A Two-Phase Approach for Building Vietnamese WordNet

Phuong-Thai Nguyen
VNU University of Engineering and Technology
thainp@vnu.edu.vn

Van-Lam Pham
VASS Institute of Linguistics
lampv.il@vass.gov.vn

Hoang-An Nguyen
Naiscorp Inc.
annh@socbay.com

Huy-Hien Vu
VNU University of Engineering and Technology
hienvuhuy@vnu.edu.vn

Ngoc-Anh Tran
Le Quy Don Technical University
anhtn69@gmail.com

Thi-Thu-Ha Truong
VASS Institute of
Lexicography and Encyclopedia
hattt.viole@vass.gov.vn

Abstract

Wordnets play an important role not only in linguistics but also in natural language processing (NLP). This paper reports major results of a project which aims to construct a wordnet for Vietnamese language. We propose a two-phase approach to the construction of Vietnamese WordNet employing available language resources and ensuring Vietnamese specific linguistic and cultural characteristics. We also give statistical results and analyses to show characteristics of the wordnet.

1 Introduction

In order to solve various problems in NLP including information retrieval, machine translation, text classification, etc. we need language resources such as corpora and dictionaries. Wordnet is one of important resources for solving such problems. The first wordnet was created at Princeton University for English language. After that, diverse wordnets were constructed such as EuroWordNet for European languages, Asian WordNet for Asian languages, etc.

There are a number of important characteristics of the Vietnamese language that impact the construction of wordnet. Firstly, the smallest unit in the formation of Vietnamese words is the syllable. Words can have just one syllable, for example 'đẹp'$_{beautiful}$, or be a compound of two or more syllables, for example 'màu sắc'$_{color}$. As shown in Table 1, single-syllable words only cover a small proportion while two-syllable words account for the largest proportion of the whole vocabulary. Forming that vocabulary is a set of 7,729 syllables, higher than the number of single words. As in many other Asian languages such as Chinese, Japanese and Thai, there is no word delimiter in Vietnamese. The space is a syllable delimiter but not a word delimiter, so a Vietnamese sentence can often be segmented in many ways. Secondly, Vietnamese is an isolating language in which words do not change their forms according to their grammatical function in a sentence.

Length	Words	Percentage
1	6,303	15.69
2	28,416	70.72
3	2,259	5.62
4	2,784	6.93
5	419	1.04
Total	40,181	100

Table 1: Word length statistics from a popular Vietnamese dictionary, made by the Vietnam Lexicography Center (Vietlex).

Constructing wordnets is a complicated task. This task involves answering questions including which approach is appropriate, how to ensure specific characteristics of the language, how to take full advantage of available resources. This paper makes an attempt to answer these fundamental questions and reports major results of a project aiming to construct a wordnet for Vietnamese language, whose database includes 30,000 synonym sets and 50,000 words with 30,000 commonly used by the Vienamese.

Figure 1 represents major steps in construction

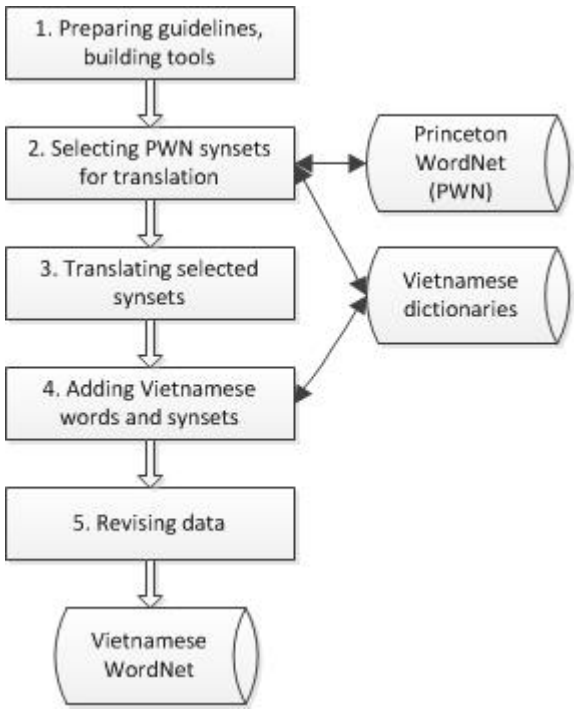

Figure 1: Steps in Vietnamese WordNet construction.

process of Vietnamese WordNet. We put these steps in two phases. Phase 1 involves steps 1-3, phase 2 involves steps 4 and 5. We exploit a number of language resources including Princeton's WordNet, a Vietnamese dictionary and an English-Vietnamese bilingual dictionary.

The class of adverbs in Vietnamese is a closed class (or a class of function words), while in English the class of adverbs is an open class (or a class of content words). Vietnamese adverbs express time (such as 'đã'$_{past}$, 'đang'$_{continuous}$), degree (such as 'rất'$_{very}$, 'hơi'$_{rather}$), and negation (such as 'không'$_{not}$). Therefore the number of adverbs in Vietnamese is much smaller than that in English. For that reason, there are only three parts of speech in Vietnamese WordNet including noun, verb, and adjective. Semantic relations in Vietnamese WordNet are similar to those in Princeton WordNet except a number of relations such as derivationally related form, participle of verb, etc.

The remaining part of this paper is organized as follows: Section 2 gives a review of several existing wordnets. Section 3 introduces our method to construct Vietnamese WordNet. Section 4 presents statistics and analyses of the wordnet being constructed. Section 5 gives a number of conclusions and future works.

2 Existing Wordnets

2.1 Princeton's WordNet

Since 1978, George Miller (Fellbaum, 1998) had researched and developed a database of words and semantic relations between words. This database was called wordnet and was considered a model of mental lexicon. Conceivably, wordnet is a large discrete graph in which nodes are synonym sets (synsets) and edges are semantic relations of synsets. A synset is a collection of synonym words of the same part of speech in which each word can be replaced by one of the others in certain contexts. For example, *car, auto, automobile, machine, motorcar* form a synset. This synset has a hyponymy relation with the synset *vehicle* because a *car* is a kind of *vehicle*.

2.2 EuroWordNet

EuroWordNet (Vossen, 2002) is a multilingual lexical database of nine European languages. Each language has its own wordnet. These component wordnets are linked via Princeton's WordNet version 1.5. More specifically, their synsets are linked to Princeton's WordNet's synsets which are equivalent or closest in meaning. EuroWordNet accepts different levels of lexicalization. For example, Princeton's WordNet contains both lexicalized and unlexicalized synsets, while Dutch WordNet contains only lexicalized ones. Component wordnets have been built by exploiting available resources such as monolingual dictionaries, bilingual dictionaries, and the Princeton's WordNet.

2.3 Asian WordNet

This project (Virach et al., 2009) aims to create wordnets for Asian languages such as Thai, Japanese, Korean, etc. Currently, there are data of 13 languages in Asian WordNet. The authors adopted a semi-automatic approach to translate Princeton's WordNet's synsets into Asian languages using bilingual dictionaries. The authors also built an online tool for editing and visualizing contents of the wordnet. By using this tool, many people can easily participate in the task of translation. They can also mod-

ify translations and can vote for the best one. In terms of wordnet design, Asian WordNet is a special case of EuroWordNet because it was built by translation approach. The major limitation of Asian WordNet is that it lacks specific concepts of Asian languages.

2.4 Laconec

This is a semantic-based multilingual dictionary available on the Internet[1]. According to the information on the website: This dictionary has been developed since 2007. The goal of Laconec is to provide multilingual lexical knowledge word lookup based on semantics. The core of the system is the large scale Princeton's Wordnet-like monolingual dictionaries linked to each other. This dictionary acknowledges Dr. Francis Bond's works (Bond and Paik, 2012) and four wordnets including English, Thai, Japanese, and Finnish.

3 A Method to Construct Vietnamese WordNet

3.1 Two Phases in Constructing Vietnamese WordNet

We construct Vietnamese Wordnet through two phases (Figure 1). In phase 1 (steps 1 to 3), we focus on translating a part of Princeton's WordNet into Vietnamese. In phase 2 (steps 4 and 5), we make use of Vietnamese resources to create the wordnet. Contents and requirements of these phases are different and separated.

The major work of phase 1 is translating a part of English Wordnet into Vietnamese. Thus, we firstly need to determine a list of English synsets to translate. Because of the significantly smaller size of our target Vietnamese wordnet, we choose to translate only a part of Princeton's WordNet. Our criteria for selecting English synsets include: (1) the lexicalization possibility in Vietnamese; (2) the connectivity of the selected part; (3) the inclusion of common base concepts.

Since the set of lexicalized concepts in English and the set of lexicalized concepts in Vietnamese are different, the data of wordnet built in phase 1 does not contain Vietnamese specific words such as *'âm dương'*$_{yin\ and\ yang}$, *'trắng ởn'*$_{white}$, *'làng xã'*$_{village}$, etc. or words relating to history, society and culture of Vietnamese such as *'truyện Kiều'*$_{a\ famous\ story\ in\ Vietnam}$, *'bánh chưng'*$_{a\ kind\ of\ cake}$, etc. Therefore in phase 2, we select coordinated compound words, reduplicative words, and subordinated compound words to add to the Vietnamese WordNet. We choose words from a popular Vietnamese dictionary, made by the Vietnam Lexicography Center (Vietlex).

3.2 Guideline Development

Editing data for wordnet is not an easy task, guideline documents are required to ensure the correctness and the consistency of data. In a wordnet, words are linked by semantic relations, therefore in the guideline document we focus on describing how to identify semantic relations especially synonymy, antonymy, hypernymy, hyponymy, holonymy, meronymy, and troponymy. We created diagnostic tests to verify relations between synsets. For instance, synonymy relation is identified on the basis of the possibility of a word being replaced by another in a specific context. This can be verified by the possibility of being mutually substitutable in sentence 'X is a *Noun*$_1$ therefore X is a *Noun*$_2$'. In addition to the tests there are a number of principles which can be used for encoding the relations, for example the Economy principle and the Compatibility principle (Fellbaum, 1998). Besides, we give guidelines as to handling Vietnamese specific linguistic and cultural characteristics. Last but not least, the guideline document contains instructions as to how to give definitions and examples, how to exploit resources such as existing dictionaries, and spelling rules.

3.3 Treatment of Vietnamese Specific Words

With regard to their structure, Vietnamese words can be divided into a number of types including single-syllable words, coordinated compound words, subordinated compound words, reduplicative words, and accidental compound words. The syllables which are not single words are bound morphemes[2], which can only be used as part of a word but not as a word on its own. The coordinated compound words (CCWs), specific to Vietnamese, are

[1] www.laconec.com

[2] They may have a meaning ('trường'$_{long}$, 'hàn'$_{cold}$) or not ('lẽo', 'nhánh')

words in which their parts– each part can be a word, single or compound words– are parallel in the sense that their meanings are similar and their order can be reversed. The meaning of a coordinated compound is often more abstract than the meanings of its parts. The proportion of this kind of words is about 10% of the number of compound words according to the statistics in the Vietlex dictionary. Reduplicative words (RWs) such as 'đất đai'$_{land}$, 'làm lụng'$_{work}$ are compounds whose parts have a phonetic relationship. This kind of words is specific to Vietnamese despite its small proportion. The identification of reduplicative words is normally deterministic and not ambiguous. Accidental compounds are non-syntactic compounds containing at least two meaningless syllables such as 'đười ươi'$_{orangutan}$, 'bù nhìn'$_{puppet}$. Subordinated compound words (SCWs) are the most problematic. A SCW can be considered as having two parts, a head and a modifier. Normally, the head goes first and then the modifiers. SCWs make up the largest proportion in the Vietnamese dictionary. Generally, discrimination between SCW and phrase is problematic because SCW's (syntactic) structure is similar to that of a phrase. This is a classical but persistent problem in Vietnamese linguistics.

The following are a number of synsets from Princeton's WordNet that were translated into Vietnamese. Words added to the synsets in phase 2 are in italics.

- (n) tree (a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown): cây; *cây cối, cây cỏ (CCW)*
- (v) laugh, express joy, express mirth (produce laughter): cười; *cười đùa (CCW), cười cợt (RW)*
- (adj) strong (having strength or power greater than average or expected): mạnh, mạnh mẽ, khoẻ; *khoẻ mạnh (CCW), khoẻ khoắn (RW)*
- (adj) black (being of the achromatic color of maximum darkness): đen, màu đen, có màu đen, mun, thâm, ô, ác, mực, huyền; *đen sì, đen sì sì, đen thui, đen trũi, đen nhẻm (SCW), đen đen (RW)*

POS	Synsets	Words	Word-synset pairs
Noun	17,084	32,122	37,452
Verb	9,483	21,180	32,273
Adjective	5,846	13,590	18,289
Total	32,413	66,892	88,014

Table 2: Vietnamese wordnet statistics.

3.4 Treatment of Vietnamese Proper Names

Proper names (place name, personal name, work name, etc.) represent important information about Vietnamese history, society, culture and thought. Vietnamese WordNet contains about 4,000 such linguistic expressions. Besides, Vietnamese WordNet has to also include worldwide famous names such as Amazon, Yangtze, Bacon, Nehru, etc. However, such names occupy only a small proportion in comparison with Vietnamese ones. The following are a few examples.

- 'nhân vật'$_{character}$ > 'nhân vật kịch'$_{drama\ character}$ > 'nhân vật chèo'$_{Vietnamese\ traditional\ operetta's/character}$ > 'hề'$_{clown}$/ 'mẹ Đốp'$_{mother\ Dop}$
- 'làng'$_{village}$ > 'Đường Lâm'$_{Duong\ Lam}$/ 'Mộ Trạch'$_{Mo\ Trach}$/ 'Hành Thiện'$_{Hanh\ Thien}$
- 'dân tộc'$_{ethnic\ group}$ > 'Kinh'$_{Kinh}$/ 'Tày'$_{Tay}$/ 'Thái'$_{Thai}$
- 'bánh'$_{cake}$ > 'bánh chưng'$_{square\ glutinous\ rice\ cake}$/ 'bánh trôi'$_{floating\ cake}$/ 'bánh rán'$_{fried\ cake}$
- 'hồ'$_{lake}$ > 'Hồ Gươm'$_{Sword\ Lake}$/ 'Hồ Tây'$_{West\ Lake}$

4 Empirical Analyses of Vietnamese WordNet

4.1 Vietnamese WordNet Statistics

Table 2 shows basic statistics of Vietnamese WordNet. Nouns take the largest proportion while the number of verbs and adjectives is smaller. Like Princeton's WordNet, Vietnamese WordNet can be considered as including three subwordnets corresponding to different parts of speech. The subwordnet of nouns has a unique root 'thực thể'$_{entity}$. The

subwordnet of verbs has 255 roots. The subwordnet of adjectives has 2,201 clusters.

As shown in Table 4, there are 61,509 semantic relations, in which 34,161 between noun synsets, 18,465 between verb synsets, and 8,883 between adjective synsets. The most frequent semantic relations include hypernymy-hyponymy, synonymy, antonymy, and similar-to. Vietnamese WordNet inherits the WordNet Domains Hierarchy (Bentivogli et al., 2004) including 164 domain labels organized as a tree structure.

4.2 Synset Size Distributions

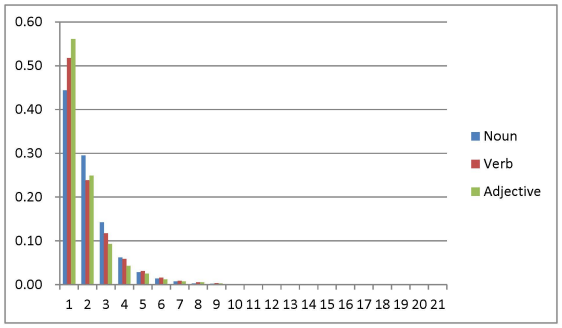

Figure 2: Synset size distributions.

Figure 2 shows synset size distributions of nouns, verbs, and adjectives. The horizontal axis represents synset size and the vertical axis represents the proportion. These distributions are not significantly different. On average each synset contains 2.42 words. When synset size increases, the corresponding proportion decreases.

4.3 Phase 2 Contributions

Table 3 represents word statistics in phase 2 of Vietnamese WordNet construction. The number of words added in this phase is 9,615. These words are specific to Vietnamese and different from words in phase 1. Besides, we also add nearly 4,000 proper nouns to Vietnamese WordNet. These nouns reflex Vietnamese anthronyms, toponyms (rivers, mountains, etc.), social events, etc.

POS	CCWs	RWs	SCWs
Noun	976	186	2,068
Verb	2,347	772	138
Adjective	1,406	1,217	505
Total	4,729	2,175	2,711

Table 3: Vietnamese WordNet statistics: phase 2.

Relation	Noun	Verb	Adjective
Antonymy	572	667	2,658
Hypernymy	15,240	8,661	
Hyponymy	15,240	8,661	
Holonymy	1,362		
Meronymy	1,362		
Entailment		307	
Cause		169	
Attribute	385		385
Similar to			5,840
Total	34,161	18,465	8,883
		61,509	

Table 4: Semantic relation statistics.

5 Conclusions

The paper has presented the most up-to-date results of the process of constructing Vietnamese WordNet. Since this project is coming to final stage, there can be slight differences between current version and the final version. We continue to revise data by lexical phenomenon or following statistical methods. Vietnamese WordNet will be published online and available for research and development purposes.

Acknowledgments

This paper has been supported by the national project number KC.01.20/11-15.

References

Luisa Bentivogli, Pamela Forner, Bernardo Magnini and Emanuele Pianta. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Proceedings of Workshop on Multilingual Linguistic Resources, COLING 2004*.

Francis Bond and Kyonghee Paik. 2012. A Survey of WordNets and Their Licenses. *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71.

Dhanon Leenoi, Thepchai Supnithi, Wirote Aroonmanakun. 2008. Building a Gold Standard for Thai WordNet. *Proceedings of IALP*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergrit Robkop, Chumpol Mokarat, and Hitoshi Isahara. 2009. Review on Development of Asian WordNet. *JAPIO 2009 Year Book*, Japan Patent Information Organization, Tokyo, Japan.

Piek Vossen. 2002. Wordnet, EuroWordnet and Global Wordnet. *Pub. linguistiques*, 2002/1 - Vol. VII, pages 27-38.

Piek Vossen. 2002. EuroWordNet General Document. *Online document*.