Automatic Acquisition of Noun Relations for Constructing Myanmar WordNet

Thandar Win, Hla Hla Htay
University of Computer Studies, Yangon
shelbygt.shelby01@gmail.com, hlahlahtay123@gmail.com

Abstract

This paper describes the acquisition of noun relations for constructing Myanmar WordNet. WordNet is a useful lexical resource where specific senses of words are clustered together into synonymy sets, and semantic relationships between the sets are specified. WordNet is used in various NLP research, such as Information Extraction, Information Retrieval and in most other NLP application. The system has three steps. First, extract the lexico semantic relations by using LexicoSyntactic Pattern method. Second, by using information theoretic notion of mutual information, the new coming word has to be estimated to identify and in which the existing word of the association with sense. Third, refine the sense of noun word by manual. We have collected the noun word list (8943 words) and 55 patterns. We have obtained 87.9 % accuracy in sense identification. The system shows noun relationship between not only word level but also sense number. The system is implemented using Java.

1. Introduction

In recent years, a number of WordNet building efforts have been initiated and carried out within a common framework for lexical representation and are becoming increasingly important resources for a wide range Natural Language Processing application. The first WordNet was developed for English language (Princeton WordNet). WordNets have since been created for many other languages, and currently there are WordNets for over 40 different languages in some shape or form. Notable efforts include EuroWordNet, BalkaNet, Asian WordNet and the establishment of the Global WordNet Association. In this paper, we describe automatic acquisition of noun relations for building WordNet for Myanmar language.

The biggest challenge in constructing a WordNet, is in identifying the words that are semantically related to one another. WordNet can be constructed by manual or automatic-semiautomatic construction. Manual construction is a cumbersome, labor-intensive, expensive and time-consuming task, whereas the automatic-semiautomatic construction is more manageable, less labor-intensive, less expensive and faster. The lack of financial and other resources, it may not be very practical to build lexical resources manually in the conventional ways. We need innovative ways to create such resources which can make use of computational power. There is a clear need for automatic constructing of WordNet.

English WordNet has semantic relations. They are noun to noun, verb to verb, adjective to adjective relations. Our system will focus for finding noun relations from Myanmar Corpus. The system has two corpora. One is pattern corpus and the other is raw corpus. Raw corpus is any kind of text. We studied the nature of the Myanmar text. Then, shows the semantic relations of words are gathered to built a pattern corpus. We have collected the noun word list (8943 words) which are obtained from Myanmar Orthography [12] and are added the noun word that not in the Myanmar Orthography. We present the LexicoSyntactic Pattern method [7] extract the lexico semantic relations from Myanmar corpus. We have manually collected 55 patterns. Next, by using information theoretic notion of mutual information [11] the new coming word has to be estimated to identify and in which the existing word of the association with sense. In this paper, we acquisitate Myanmar noun word relation and identify the word sense number.

2. Related Work

WordNets are built manually or semi-automatic or automatically. The first ever WordNet is Princeton WordNet (PWN) [5] and it is for English language.

WordNet built by manual are Euro WordNet (EWN) [14] for a multilingual database with

WordNets, Arabic WordNet [15] which can be linked directly PWN 2.0 and (EWN), Indonesian WordNet [3], based on PWN and Hindi WordNet [4] linked with EWN and PWN and Telugu WordNet [18] based on Hindi WordNet.

WordNets built by semi-automatic are Thai WordNet [16] by using the expand approach and a bilingual dictionary, align the PWN synsets, English-Russian WordNet [17] by using expand approach and mapping of PWN to RWN (Russian WordNet).

WordNets built by automatic are French WordNet [1] by combining multilingual resources, Korean WordNet [10] by using word sense disambiguation techniques and bilingual dictionary, Japanese WordNet [9] by using unsupervised word-sense disambiguation and bilingual comparable corpora, Persian WordNet [13] by using Persian and English corpora and bilingual dictionary, Slovene WordNet [20] by using expand model, Seriban WordNet and bilingual dictionary, based on PWN.

3. Extraction of Myanmar Noun Relations

The system flow is discussed in below. See in figure 1.

3.1. Corpora Preprocessing

Myanmar language, contain words and has mainly 9 part of speech: noun, pronoun, verb, adjective, adverb, particle, conjunction, post-positional marker and interjection. Myanmar text is written without natural delimiters. Myanmar corpus is therefore needed to segment into words. Sentences for Myanmar corpus are collected from internet in Zawgyi font script. Some sentences are gathered from Myanmar text book. First, Myanmar corpora (raw corpus and pattern corpus) are segmented into sentences based on Myanmar sentence boundary marker ပုဒ်ဖ (။). Second, these sentences are splitted into words with ThaiLucene [19] based on noun word list (8943 words). These words frequencies are counted for calculating the mutual information [Section 4].

3.2. Semantic Relations for Noun of Myanmar Language

These noun word can be designed to capture and describe word sense, inter-connected them through a variety of lexical and semantic relations. Noun

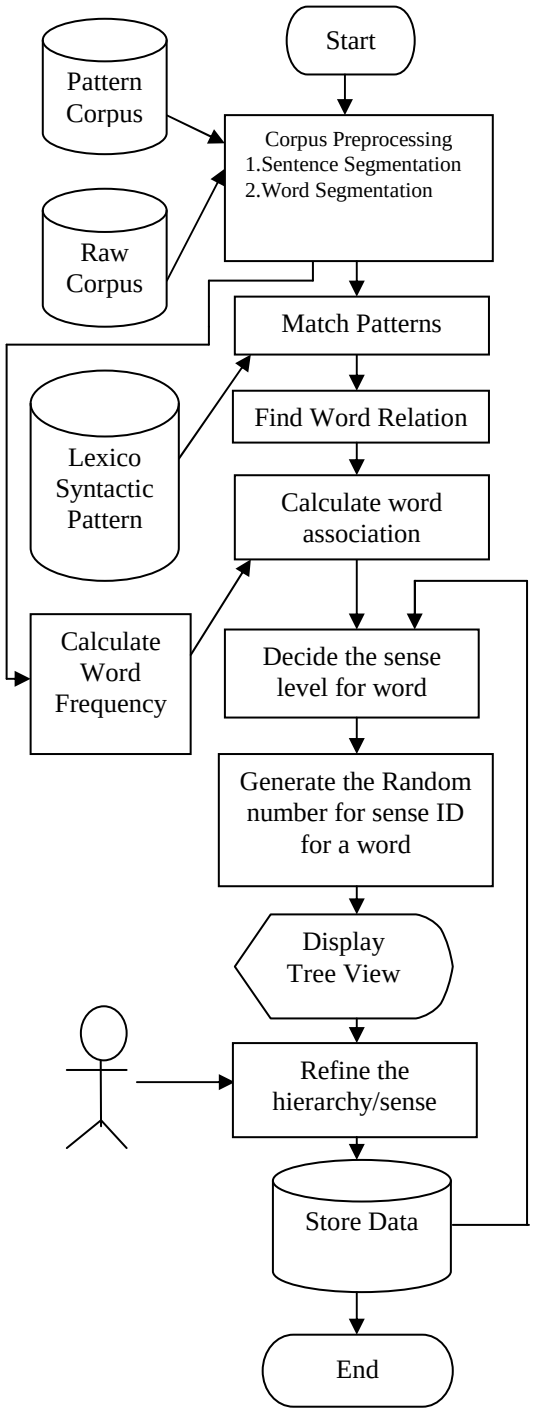


Figure 1. System Flow Diagram

relations are classified into six categories in English WordNet [6]: Hypernymy- Hyponymy, which the relation of subordination (or class inclusion of subsumption), which in this context we will call hyponymy. X is a hyponymy of Y if X is a kind of Y. Hypernymy is inverse of Hyponymy. For example, the noun 'bird' is a hyponymy (subordinate) of the noun 'animal', or, conversely, 'bird' is a hypernymy (superordinate) of 'bird'. Meronymy-Holonymy, the part-whole relation between nouns is generally considered to be a semantic relation, called meronymy. This relation also has an inverse: if W_m is a meronymy of W_h, then W_h is said to be a holonymy of W_m. The conventional test phrases are is a part of or has a. If W_m is a part of W_h is acceptable, then W_m is a meronymy of W_h; if W_h has a W_m (as a part) is

acceptable, then W_h is a holonymy of W_m. For example, the noun 'branch' is a holonymy of 'tree', or, conversely, 'tree' is a meronymy of 'branch'. The two relation of the structure is hierarchy and sense level. And, Synonymy is same meaning of the word and Antonymy is opposite meaning of the word which of the design is word level. These relations are extracted from pattern corpus by using LexicoSyntactic Pattern approach [Section 3.3].

3.3. LexicoSyntactic Pattern Approach

A method for the automatic discovery of lexico semantic relations by searching for corresponding lexicosyntactic patterns in large text collections. The sentence like *Gelidium* is a kind of *red algae*. Therefore, *Gelidium* is a hyponymy of *red algae*. In Myanmar language, to extract lexico semantic relations, we define a set of patterns which manually collect from book, newspaper, journal and website.

Table 1. Example of pattern in English

1	Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.
2	a. NP_0 such as NP_1 {, NP_2…, (and/or)NP_i}i >= 1 b. for all NPi, i >= 1, Hyponymy (NPi , NP0)
3	Hyponymy (Gelidium , red algae)

Table 2. Patterns for Extracted Relations

	Pattern	Relation
1	w_1၊w_2၊….၊/(နှင့်)w_n စသော w	Hypernymy
2	w_1၊w_2၊…. ၊/(နှင့်) w_n အစရှိသော w	Hypernymy
3	w_1၊w_2၊… ၊/(နှင့်) w_n စသည့် w	Hypernymy[1]
4	W ၏ w_1	Meronymy
5	w_1 ၊ w_2 ၊ …… w_n စုပေါင်း၍ w	Meronymy
6	w ကို w_1 ဟုခေါ်သည်။	Synonymy
7	w(w_1)	Synonymy

For extracted relation, the patterns are seen in Table 2. Hyponymy is inverse of hypernymy relation and holonymy is also inverse of meronymy relation. Manually collect 26 patterns in hypernymy relation and 14 patterns in meronymy and 15 patterns in synonymy relations. The semantic relations of noun word are shown in figure 2.

1. (၊) (ပုဒ်မ) which is same as ",", comma in English.
2. w_1,w_2,….,w_n and w means the noun word of Myanmar language.

Table 3. Example of Lexico Semantic Relations

1	ငှက်၊ကျား၊ဆင် စသော တိရစ္ဆာန် များကိုတွေ့ရသည်။ w_1၊w_2၊w_3 စသော w w_1,w_2 and w_3 is hyponymy and w is hypernym relation.
2	ကျောက်ရည်ပူ ကို ချော်ရည် ဟုခေါ်သည်။ W ကို w_1 ဟုခေါ်သည်။ W and w_1 is a synonymy relation.
3	ငှက်တော ၏ အမွေး သည်မည်းနက်၏။ W ၏ w_1 W is meronymy and w_1 is holonymy relation.

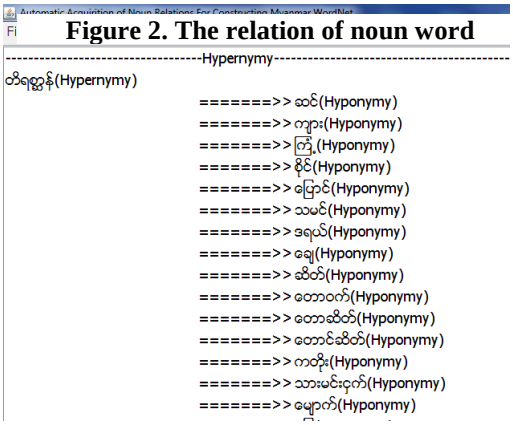


Figure 2. The relation of noun word

4. Identifying the different meanings of the word

The words which will be appeared as the semantic relations are assumed that the target word calculates with the associate word. One word has one or more than meanings. The words are same but different meanings. Example, the ဂေါရခါး word has 2 senses (ဂေါရခါး #n#1 and ဂေါရခါး #n#2). For the first meaning (ဂေါရခါး #n#1) will highly associated with (လူမျိုး #n#1-hypernymy). The mutual information score of two words greater than 3, we assume that they are highly associated [11]. For the second meaning (ဂေါရခါး #n#2) will highly associated with (သီးနှံ #n#1 - hypernymy). In automatic determining the different senses of a word, the information theoretic concept of mutual information [11] is used. Mutual information, I(x,y), is defined as

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)\,P(y)}$$

f(x) = frequency of x word in raw corpus
f(y) = frequency of y word in raw corpus

N = no of noun word in raw corpus

If I(x,y) is greater than 3, the pairs of word are related. (N =10432)

Calculate for first ဂေါရခါး word relates လူမျိုး word,

P(x), f(x) / N = 25/10432 = 0.002

P(y), f(y) / N = 15/10432 = 0.001

P(x,y), f(x,y) / N = 8/10432 = 0.001

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

$$= \log_2 \frac{0.001}{(0.002)(0.001)}$$

$$= 8.97$$

Calculate for second ဂေါရခါး word relates သီးနှံ,

P(x), f(x) / N = 14/10432= 0.001

P(y), f(y) / N = 15/10432= 0.001

P(x,y), f(x,y) / N = 5/10432= 0.0004

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

$$= \log_2 \frac{0.0004}{(0.001)(0.001)}$$

$$= 8.64$$

Table 4. Calculate the word association result

x	f(x)	y	f(y)	f(x,y)	I(x,y)
လူမျိုး	25	ဂေါရခါး	15	8	8.97
သီးနှံ	14	ဂေါရခါး	15	5	8.64

The first ဂေါရခါး word and လူမျိုး word of mutual information is 8.97. Therefore, this is interested relation (လူမျိုး #n#1- ဂေါရခါး #n#1). And, the second ဂေါရခါး word and သီးနှံ word of mutual information is 8.64, (သီးနှံ #n#1- ဂေါရခါး #n#2). (n=Noun). In this way, different meaning of a word can be identified. The result of word associations are described in figure 3.

4.1. Storing the concept/sense hierarchy

Noun relations are stored in the database. In the database,

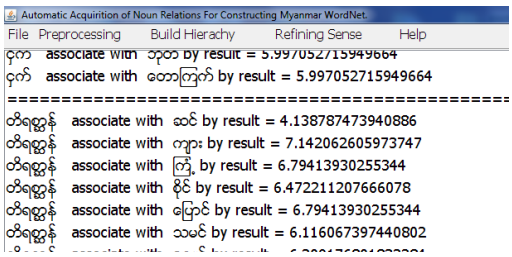
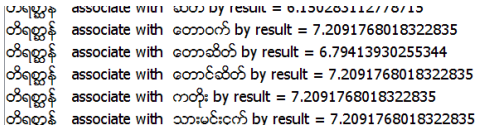


Figure 3. The result of word association



(1) Nouns are organized into synsets (synonym sets), which are arranged into a set of lexical semantic relations by using the Lexico Syntactic Pattern [7]. (e.g. တိရစ္ဆာန်@ ဆင်~၊ကျား~).

(2) Different meaning of a word using the theoretic notion of mutual information [11]. The database lists Myanmar synsets. A each entry consists of random number, relation pointers, sense no and the word. The relevant relations and encodes them in the synset as relational pointers. Synonymy of word is implicit by inclusion in the same synset. Other relations are represented by either semantic (between synsets) or lexical (between individual word forms) pointers.

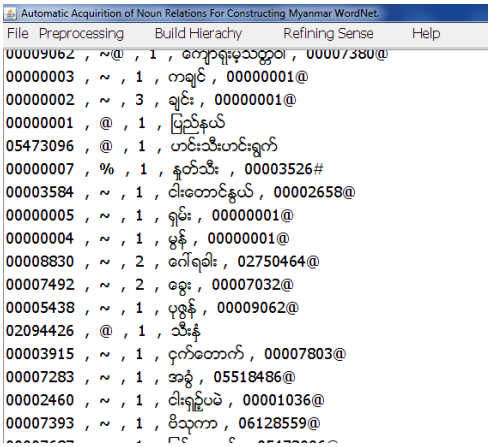
Table 5. Noun Relations of pointer symbols

Noun	
Antonymy	!
Hyponymy	~
Hypernymy	@
Meronymy	#
Holonymy	%

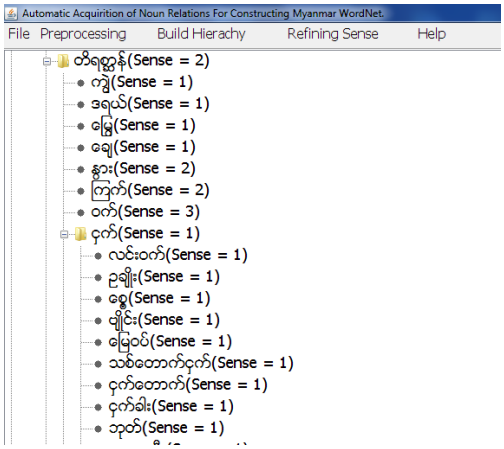
The concept/sense hierarchy data are stored in database as shown in figure 4.

Figure 4. Loading the stored data

5. Manual refining the concept hierarchy



It is necessary to refine the word of concept hierarchy. The word (တိရစ္ဆာန်#n#1) has only one sense (meaning). However, there may have different hierarchies. They have semantic relation but currently our corpus is small in size. Therefore, the pair of words cannot occur frequently in our corpus. Therefore, the word hierarchies are along to correct



manually. The noun word concept hierarchies are described as tree view. See in figure 5.

6. Limitations and Further Extension

The noun word which has two or more syllables in length is considered for automatic defining the meaning (sense). Because of, one syllable noun word is more ambiguous for defining automatic meaning (sense). Antonymy, attribute noun relations are not yet included in this thesis because pattern based sentences of these relations are difficult to find. In this system, Verb, Adjective and Adverb relations are not yet included. Searching for the information of individual word is not yet included in this system. And then, definition of sense and frequency of sense are not yet included in this system. These limitations can be extended. The pattern corpus and normal corpus sizes also can be extended.

7. Evaluation

In this system, we calculate the accuracy by using the following equation.

$$\text{Precision} = \frac{\text{number of correctly identified senses}}{\text{number of all senses reported by the system}}$$

Table 6. The evaluation of noun relations

Relations	System Reported senses	Number of correct senses
hypernymy	194	155
hyponymy	683	610
meronymy	40	33
holonymy	84	76
synonymy	52	52
Total number of senses	1053	926

The Table 6 shows the evaluation of hierarchy obtained for the 300 pattern based sentences after the system has automatically defined sense number. We have obtained 87.9% accuracy in sense identification.

8. Conclusion

In this paper, we have described the methods for building noun relations for Myanmar WordNet. Construction WordNet is a kind of creating lexical resource. Creating lexical resources is important for building Natural Language Processing (NLP) applications such as Lexicography, Information Extraction, Information Retrieval, Machine Translation, Knowledge Engineering etc.

The system has three steps. First, the noun relations are extracted from running text using lexicosyntactic pattern method. 55 patterns are collected for semantic relations and 8943 noun words are gathered for segmenting the corpus into words. Second, identify different senses of each noun word by using theoretic notion of mutual information. Third, we have also implemented manual refining of the sense if desired. We have built the hierarchical structure with the 300 sentences sized pattern corpus. 1053 noun senses are put in hierarchy. The depth of hierarchy is 5.

References

[1] Benoit Sagot et.al, Building a free French WordNet from multilingual resources, University of Paris, France.

[2] Caraballo, S.A, Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text, Brown University Ph.D. Thesis, 2001

[3] Desmond Darma Patra et.al, Building an Indonesian WordNet, Proceedings of the 2nd International MALINDO Workshop, Faculty of Computer Science University of Indonesia.

[4] Dipak Narayan et.al, A WordNet for Hindi, First International Conference on Global WordNet, Mysore, India, January, 2002.

[5] Fellbaum, C. , WordNet: An Electronic Lexical Database, Cambridge, MA: MIT Press, 1998.

[6] George A.Miller, Nouns in WordNet, Princeton University.

[7] Hearst, M. & Suchtze, H. Customizing a lexicon to better suit a computational task. Proceedings of the ACL SIGLEX Workshop on Acquisition of lexical knowledge from Text, 1993.

[8] Hearst, M., Automatic Acquisition of Hyponyms from Large Text Corpora. Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, 1992.

[9] Hiroyuki Kaji et.al, Automatic Construction of Japanese WordNet, Department of Computer Science, Shizuoka University, Japan.

[10] Juho Lee et.al, A Korean Noun Semantic Hierarchy (WordNet) Construction, Department of EECS, Korea.

[11] Kenneth Ward Church, Patrick Hanks, Word Association Norms, Mutual Information, and Lexicography to identify word of sense level, Scotland, Collins Publishers, pg. 1-3.

[12] Myanmar Orthography, http://myanmar words. pikay. org/ feeds/ posts/ default.

[13] Mortaza Montazery, Automatic Persian WordNet Construction, School of Electrical and Computer Engineering College Engineering, Tehran University.

[14] Piek Vossen, Euro WordNet (EWN), Published in: The ELRA Newsletter, Vol. 3 n.1, ISSN: 1026-8300. Paris. Pg. 7-10, February 1998.

[15] Sabri Elkateb et.al, Building a WordNet for Arabic, Manchester University, UK, 2006.

[16] Sareewan Thoongrup et.al, Thai WordNet Construction, Proceedings of the 7th Workshop on Asian language Resources, Semi-automatic Compilation of Asian WordNet, Proceedings of the 14th NLP 2008, 24 May 2010.

[17] Sergey Yablonsky et.al, English-Russian WordNet, Proceedings of the third international WordNet conference, Seagwipo KAL Hotel, Sucth Jeju Island, Korea, January 22-26, 2006.

[18] S. Arulmozi, Telugu WordNet, Department of Dravidian & Computational Linguistics, Dravidian University, Kuppam 517425, India.

[19] ThaiLucene, Word Tokenizer, http://rcmuir. Wordpress.com, 2009.

[20] Tomaz Erjavec, Building Slovene WordNet, Proceedings of the 5th International Conference on language Resources and Evaluations, 22-28 May, 2006, Genoa, Italy.

[21] WordNet 2.1, stored the data, Princeton University, released in March 2005.