary was not available at that time. Later on the validation of Serbian wordnet has been done with monolingual dictionaries, Serbian/French, Serbian/English corpora, morphological e-dictionaries.
MultiDic tool: *Advantages*: Addition, deletion, modification of synsets are easy with this tool. *Disadvantages*: Only Hindi records can be used as a source. Sanskrit, Punjabi, Tulu, and Konkani wordnets are created using MultiDic tool.	
Sanskrit WN [17]	Initially Sanskrit wordnet started creating synsets using MultiDic tool by referring Hindi synsets. Later on, lists of important Sanskrit words were acquired from University of Hyderabad, Sanskrit words from Ramayana, Mahabharata and from Naravane's Bhasha Vyavahar Kosh. And synsets for these concepts were created.
Punjabi WN [23]	Punjabi wordnet was created by expansion method using MultiDic tool of IITB. Punjabi synsets were created by referring Hindi synsets and by adding, deleting, modifying the Punjabi synsets as per the requirement.
Tulu WN	Tulu wordnet was discussed by Shivkumar B. in his paper *A Wordnet for Tulu* in 2009. He has used Hindi wordnet as base wordnet for Tulu. MultiDic tool has been used for synset entry. Concepts in Tulu has been entered which are equivalent to Hindi concepts and English too.
Konkani WN [32]	Hindi and Konkani are close languages. Maintaining identical concepts in Konkani language by referring Hindi Wordnet concepts was not much difficult. Hence to create Konkani Wordnet, MultiDic tool was used.
Merge Approach: *Advantages:* More accurate and no other language influence. *Disadvantages:* Time-consuming and tedious approach. Hindi, Assamese, Marathi, Kannada, and Czech wordnet have followed Merge approach.	
Hindi WN [7]	Hindi Wordnet was created at IITB using Merge approach. Synsets were created by looking into various resources of Hindi language. Later on it has been linked to English wordnet.
Assamese WN [13]	Wordnet Database has been created by entering data word by word using simple data entry interface (Web interface and offline interface). Data has been collected from online dictionary, Samartha Sabdakosh- An Assamese thesaurus, CIIL-EMILLE corpus and Assamese Pratidin corpus.
Marathi WN [22]	Marathi words are grouped together based on their similarity of meanings. Further, synsets are connected by considering their lexical and semantic relationship.
Kannada WN [28] / Gujarati WN [25]	Kannada wordnet was inspired by English wordnet and Hindi wordnet. To enter wordnet data, lexicographer's interface was created. Data was stored in RDBMS with several tables as Synset_table, noun_table, adjective_table, adverb_table, verb_table. Gujarati Wordnet has been developed by manually entering each word with its all details into database by lexicographers using data entry interface. MySQL database was used for storing synset and semantic relations using different tables.
CZech WN [24]	The first set of core synsets obtained from Dictionary of Literary Czech, DLC. Czech-English/English-Czech electronics dictionary was used later to compare resultant synsets with Princeton wordnet. In addition, synsets are manually checked and final version of core wordnet has been compiled. This cycle has been repeated 3 times. Most of the efforts were spent on sense discrimination.

VII. CONCLUSION

Availability of resources for Pali language is very less, we need to make a wise decision to utilize these resources and make use of it as much as possible. Creation of Pali wordnet will not only increase the number of Pali resources but also encourage future research in this language.

REFERENCES

[1]. Apidianaki M (2009) Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 77-85). Association for Computational Linguistics.

[2]. Apidianaki M, Sagot B (2012) Applying cross-lingual WSD to wordnet development. In LREC 2012-Eighth International Conference on Language Resources and Evaluation.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

[3]. Atserias J, Climent S, Farreres X, Rigau G, Rodr´ıguez H (2000) Combining multiple methods for the automatic construction of multilingual WordNets. Amesterdam Studies in the Theory and History of Lingustic Science Series 4, 327-340.

[4]. Barbu E (2007) Automatic Building of Wordnets EdUArd BarbU* &: Verginica BarbU MiTiTElU*** Graphitech Italy" Romanian Academy, Research Institute for Artificial Intelligence. Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005, 292, 217.

[5]. Bilgin, Orhan et al. (2004) Building a wordnet for Turkish. Romanian Journal of Information Science and Technology, 7(1-2), 163-172.

[6]. Bizzoni Y, Boschetti F et al (2014) The Making of Ancient Greek WordNet. In LREC (Vol. 2014, pp. 1140-1147).

[7]. Chakrabarti D, Narayan D, Pandey P, Bhattacharya P (2002) An experience in building the indo wordnet-a wordnet for hindi. In First International Conference on Global WordNet, Mysore, India.

[8]. Chatterjee A, Joshi S, Khapra M, Bhattacharyya P (2010) Introduction to Tools for IndoWordNet and Word Sense Disambiguation. In 3rd IndoWordNet workshop, International Conference on Natural Language Processing.

[9]. Cilibrasi R, Vit´anyi P (2007) The google similarity distance. IEEE Transactions on Knowledge & Data Engineering, 19(3), 370-383

[10]. Elkateb S, Black W et al (2006) Building a wordnet for Arabic. In Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006) (pp. 22-28).

[11]. Fellbaum C (2013) George A. Miller. Computational Linguistics March 2013, Vol. 39, No. 1, Pages: 1-3.

[12]. Horák A and Smrž P (2004) VisDic–wordnet browsing and editing tool. In Proceedings of the Second International WordNet Conference–GWC (pp. 136-141).

[13]. Hussain I, Saharia N, Sharma U (2011) Development of assamese wordnet. Machine Intelligence: Recent Advances, Narosa Publishing House, Editors. B. Nath, U. Sharma and DK Bhattacharyya, ISBN-978-81-8487-140-1.

[14]. Kaji H, Watanabe M (2006) Automatic Construction of Japanese wordnet. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).

[15]. Koeva S, Mihov S, Tinchev T (2004) Bulgarian Wordnet–Structure and Validation. Romanian Journal of Information Science and Technology, 7(1-2), 61-78.

[16]. Krstev C, Pavlovič G, Lažetič, Vitas D (2004) Using textual and lexical resources in developing Serbian wordnet. Romanian Journal of Information Science and Technology, 7(1-2), 147-161.

[17]. Kulkarni M, Dangarikar C, Kulkarni I, Nanda A, Bhattacharya P (2010) Introducing Sanskrit wordnet. In Proceedings on the 5th Global Wordnet Conference (GWC 2010), Narosa, Mumbai (pp. 287-294).

[18]. Lee C, Lee G, Yun S (2000) Automatic WordNet mapping using word sense disambiguation. In Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13 (pp. 142-147). Association for Computational Linguistics.

[19]. Miháltz M, Hatvani C, Kuti J, Szarvas G, Csirik J, Prószéky G, Váradi T (2008) Methods and results of the Hungarian WordNet project. In Proceedings of the Fourth Global WordNet Conference (GWC 2008) (pp. 311-320).

[20]. Miller G (1995) WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

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

[22]. Naik R, Mahender (2014) Marathi WordNet Development. International Journal of Engineering And Computer Science ISSN: 2319–7242, 3(8), 7622-7624.

[23]. Narang A, Sharma R, Kumar P (2013) Development of Punjabi WordNet. CSI transactions on ICT, 1(4), 349-354.

[24]. Pala K, Smrž P (2004) Building czech wordnet. Romanian Journal of Information Science and Technology, 7(2-3), 79-88.

[25]. Panchal S, Shukla P, Panchal P, Kolte J, Bharati (2015) Gujarati WordNet a [euro]" A Lexical Database. International Journal of Computer Applications, 116(20).

[26]. Pianta E, Bentivogli L, Girardi C (2002) Developing an aligned multilingual database. In Proc. 1st Int'l Conference on Global WordNet.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 6, Issue 4, July 2017

[27]. Prabhu, Venkatesh, Desai S, Redkar H, Prabhugaonkar N, Nagvenkar A, Karmali R (2012) An effcient database design for IndoWordNet development using hybrid approach.

[28]. Sahoo K, Vidyasagar E (2003) Kannada WordNet-A lexical database. In TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region (Vol. 4, pp. 1352-1356). IEEE.

[29]. Saveski M, Trajkovski I (2010) Automatic construction of wordnets by using machine translation and language modeling. In 13th Multiconference Information Society, Ljubljana, Slovenia.

[30]. Vossen P (2002) Version 3 Final July 1, 2002 Piek Vossen (ed.) University of Amsterdam.

[31]. Vossen P, Bloksma L, Rodriguez H, Climent S, Calzolari N, Roventini A, Bertagna F, Alonge A, Peters W (1998) The eurowordnet base concepts and top ontology. Deliverable D017 D, 34, D036.

[32]. Walawalikar S, Desai S et al (2010) Experiences in building the konkani wordnet using the expansion approach.

[33]. Magnini, Bernardo, and Gabriela Cavaglia. "Integrating Subject Field Codes into WordNet." LREC. 2000.

[34]. http://www.cfilt.iitb.ac.in/indowordnet/.

[35]. http://www.dblab.upatras.gr/balkanet/resources.htm.

[36]. Catalog.elra.info/product_info.php?products_id=549.