ISSN 2007-9737

Constructing Vietnamese WordNet: A Case Study

Khang Nhut Lam[1], Jugal Kalita[2]

[1] Can Tho University,
Vietnam

[2] University of Colorado,
USA

lnkhang@ctu.edu.vn, jkalita@uccs.edu

Abstract. WordNets are commonly used in tasks such as summarizing documents, extracting information, translating and creating other lexical resources. This paper presents experiments in constructing a Vietnamese WordNet (VWN) from a variety of freely published resources in several languages. The VWN has the same structure as the Princeton WordNet. Our algorithm translates several existing WordNets to Vietnamese using a freely available machine translator, removes translation ambiguities by applying ranking methods based on occurrence counts and Google distances on translation candidates. We also establish connections between synsets and extract glosses for synsets. Finally, we carefully look at the VWN created and identify problematic issues in the VWN due to differences in culture and agglutinative morphology of Vietnamese and other languages used.

Keywords. WordNet, Vietnamese, ontology construction.

1 Introduction

A WordNet is a large lexical database where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, the so-called synsets [17]. Each synset represents a distinct concept and consists of a unique synsetID, synset members, and a gloss consisting of a brief definition and one or more examples showing the use of members in the synsets. Synsets are connected to others by means of semantic relations such as hypernymy or generalization, hyponymy or particularization, and meronymy or part-whole relation. Currently, the biggest WordNet is the Princeton WordNet[1] (PWN) constructed manually since 1990. The PWN version 3.0 has 117,659 synsets including 82,415 noun synsets, 13,767 verb synsets, 18,156 adjective synsets and 3,621 adverb synsets.

In this paper, we discuss the feasibility of creating a Vietnamese WordNet (VWN) having the same structure as the PWN by bootstrapping from freely available resources. The remainder of this paper is organized as follows. In Section 2, we discuss related work. Section 3 describes the proposed approaches to build the VWN from existing resources. Results of our experiments and discussion are presented in Section 4. Section 5 concludes the paper.

2 Related Work

The research presented in this paper discusses an efficient method to generate a VWN with the same structure as the PWN. Therefore, this section highlights prior work on constructing WordNets based on the PWN. According to Vossen [25], the two common approaches to build a new WordNet in a target language *T* are the *expand* approach and the *merge* approach. Using the expand method, a new WordNet is created by simply translating the PWN to *T*, whereas using the merge method, an independent WordNet in *T* is firstly built and then aligned to the PWN. There have been a large number of efforts in various languages with the

[1] https://wordnet.princeton.edu/

goal of constructing WordNets. We present a few prominent ones in this section.

2.1 WordNets Created Using the Merge Approach

A French WordNet was constructed from multilingual resources by Sagot and Fiser [20]. The authors performed word alignment and extracted bilingual lexicons from a multilingual corpus; then, every lexical entry was assigned a synsetID obtained from the Balkan WordNet [23]. They also translated the English WordNet to French using dictionaries and thesauri. The French WordNet was finally generated by merging synsets collected from the two methods. Their WordNet contains 32,351 non-empty synsets, and its accuracy based on manual evaluation is 80%.

Gunawan and Saputra [7] generated a prototype version of synsets for an Indonesian WordNet from a monolingual dictionary of Bahasa Indonesia and an Indonesian thesaurus. They first extracted synonym concepts from the thesaurus, combined them with entries in the monolingual dictionary and removed duplicate entries. Finally, a hierarchical clustering technique was applied to merge synsets. Their Bahasa WordNet consists of 60,673 synsets. No evaluation was performed.

A Hindi WordNet[2] has been constructed manually by 'looking up the various list meanings of words in different dictionaries' [4]. The current version has 105,352 unique words and 40,457 synsets. The Hindi WordNet is the first WordNet for Indian languages and has been used to construct WordNets for other Indian languages (e.g., Marathi, Sanskirt and Gujarati) in the IndoWordNet project.

2.2 WordNets Created Using the Expand Approach

Oliver and Climent [18] compared the accuracies of WordNets created by several methods. The first WordNet was created using the Google translation machine to translate a sense-tagged corpus in English to Spanish. The generated WordNet had about 8,000 synsets with accuracy of 80%. In the second method, given a parallel corpus, an analyzer was used to tag senses of words with the English WordNet. Then, constructing a WordNet for Spanish became a word alignment problem. The accuracy of the second approach was lower than that of the first approach, and it depended on the size of the corpus. A bigger corpus increased the accuracy of the created WordNet. They also concluded that sense tagging introduced more errors than statistical machine translation.

Kaji and Watanable [9] constructed a Japanese WordNet by translating the PWN synsets to Japanese, by using a correlation matrix to deal with translation ambiguity. Later, Bond et al. [3] and Isahara et al. [8] constructed another Japanese WordNet by extracting synsets from the PWN and translating them to Japanese using bilingual dictionaries. They enriched the Japanese WordNet using the most common words obtained from different resources. This Japanese WordNet contained 57,238 synsets with 93,834 words.

Sathapornrungkij and Pluempitiwiriyawej [21] proposed a semi-automatic method to construct a Thai WordNet from machine readable dictionaries. They designed a WordNet Builder system which extracted lexical, semantic, and translation relations from the English WordNet and a dictionary. The extracted data was then evaluated according to 13 criteria (e.g., monosemic one-to-one, polysemic one-to-one and polysemic many-to-one). The created Thai WordNet contained 19,582 synsets with a coverage of 80% at 76% accuracy. Later, Akaraputthiporn et al. [1] and Leenoi et al. [14, 15] constructed Thai WordNets from several bilingual dictionaries using a bi-directional translation method. They noted that using different input dictionaries created by different methods such as corpora-based methods or author's expertise produced WordNets with different accuracies. In addition, cultural issues such as categorization, gender, and collective perception needed to be taken into account to maintain the structure of Thai data.

Saveski and Trajkovski [22] constructed a Macedonian WordNet using the expand approach. To remove irrelevant translations, the English synset gloss was translated into Macedonian, and then the Google similarity metric [5] was

[2] http://www.cfilt.iitb.ac.in/wordnet/webhwn/index.php

applied to compute the similarity scores showing the semantic relatedness between the translated gloss and the candidate words. The selected words were words with Google similarity distance with the translated gloss greater than a threshold. The Macedonian WordNet they created had 33,276 synsets.

Lam et al. [13] proposed several methods to create WordNets in many languages with limited resources. The authors generated WordNet synsets for a target language *T* by translating PWN synsets to *T* using the Microsoft Translator. The approach using direct translation (DR), the approach using intermediate WordNets (IW) and the approach using intermediate WordNets and a dictionary (IWND) were introduced to remove translation ambiguities. In the DR approach, synsets in the *T* WordNet were built by simply translating PWN synsets to *T*. The IW approach handled translation ambiguities by using different WordNets with the same structure as the PWN. For each synsetID in PWN, they extracted all synsets of intermediate WordNets and translated to *T*. The objects of their study included resource poor and endangered languages, which do not have many existing lexical resources. Hence, the IWND approach translated synsets having the same synsetID to English, and then translated them to *T*. The correct members of synsets were selected based on the occurrence counts of translation candidates. The authors claimed that the IW approach with 4 intermediate WordNets helped construct better WordNet synsets. They did not establish connections between synsets created.

WordNets created using the expand approach have the same structure as the PWN; however, their quality considering complex agglutinative morphology, presence of culture specific meanings and usages of words is not good compared to those of WordNets built using the merge approach. Generally, the expand approach is more widely used than the other.

3 Proposed Approaches

Generating a new WordNet for a language using the merge approach needs linguistic experts in the language. In addition, the VWN we want to create will have the same structure as the PWN. Therefore, the expand approach is the best choice to construct a VWN. Our work is based on the study of Lam et al. [13], and is divided into 3 parts: creating synsets, establishing connections among synsets and extracting glosses of synsets.

Table 1. Information about WordNets used

WordNet	Synsets	% coverage
FinnWordNet (FWN) [16]	116,763	100%
Japanese WordNet (JWN) [8]	57,184	95%
PWN	117,659	100%
Thai WordNet (TWN) [24]	73,350	81%
WOLF WordNet (WWN) [20]	59,091	92%

3.1 Creating Synsets

To create synsets for the VWN, we use the intermediate WordNets (IW) approach. Lam et al. [13] experimented using the IW approach with different numbers of intermediate WordNets, but they did not know how many intermediate WordNets are good enough to create a new WordNet of high quality. In addition to the WordNets used in their studies, we experiment with one more WordNet, the Thai WordNet. Table 1 presents information about WordNets used. All WordNets used are linked to the PWN version 3.0 and are obtained from the Open Multilingual WordNet [2].

First, we query synsetIDs of all synsets in the PWN. For each synsetID, we extract all members belonging to that particular synset in the PWN and other intermediate WordNets. Then, we translate all synset members in different languages to Vietnamese using a machine translator. As a result of this step, for every synsetID we have a list of translation candidates in Vietnamese. One drawback of the IW approach is that the coverage percentage of synsets created using the IW approach is lower than using the DR and IWND approaches.

To increase the coverage percentage of synsets in the VWN, we improve the method to select translation candidates. The ranking method based on occurrence count is still applied to calculate the

ranking value of translation candidates. The rank of a candidate w is calculated as below:

$$rank_w = \frac{occur_w}{numCandidates} \times \frac{numDstWordNets}{numWordNets}. \quad (1)$$

where:

— $numCandidates$ is the total number of translation candidates of members belonging to a synsetID.

— $occur_w$ is the occurrence count of the word w in the $numCandidates$.

— $numWordNets$ is the number of intermediate WordNets used.

— $numDstWordNets$ is the number of distinct intermediate WordNets that have members translated to the candidate w.

The rank value of each translation candidate is in the range from 0.000 to 1.000. The greater the rank value of the candidate, the higher the possibility that it will become a synset member. Lam et al. [13] select translation candidates based on 3 scenarios: (i) All candidates with the rank values of 1.000 are accepted as correct translations. (ii) If there is no candidate with rank values of 1.000, the candidates with the highest rank value are selected as correct translations. (iii) For each synsetID, if all candidates have the same rank value, they skip all these candidates.

Their approaches to select candidates for each synsetID significantly reduce translation ambiguities; however, an issue is that they discard many correct translations. For instance, members of the synsetID 110399491, with a gloss 'a father or mother; one who begets or one who gives birth to or nurtures and raises a child; a relative who plays the role of guardian', obtained from PWN and JWN are {parent} and {ペアレント}.

Translations of these members are {cha mẹ} and {phụ huynh}, respectively. The criteria for selecting candidates by Lam et al. discard these two candidates which are both correct translations. So, we change the selection method: if all translation candidates of a synset have the same rank value, we compute the Google distance between each translation candidate pair to find the semantic relation among candidates using the NGD formula [6]:

$$NGD(w_1, w_2) = \frac{\frac{max\{logf(w_1), logf(w_2)\} - logf(w_1, w_2)}{logM - min\{logf(w_1), logf(w_2)\}}}{0.7}. \quad (2)$$

where:

— M is the total number of pages indexed by Google[3], nearly 50,500,000,000 at the time we experiment.

— $f(w_1)$ and $f(w_2)$ are the numbers of pages containing w_1 and w_2, respectively.

— $f(w_1, w_2)$ denotes the number of pages containing both w_1 and w_2.

A pair of candidates is accepted as correct translations if the Google distance is smaller than a threshold α, which is 0.450 and is set by experiment. For example, the numbers of pages containing the words (cha mẹ), (phụ huynh) and (cha mẹ, phụ huynh) are respectively 655,000, 515,000 and 20,700. Applying the NGD formula, the NGD value of the pair (cha mẹ, phụ huynh) is 0.420. Therefore, we accept 'cha mẹ' and 'phụ huynh' as correct translations of synset members of synsetID 110399491 in the VWN.

3.2 Establishing Connections Among Synsets

Synsets in PWN are linked to others by semantic relations, which are of 28 types in the PWN version 3.0. There are 285,348 relations among synsets. Lam et al. [13] did not establish connections among the synsets created. We establish connection among synsets in the VWN based on relations among synsets in the PWN using Algorithm 1. First, each Vietnamese synset created $synset_{Vi}$ is mapped to a corresponding $synset_{Pj}$ in the PWN through a synsetID (lines 1-2). Then, for every $synset_{Pj}$ in the PWN, we extract all connections $semRelation_r$ between it and other synsets $synset_{Pk}$ (lines 3-4). Next, we check for the existence of $synset_{Vu}$, which corresponds to $synset_{Pk}$, in the VWN (lines 5-6). If there exists $synset_{Vu}$ in the VWN, we accept and establish the $semRelation_r$ between $synset_{Vi}$ and $synset_{Vu}$ in the VWN (lines 7-8).

[3] http://www.worldwidewebsize.com/

Algorithm 1 Establish connection among synsets in the VWN

Input: synsets in the VWN, synsets in the PWN and their sematic relations

Output: semantic relations among synsets in the VWN

for all $synset_{Vi}$ in the VWN created **do**
  $synset_{Pj} \leftarrow$ map ($synset_{Vi}$,PWN)
  for all $synset_{Pj}$ in the PWN **do**
    Extract all $semRelation_r$ ($synset_{Pj}$, $synset_{Pk}$)
    for all $semRelation_r$($synset_{Pj}$, $synset_{Pk}$) **do**
      $synset_{Vu} \leftarrow$ map ($synset_{Pk}$,VWN)
      if exist $synset_{Vu}$ **then**
        add $semRelation_r$($synset_{Vi}$, $synset_{Vu}$)
      end if
    end for
  end for
end for

Table 2 shows an example of establishing connections between synsetID 110399491 in the VWN with 2 synset members {cha mẹ, phụ huynh}. We note that we do not translate semantic relations to Vietnamese. Currently, the VWN constructed is managed based on the WNSQL project[4].

3.3 Extracting Glosses of Synsets From the Viet WNMS

The project called Viet WNMS[5] has constructed a Vietnamese WordNet for nouns, verbs and adjectives. This Viet WNMS project has been developed using the WNMS tool of the Asian WordNet project (AWN) [19] which provides a platform for building and sharing WordNets in Asian languages based on the PWN. The target of the Viet WNMS project is to build a Vietnamese WordNet consisting of 30,000 synsets and 50,000 words, including the 30,000 most common words in Vietnamese. The Viet WNMS project is divided into 2 parts[6]:

[4]http://wnsql.sourceforge.net/

[5]http://viet.wordnet.vn/wnms/

[6]http://wordnet.vn/vi/chi-tiet/tong-quan-ve-xay-dung-mang-tu-tieng-viet-18-1.html

Table 2. Example of synsets having connections to the synsetID 110399491 in the VWN

Synset ID	Synset member		Gloss	Semantic relation
	PWN	VWN		
107970406	family, family unit	gia đình, hộ gia đình	primary social group; parents and children	member meronym
109772448	adopter, adoptive parent	cha mẹ nuôi	a person who adopts a child of other parents as his or her own child	hyponym
110332385	female parent, mother	mẹ	a woman who has given birth to a child (also used as a term of address to your mother	hyponym
110126708	genitor	cha mẹ ruột	a natural father or mother	hypernym
110654932	stepparent	cha dượng	the spouse of your parent by a subsequent marriage	hyponym
109918248	kid, child	đứa trẻ	a human offspring (son or daughter) of any age	antonym

— Translating the core of the PWN to Vietnamese. According to authors, the core of the PWN are words with high occurrence counts obtained from the BNC corpus[7].

— Manually adding concepts that exist only in Vietnamese. Currently, the Viet WNMS has 40,788 synsets and 67,344 words.

The approach to create the VWN, discussed in this paper based on the IW approach in [13], takes advantages of lexicons in several WordNets having the same structure as the PWN. As a result, our VWN has a better synset coverage percentage and includes common words not only in English but also in several other languages such as French, Finnish, Japanese and Thai. Moreover, our VWN has 4 POSes, including adverbs, whereas the Viet WNMS has 3 POSes. To the best of our knowledge, there is no published paper on this Viet WNMS project. We do not know anything about the structure of this WordNet. However, by manually checking several synsetIDs, we understand that these synsetIDs or synsetOffsets in the Viet WNMS are not the same as in the PWN. Hence,

[7]http://www.natcorp.ox.ac.uk/

Algorithm 2 Extract glosses to synsets in the VWN

Input: the VWN and the Viet WNMS
Output: glosses of synsets in the VWN

1: **for all** words w in the VWN **do**
2: Extract all $synsets_{Ei}$ having w as a synset member from the Viet WNMS
3: $glossViet_i \leftarrow$ getGloss($synsets_{Ei}$)
4: Extract all $synsets_{Vj}$ having w as a synset member from the VWN
5: $glossTrans_j \leftarrow$ getGloss($synsets_{Vj}$)
6: Compute $CosineSim$ of each pair $glossViet_i$ and $glossTrans_j$
7: **if** ($CosineSim$ > β) AND ($CosineSim$ is the greatest) **then**
8: Accept $glossViet_i$ as a gloss of $synset_{Vj}$ in the VWN
9: **end if**
10: **end for**

the Viet WNMS is likely to have a different structure compared to the PWN and our VWN.

We notice that synsets in the Viet WNMS have glosses in Vietnamese, which we believe are constructed manually by experts. Therefore, we extract these glosses and add them to synsets in our VWN using Algorithm 2. We could not use synsetIDs or synsetOffsets to retrieve data from the Viet WNMS. Hence, for each word w in the VWN we created (line 1):

(i) We query all synsets, including their glosses (each of which is called $glossViet$), having w as a synset member in the Viet WNMS (lines 2-3).

(ii) We trace back to all synsets having w as a synset member and translate the corresponding glosses to Vietnamese using a machine translator, the so-called $glossTrans$ (lines 4-5).

Then, we compute a cosine similarity score between each pair of $glossTrans$ and $glossViet$ (line 6). If this score is greater than a threshold β, we accept the $glossViet$ as a correct gloss of that corresponding synset and add them to our VWN. For each $glossTrans$, if there are several $glossViet$s with cosine similarity scores greater than the threshold, we keep the one with the greatest cosine similarity score (lines 7-8).

4 Experiments and Discussion

4.1 Experiments

The synsets and the semantic relations among them in the VWN are evaluated by 8 volunteers who use Vietnamese as mother tongue. We use the same set of 300 synsetIDs, randomly chosen from the synsets we create, and connections among them. Each volunteer is requested to evaluate using a 5-point scale: 5: excellent, 4: good, 3: average, 2: fair and 1: bad.

The VWN is built by translating the PWN and several intermediate WordNets to Vietnamese. The quality of translations and quantity of synsets are highly dependent on machine translators used. Lam et al. [13] used the Microsoft Translator API for translation. When we performed experiments in 2017 for this paper, the Microsoft Translator API was not available for free, and therefore we use the Yandex Translate API[8].

We experimented by constructing VWNs using both our approaches, denoted by IW-NGD, and the IW approach [13] with 4 intermediate WordNets (PWN, FWN, WWN and JWN) and 5 intermediate WordNets (PWN, FWN, WWN, JWN and TWN) using the Yandex Translate API. Table 3 presents the number of synsets, their coverage percentages and average scores of the VWNs built. The VWNs generated using 5 intermediate WordNets have greater numbers of synsets and average scores.

Moreover, the IW-NGD approach creates VWNs of better quality in terms of the numbers of synsets and coverage percentages than the IW approach. The IW-NGD approach with 5 intermediate WordNets creates the best VWN in our experiment. So, we establish links among synsets in the best VWN created. There exist 80,413 semantic relations among 78,285 synsets created in the VWN. The average evaluation score of relations is 3.60.

The Viet WNMS has been published on a website but has limited web service capability. In addition, words in our VWN are not the same as words in the Viet WNMS. In particular, our VWN has many words which do not exist in the Viet WNMS; and contrarily, the Viet WNMS consists

[8] https://tech.yandex.com/translate/

Table 3. VWNs created using different approaches

Approach	Number of intermediate WordNets	Synsets	Average score	% coverage
IW	4	55,048	3.21	46.79%
IW	5	61,808	3.61	52.53%
IW-NGD	4	61,348	3.23	52.14%
IW-NGD	5	78,285	3.73	66.54%

of many words that do not exist in our VWN. Currently, we have queried 2,094 words from the Viet WNMS, and then extracted synsets' glosses for these words.

We carefully evaluate the glosses extracted and find that a value of 0.30 or higher for threshold β finds very good mapped glosses, with an average evaluation score of 4.60. Hence, such synset glosses (the ones extracted from the Viet WNMS) are accepted as the correct glosses and are aligned to the corresponding synsets in our VWN. We have extracted 4,555 glosses for synsets in our VWN. We believe that cooperation between the two Vietnamese WordNets is likely to produce a more extensive WordNet.

Table 4 presents some glosses extracted and aligned to the corresponding synsets in our VWN. In this table, *Member* means the synset member of the *SynsetID* in our VWN, *Gloss in the PWN*: the gloss of the *SynsetID* extracted from the PWN, *GlossTrans*: the translation of the *Gloss in the PWN* generated by a machine translator, *CosineSim*: the cosine similarity score between the *GlossTrans* and the *Gloss extracted* from the Viet WNMS.

4.2 Discussion

Lam et al. [13] and we create VWNs using the IW approach and the same 4 intermediate WordNets. The only different resource used in the prior published experiments and experiments reported in this paper is the machine translator. The previously reported VWN had 72,010 synsets (61.20% coverage percentage) with an average score of 4.26, which is higher than the VWN reported in this paper. The VWN created by Lam et al. [13] was evaluated by native Vietnamese speakers in the US whereas the VWN created in this paper has been evaluated by native Vietnamese speakers in Vietnam. We claim that the translation quality significantly affects the VWN created. Then, an initial important step to build a good WordNet is to use a very good machine translator or dictionaries for translation.

The VWN we created for this paper is managed using WNSQL with 18 tables. The main tables in our project are: linktypes, lexlinks, semlinks, senses, synsets and words. In addition, as mentioned earlier, the PWN has 28 types of semantic relations. We have established only 15 relation types among the synsets we created. One reason for limited connectivity is that many synsets do not exist in the VWN.

Constructing a VWN using the expand approach may lead to problematic issues regarding language gap as discussed below.

— The PWN has concepts which cannot be translated to Vietnamese. For instance, synsetID 107573347 with a gloss 'a canned meat made largely from pork' has one member {Spam} which does not translate well to Vietnamese, although it could possibly be translated to 'một dạng thịt heo đóng hộp' or 'đồ hộp Mỹ[9]'.

— Many concepts in Vietnamese do not exist in English. For example, synsetID 107804323 with a gloss 'grains used as food either unpolished or more often polished' has one member {rice}, which should be translated to 'gạo' in Vietnamese. To the best of our knowledge, in English, 'rice' can be also used for 'cooked rice' or 'boiled rice' which are both translated to 'cơm'. The PWN does not contain synsets pertaining to 'cooked rice' or 'boiled rice'. In Vietnamese, 'gạo' is different from 'cơm'. A similar issue is identified by Sathapornrungkij and Pluempitiwiriyawej [24] when building a Thai WordNet.

— Parts-of-speech (POS) of words in English and their translations in Vietnamese may not be similar. For instance, the word 'sad' in the PWN has only one POS of adjective. This

[9] https://vi.wiktionary.org/wiki/spam#Ti%E1%BA%BFng_Anh

Table 4. Examples of glosses extracted

SynsetId	Member	Gloss extracted	GlossTrans	Gloss in the PWN	Cosine Sim
100887081	sư phạm	nghề của một giáo viên	nghề của một giáo viên	the profession of a teacher	1.00
104161981	ghế	đồ đặc, được thiết kế để ngồi	đồ nội thất, được thiết kế để ngồi	furniture that is designed for sitting on	0.76
300230843	điều chỉnh	sửa đổi để chức năng tốt hơn	sửa đổi cho tốt hơn	modified for the better	0.68
113548105	lọc	loại bỏ các tạp chất	quá trình loại bỏ các tạp chất (như dầu hoặc kim loại hoặc đường)	the process of removing impurities (as from oil or metals or sugar etc.)	0.62
300128572	chưa từng có	không có ví dụ, tiền lệ hoặc sự tương tự trước đây	không có tiền lệ	having no precedent; novel	0.58
301711614	đau đớn	vô cùng đau khổ	thể hiện đau đớn hoặc đau đớn	expressing pain or agony	0.30

word is translated to 'buồn' in Vietnamese. In addition to the POS of adjective, the word 'buồn' has a POS of verb, meaning 'having strong need to do something'[10] and the PWN does not have this concept. Some examples showing the uses of the word 'buồn' are 'buồn ngủ' (sleepy or need to sleep) and 'buồn cười' (to feel like a laugh coming because of something funny (to need to laugh at something)).

5 Conclusion

The purpose of our work presented in this paper has been to study the feasibility of constructing a Vietnamese WordNet with as many synsets as possible by bootstrapping from free lexical resources. We have created synsets and established connections among them.

We intend to improve translation by changing the Yandex Translate API to another better freely available machine translator (if we can find one), and freely available dictionaries [11, 12].

We are contemplating several potential approaches to translate glosses of synsets in the PWN to Vietnamese or to extract glosses of synsets from a Vietnamese corpus. To improve translation quality between English and Vietnamese of glosses, we will use the approach proposed in [10].

In addition, finding a good method to mine or combine information from the Viet WNMS as we have done will definitely improve the quality of our VWN.

[10] https://en.wiktionary.org/wiki/bu%E1%BB%93n

References

1. **Akaraputthiporn, P., Kosawat, K., Aroonmanakun, W. (2009).** A bi-directional translation approach for building Thai WordNet. Asian Language Processing, 2009. IALP'09. International Conference on, IEEE, pp. 97–101.
2. **Bond, F., Foster, R. (2013).** Linking and extending an open multilingual WordNet. Proceedings of the 51st Annual Meeting, volume 1, pp. 1352–1362.
3. **Bond, F., Isahara, H., Kanzaki, K., Uchimoto, K. (2008).** Boot-strapping a WordNet using multiple existing WordNets. Proceedings of the 6th International conference on Language Resources and Evaluation, pp. 1–6.
4. **Chakrabarti, D., Sarma, V., Bhattacharyya, P. (2007).** Complex predicates in Indian language

WordNets. Lexical Resources and Evaluation Journal, Vol. 40, No. 3–4.

5. **Cilibrasi, R. L., Vitanyi, P. M. (2007).** The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 3.
6. **Evangelista, A., Kjos-Hanssen, B. (2009).** Google distance between words. Frontiers in Undergraduate Research.
7. **Gunawan, Saputra, A. (2010).** Building synsets for Indonesian WordNet with monolingual lexical resources. Asian Language Processing IALP 2010 International Conference on, IEEE, pp. 297–300.
8. **Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., Kanzaki, K. (2008).** Development of the Japanese WordNet. Proceedings of the 6th International Conference on Language Resources and Evaluation, pp. 2420–2423.
9. **Kaji, H., Watanabe, M. (2006).** Automatic construction of Japanese WordNet. Proceedings of the 5th International Conference on Language Resources and Evaluation.
10. **Lam, K. N., Al Tarouti, F., Kalita, J. (2015).** Phrase translation using a bilingual dictionary and n-gram data: A case study from Vietnamese to English. Proceedings of the 11th Workshop on Multiword Expressions, pp. 65–69.
11. **Lam, K. N., Al Tarouti, F., Kalita, J. K. (2015).** Automatically creating a large number of new bilingual dictionaries. Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2174–2180.
12. **Lam, K. N., Kalita, J. (2013).** Creating reverse bilingual dictionaries. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 524–528.
13. **Lam, K. N., Tarouti, F. A., Kalita, J. (2014aaro).** Automatically constructing WordNet synsets. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, volume 2, pp. 106–111.
14. **Leenoi, D., Supnithi, T., Aroonmanakun, W. (2008).** Building a gold standard for Thai WordNet. Proceeding of The International Conference on Asian Language Processing 2008 (IALP2008), COLIPS, pp. 78–82.
15. **Leenoi, D., Supnithi, T., Aroonmanakun, W. (2009).** Building Thai WordNet with a bi-directional translation method. Asian Language Processing. IALP'09. International Conference on, IEEE, pp. 48–52.
16. **Linden, K., Carlson, L. (2010).** Finnwordnet: Finnish WordNet by translation. LexicoNordica - Nordic Journal of Lexicography, Vol. 17, pp. 119–140.
17. **Miller, G. A. (1995).** WordNet: a lexical database for English. Communications of the ACM, Vol. 38, No. 11, pp. 39–41.
18. **Oliver, A., Climent, S. (2012).** Parallel corpora for WordNet construction: machine translation vs. automatic sense tagging. International Conference on Intelligent Text Processing and Computational Linguistics, Springer, pp. 110–121.
19. **Robkop, K., Thoongsup, S., Charoenporn, T., Sornlertlamvanich, V., Isahara, H. (2010).** Wnms: Connecting the distributed WordNnet in the case of Asian WordNet. Proceedings of the 5th Global WordNet Conference, Narosa Publishing.
20. **Sagot, B., Fiser, D. (2008).** Building a free French WordNet from multilingual resources. Proceedings of OntoLex.
21. **Sathapornrungkij, P., Pluempitiwiriyawej, C. (2005).** Construction of Thai WordNet lexical database from machine readable dictionaries. Proceedings of the 10th Machine Translation Summit, Phuket, Thailand, pp. 78–82.
22. **Saveski, M., Trajkovski, I. (2010).** Automatic construction of WordNets by using machine translation and language modeling. Proceedings of the 13th Multiconference Information Society, Ljubljana, Slovenia.
23. **Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., Grigoriadou, M. (2002).** Balkanet: A multilingual semantic network for the Balkan languages. Proceedings of the International WordNet Conference, Mysore, India, pp. 21–25.
24. **Thoongsup, S., Robkop, K., Mokarat, C., Sinthurahat, T., Charoenporn, T., Sornlertlamvanich, V., Isahara, H. (2009).** Thai WordNet construction. Proceedings of the 7th workshop on Asian language resources,

Association for Computational Linguistics, pp. 139–144.

25. **Vossen, P. (2005).** Building WordNets. Irion Technologies. Diaporama électronique.

Article received on 15/02/2018; accepted on 16/01/2020. Corresponding author is Khang Nhut Lam.