Institute of Advanced Engineering and Science

IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 2, No. 3, September 2013, pp. 107~116
ISSN: 2252-8938

Unknown Word Detection via Syntax Analyzer

Soe Lai Phyue*, Aye Thida
* University of Computer Studies, Mandalay

Article Info

Article history:

Received Nov 15, 2012
Revised Feb 10, 2013
Accepted Feb 16,2013

Keyword:

Natural Language Processing,
Myanmar Language Lexico-conceptual Knowledge Resource,
WordNet,
Syntax Analyzer,
Context Free Grammar,
Name Entity Recognition

ABSTRACT

A knowledge resource is the central repository of data for all Natural Language Processing (NLP) applications and development of NLP applications mostly depend on coverage of knowledge resources. The multipurpose Myanmar Language Lexico-conceptual Knowledge Resource (ML2KR) and Myanmar function tagged corpus were developed as initial resources by using semiautomatic approach. ML2KR consists of Myanmar WordNet, Myanmar English bilingual computational lexicon and morphological processor. Myanmar language is morphologically rich and agglutinative language. Therefore, it is usually required to segment Myanmar texts prior to further processing. Segmentation has two main problems, word ambiguity that more than one meaning and unknown word occurrence that a word does not have in the lexicon. In this paper, we address on the unknown word occurrence issue. To detect the new unrestricted character patterns of words, character based parsing syntax analyzer is built by using Context Free Grammar (CFG). Firstly, unknown words are considered as a Name by Name Entity Recognition with forward and backward rule based approach. If the name does not agree with syntax analyzer, all possible unknown words are verified to update the lexicon and Myanmar WordNet. The output of syntax analyzer for correct sentence is added to create the function tagged corpus. Our function tagged corpus is very useful in Myanmar to English machine translation system.

Corresponding Author:

Aye Thida,
Departement of Research and Development,
University of Computer Studies, Mandalay.
Mandalay Division, Myanmar.
Email: ayethida.royal@gmail.com

1. INTRODUCTION

Nowadays, the development of context of NLP in general is rapidly growing as computational linguistic field. Grammars and words belong to the province of linguistics, but the concepts they express belong to the extra- linguistic knowledge base about the word [1]. For each language, the lexicon must provide the links that enable a language processor to carry messages from one province to the other. The demands on the lexicon also vary with the type of application in NLP. Each application can also be processed at levels of detail ranging from a rough approximation triggered by keywords to a deep understanding that applies all the resources of syntax, semantics and pragmatics. The language resources like lexicon is a bridge between a language and the knowledge base expressed in that language.

For this reason, ML2KR is already developed for various NLP applications. In the previous work, MLR is constructed by using semiautomatic methodology by acquiring the lexical and conceptual knowledge from WordNet and Myanmar<->English Machine Readable Dictionaries (MRDs) [2]. To build the MLRs, the translation links are collected from existing bilingual MRDs and semantic meaning and synset links are collected from English WordNet. The collected links and their meaning are manually verified. The computational lexicon stores the word according to their part of speech. Beside then, this work needs to deal the inflectional cases for MLRs to improve the coverage by using the rule-and-feature based model of Myanmar inflectional morphology [3]. Through a detailed study of the Myanmar language, we developed an analyzer that incorporates many of the unique features and challenges present in Myanmar. This resource is used in Machine Translation (MT) system.

Almost all MT approaches use Part Of Speech (POS) tagging and parsing as preliminary step [4]. Since Myanmar sentences are strings of characters with no delimiters to mark word boundaries, segmenting Myanmar texts becomes an essential task for Myanmar language processing. Besides word segmentation, we also need to identify the part-of-speech (POS) tags of the words. During the process of word segmentation, two main problems occur: segmentation ambiguities and unknown word occurrences. For the unknown word problem, we need to detect and identify them from the input text and verify them to add the lexicon. New word identifying and verifying to update lexicon is also helpful for increment of our knowledge resources [5].

Statistical methods for extracting Myanmar unknown words usually suffer a problem that superfluous character strings with strong statistical associations are extracted as well. To solve this problem, this paper proposes to use a set of general morphological rules to broaden the coverage and on the other hand, the rules are appended with different linguistic. To reduce the complexity of the rule matching, early parsing algorithm for extraction is proposed, which merges possible morphemes recursively by consulting above the general rules.

An efficient bottom-up merging algorithm by consulting the general rules to extract unknown words and using priority measures to resolve the rule matching ambiguities for Chinese word is proposed in [6]. They compare effects of different priority strategies. It is found that the performance of unknown word detection would affect the entire performance significantly. Although the performance of unknown word detection is not bad, there is still room for improvement.

A character-based chunking for unknown word identification in Japanese and Chinese text is introduced in [7, 8]. The method is built upon SVM-based chunking, by use of character *n*-gram and surrounding context of *n*-best word segmentation candidates from statistical morphological analysis as features.

In proposed unknown word identification and detection system consider with segmentation and part of speech tagging problem. The detected word is needed to register in lexicon. First, a morphological analysis is done to obtain initial segmentation and POS tags and then a parser is used to detect unknown words.

The paper is organized into 7 sections. In the next section, we provide an overview of our system. Section 3 briefly introduces the existing structure of ML2KR and Section 4 provides the syntax analyzer for Myanmar sentence. Unknown word detection as name need to makes some analysis for helping the derivation of general rules and verification of unknown words to update MLRs need to automatic sensation network for Myanmar WordNet. Therefore, we derive a set of general rules to represent all kinds of unknown words, and then modify it by appending rules and taking relation of word from WordNet are described in Section 5. In section 6, the evaluation of extraction is presented. Finally, in section 7, we make the conclusion and propose some future works.

2. OVERVIEW OF THE SYSTEM

For unknown words with more regular morphological structures, such as personal names, morphological rules are commonly used for improving the performance by restricting the structures of extracted words [9]. However, it's not possible to list morphological rules for all kinds of unknown words, especially those words with very irregular structures, which have the characteristics of variable lengths and flexible morphological structures, such as proper names, abbreviations etc [9].

Identifying Myanmar unknown words from a sentence is difficult; since (i) there is no blank to mark word boundaries; (ii) almost all Myanmar characters and words are morphemes; (iii) Morphemes are syntactic ambiguous and semantic ambiguous; (iv) words with same morpho-syntactic structure might have different syntactic categories.

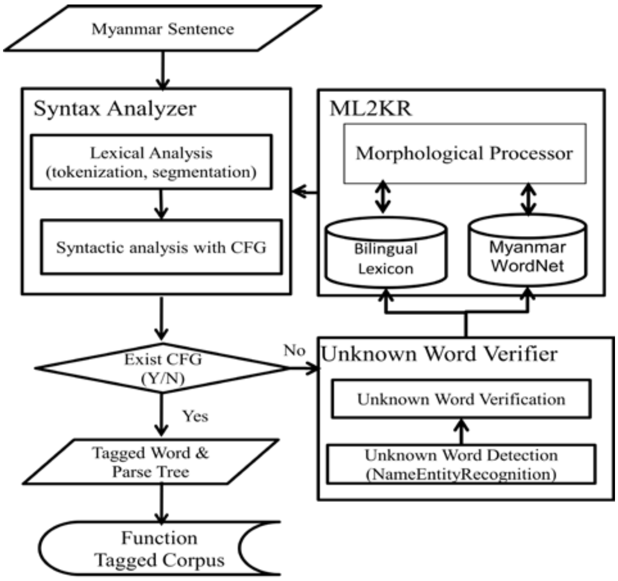

Figure 1. Overview of the New Word Detection and Identification for Improvement of Language Resources

A framework of the new word detection and identification system for improvement of language resources illustrates in Figure 1. Initially, the input sentence is segmented and analysis by using the lexical analyzer proposed in [10]. In this system, each unknown word in the sentence will be segmented into several adjacent tokens (known words or monosyllabic morphemes) and generate the possible pattern. Afterward, Name Entity Recognition algorithm applies to identify name and tag again with syntax analyzer. At unknown word detection stage, every monosyllable is decided whether it is a word or an unknown word morpheme by a syntax analyzer. If analyzed tree is not generated, all possible combined unknown word is extracted and verifying. The verified words are update to the lexicon to improve the coverage of lexicon. Finally, the input text is re-segmented using name word and updated lexicon. The syntax tree is stored in corpus for further NLP application. This system may improve the coverage of both the MLRs and Myanmar function tagged corpus. The example input sentence of "သူတို့ ဒဂုံစင်တာ သို့ သွားသည်", immediate result of each process and output tagged tree shown in table 1.

Table 1. Example of Unknown Word Tagger via Syntax Analyzer

input sentence	သူတို့ ဒဂုံစင်တာ သို့ သွားသည်
tokenize word	သူ + တို့ + ဒ + ဂုံ + စင် + တာ + သို့ + သွား + သည်
possible Word List	သူ< PRON_PERSON,>
	သူတို့< PRON_PERSON,>
	တို့<>
	ဒ<>
	ဂုံ< >
	စင် <noun.artifact,noun.artifact,noun.artifact,noun.artifact,noun.artifact, noun.artifact,noun.artifact,noun.artifact,noun.artifact,>
	တာ< >
	သို့< PREP_ARRIVAL,>
	သွား <noun.body,noun.artifact,noun.artifact,verb.motion,verb.motion,verb.motion,verb.stative,verb.social,verb.motion,verb.motion,verb.motion,verb.motion,verb.stative,verb.social,verb.motion,verb.communication,verb.change,verb.consumption,verb.stativ

	e,verb.stative,verb.stative,verb.stative,verb.stative,verb.motion,verb.contact,verb.contact,verb.consumption,verb.consumption,verb.competition,verb.cognition,verb.change,verb.change,verb.change,verb.change,verb.change,verb.change,verb.motion,verb.motion,verb.perception,verb.social,verb.stative,verb.stative,verb.stative,verb.stative,verb.stative,verb.stative,verb.stative,verb.stative,PARTICLE_VERB,PARTICLE_VERB,>
	သွားသည် <verb.stative,verb.stative,verb.stative,verb.stative,verb.stative, verb.stative,verb.stative,verb.stative,verb.social,verb.perception,verb.motion,verb.motion,verb.change,verb.change,verb.change,verb.change,verb.change,verb.change,verb.cognition,verb.competition,verb.communication,verb.motion,verb.social,verb.stative,verb.motion,verb.motion,verb.motion,verb.motion,verb.social,verb.stative,verb.motion,verb.motion,verb.motion,verb.change,verb.consumption,verb.stative,verb.stative,verb.stative,verb.stative,verb.stative,verb.motion,verb.contact,verb.contact,verb.consumption,verb.consumption,>
	သည်< PREP_VERB,PREP_NOM,>
Pre-segmented sentence	သူတို့ <PRON_PERSON >+ ဒ< > + ဂုံ < > + စင် < noun.artifact> + တော< > + သို့ < PREP_ARRIVAL,> +သွားသည်< verb.stative>
Tagged Segmented sentence	သူတို့ <PRON_PERSON >+ ဒဂုံစင်တော< UnKnown > + သို့< PREP_ARRIVAL,> +သွားသည်< verb.stative>
Function Tagged Tree	0:[S] 0:[NP_NOM] 0:[NP] 0:[PRON_PERSON] > 0:[သူတို့] 0:[PREP_NOM] > 0:[သည်] 0:[NP_ARRIVAL] 0:[NAME_P] 0:[NAME_PLACE] > 0:[ဒဂုံစင်တော] 0:[PREP_ARRIVAL] > 0:[သို့] 0:[VEND] 0:[V] > 0:[သွားသည်]

3. MYANMAR LANGUAGE LEXICO-CONCEPTUAL KNOWLEDGE RESOURCES (ML2KR)

ML2KR is developed by the semiautomatic approach using the existing language resource of WordNet and Myanmar-English Machine Readable Dictionaries (MRDs) [2]. ML2KR consists of three main resources (Myanmar WordNet, bilingual computational lexicon and Morphocon) as shown in Figure 2. The components of the ML2KR are separated component, so it can be sense as multifunctional. Thus, it has been designed to be potentially reused in many NLP tasks (e.g. Information Retrieval (IR) and extraction (IE), machine translation, dialogue-based systems, etc). These resources are formed several independent but interrelated modules.

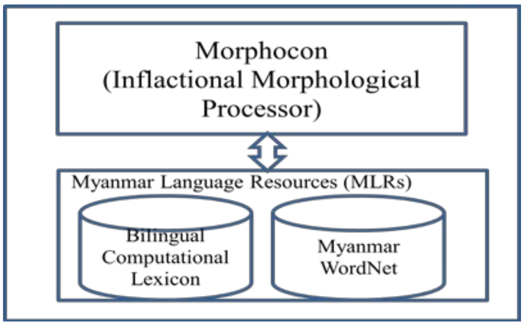

Figure 2. Myanmar Language Lexico-Conceptual Knowledge Resource (ML2KR)

3.1. Myanmar WordNet

Myanmar WordNet is built by using existing of WordNet from Princeton applying the semiautomatic approach [11]. The distinct principles of organization of WordNet from Princeton organized the word as concepts, viz. synsets, and act as the basic units of lexical semantics, and the hyponymy of the concepts acts as the basic relation among others. Upon this tree structure of hyponymy, there also exist some other semantic relations like holonymy, antonymy, attribute, entailment, cause, etc., which further interweave all the concepts in the lexicon into a huge semantic network. According to the specification of WordNet, the noun is categorized into 26 broad, the verb is categorized into 15 broad and the adjective is 3 broad.

3.2. Bilingual Computational Lexicon

By reusing the existing resources and manage this resources, not only the building of the monolingual Myanmar WordNet but also Myanmar-English computational lexicon benefits. The bilingual computational lexicon [6] is constructed for the further NLP application. This bilingual lexicon is built base on the Myanmar WordNet lexical database [3]. Therefore the design is greatly depend on Myanmar WordNet lexical database structure and information. In this lexicon defined the noun as 26 tag set, verb as 15 tag set, adjective as 3 tag set and adverb has 2 tag set, proposition as 17 and conjunction as 8 tag set. Beside then particle are used as indicator for defining the definite POS and produced the inflected form of word.

3.3. Morphological Processor

An important class of lexical relations is the morphological relations between word forms. Computational lexicon and WordNet as a language resource became increasingly obvious that have to deal with inflectional morphology. For this reason morphological processor is proposed in [3]. Morphocon emphasize on inflectional case of Myanmar morphology. The proposed rule based Morphocon include twofold: morphological analyzer and morphological generator. The morphological analyzer interacts with, but is separated from the lexicon. Proposed Morphocon performs analyzing Myanmar words and generating the equivalent English words: this is basically the rule of the morpheme of the Myanmar word for WordNet and grammar pattern relation between Myanmar and English word for lexicon. By supporting with Morphocon in MLRs, it can reduce the time and storage consumption. The evaluation of coverage for lexical acquisition increased to nearly tenfold of existing data.

4. SYNTAX ANALYZER

Syntax analyzer is separated into two distinct parts, lexical analysis and syntactic analysis. Lexical analysis deals with small-scale language construct which names and literals. Lexical Analysis is of central importance in computational language processing. The reason is that anything a machine, knows' about a string must be coded in a lexicon. Therefore, mapping of strings (tokens) to the lexicon is a central task in language processing.

Syntactic analysis deal with large-scale language constructs which provides information necessary for Machine Translation. Syntax refers to the relationship that sentence's element bear to each other in a sentence and syntactic analysis corresponds to recognizing various grammatical functions for example, subjects, objects, main verbs and various complements. Grammatical Functions are those functions which depend on morpholexical analysis and/or the subsequent step of parsing.

4.1. Lexical Analysis

Lexical analysis (LA) is determining the meaning of individual words, and identifying non-word tokens. To understand the morphology of each word, first tokenize the sentence and determine the word relationships. It is working together with the Myanmar-English computational lexicon [6]. The portion of the system holds all specific attributes to each word of the source sentence. Basic method of lexical analysis is the word lookup in a lexicon and it has some problem which is word-level ambiguity that words may have several meanings, and the "correct" one cannot be chosen based on the word itself for example: the word [illegible] it may become verb, noun and particle. To define the definite POS of the word, rule based approach of CFG is used.

4.1.1. Tokenization

Myanmar script uses no space between words and syllable, therefore segmentation represents a significant process in many NLP tasks such as word segmentation, sorting, line breaking and so on. Thus, the computer has to determine syllable by means of an algorithm. Moreover, a Myanmar syllable can be composed of multiple characters. Our proposed tokenization algorithm is taking into account the advantage of storage of Myanmar3. Myanmar Unicode 5.0 is stored start with consonant and one syllable can be contained one or more consonants. However, we can find out that if Myanmar syllable has more than one consonant, the followed consonant has always followed by final "☐".

4.1.2. Segmentation and Pattern Merging

Segmentation and pattern merging is done by proposed algorithm [10], which produce possible word pattern for sentence. The tokenized word is use as input to the algorithm. The output of the algorithm is the segmentation word and possible pattern. They are sent to the part of speech tagging of first phase (rule based tagging process). First, the input text of segmented syllables is broken down into sentences and phrases by looking at punctuation marks and spaces. For each sentence or phrase, all possible combinations of merged words are generated by matching segmented syllables in the sentence or phrase with word entries in the dictionary. From the resulting combinations, the one with the minimum number of merged words is selected, and taken as the correctly merged words of the sentence or phrase. This approach is biased to prefer longer word matching in the dictionary since the dictionary-based approach with longest matching works well in our internal tests for syllable merging. When there are two or more combinations with the same minimum number of merged words, our algorithm generate all possible segmented sentence of phase pattern as their score for further processing. The segmentation and pattern merging algorithm can also handle the unknown case of word and it does not only depend on lexicon.

4.2. Syntactic Analysis

Syntactic analyzing is the words in a sentence so as to uncover the grammatical structure of the sentence. This requires both a grammar and a parser. The input utterance is being checked to ensure that its syntax is correct and structured representations of the possible parses are generated. Complete syntactic analysis involves the identification of relationships among phrases and clauses within sentences. Various models of syntactic structure and methods of parsing have been adopted in MT systems. In our case, phrase structure analysis is applied. It provides labels for constituent groups in sentences: noun phrase (NP), verb phrase (VP), prepositional phrase (PP), etc. The phrase structure approach is associated most closely in the early period of MT research.

4.2.1. Part of Speech and Function Tagging

CFG is an abstract model for associating structures with strings but it is not intended as model of how humans produce sentences. Sentences that can be derived by a grammar G belong to the formal language defined by G, and are called grammatical sentences with respect to G. Sentences that cannot be derived by G are ungrammatical Sentences with respect to G. The language L_G defined by grammar G is the set of strings composed of terminal symbols that are derivable from the start symbol:

$L_G = \{w \mid w \in T \text{ and } S \text{ derives } w\}$

The generation of CFG rules has two steps.

1. Lexical rules recognizing POS from the Myanmar words are generated.
2. CFG rules recognizing phrase from POS are generated.

Some of the lexical rules are shown in table 2.

Table 2. Example of Lexical Rules

N::=PARTICLE_CHANGE_NOUN_START<&>V
N::=PARTICLE_CHANGE_NOUN_START<&>ADJ
N::=V<&>PARTICLE_CHANGE_NOUN_FOLLOW
N::= ADJ<&>PARTICLE_CHANGE_NOUN_FOLLOW
ADJ::=V<&>PARTICLE_CHANGE_ADJ
ADV::=V<&>PARTICLE_CHANGE_ADV
ADV::= ADJ<&>PARTICLE_CHANGE_ADV

For function tagging, sentence level grammar rules have to define to build parse for POS tagged words. We have to combine adjective and noun tags to form noun phase and also adverb and verb tags to verb phase. Some of the parsing rules are described using CFG as follows in table 3.

Table 3. Example of Parsing rules

NP_ARRIVAL::=NOUN_LOCATION<&>PREP_ARRIVAL
NP_ARRIVAL::=NOUN_ARTIFACT<&>PREP_ARRIVAL
NP_REASON::=NP<&> PREP_
NP_NOM::=NOUN_PERSON<&>PREP_NOM
NP_NOM::=NOUN_PERSON
NP ::=PRON_PERSON
NP::=PRON_PPP
NP::=NP_NADJ<&>PREP_POSS<&>N
NP::=N<&>PREP_POSS<&>N
NP::=N<&>PREP_POSS<&>NP_NADJ
NP::=N<&>PREP_POSS<&>NP_ADJN
VEND::=ADV<&> V
VEND::=ADV<&> V<&>PREP_VERB
VEND::=ADV<&>PARTICLE_VERB<&>V<&>PREP_VERB

5. UNKNOWN WORD DETECTION AND VERIFICATION

5.1. Name Entity Recognition (NER)

NER in Myanmar is a challenging problem for language processing. This system will extract the UNK tag set and identify the boundary of candidate NE that composes of many words by using the pre-defined rules, the proposed NER algorithm and Myanmar-English bilingual lexicon. Then, this system will display NEs and correctly classify the class of these NEs. It is important to note that there are so many available features in the Myanmar NER such that the ones concerned in this study are actually not enough for expressing all possible situations of the Myanmar NER.

Named Entity Identification includes locating named entities and classifying those names in text. Myanmar Named Entity Identification is done using ruled based method identification and verifying with syntax analyzer. Some of the naming rules are described using CFG as follows.

Table 4. Example of Name Rules

NAME_P::=NAME
NAME_P::= ADJ_P <&> PARTICLE_CHANGE_ADJ <&>NAME
NAME_P::=NAME<&>NOUN_ARTIFACT
NAME_P::=NAME<&>NOUN_LOCATION
NAME_P::=NAME_P<&>CONJ_NOUNPHRASE<&>NAME_P
NAME_P::=NAME_P<&>PREP_POSS<&>NP
NAME_P::=NOUN_PERSON<&>NAME

In Myanmar name which may or may not contained the particle prefix as (ဦး , ဒေါ်, ကို, မ etc). If we fine this particle, forward merging is used. If not we used the prepositional and noun (eg. သည်, သို့, မှ, ဖြင့်, အစိစု, etc)as marker and backward merging approach is used.

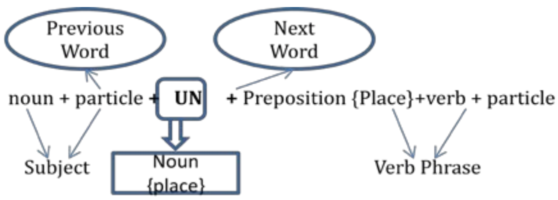

Figure 3. Example Pattern for Name of Place

In our example the preposition "သို့" is evident the unknown tag of "ဒဂုံစင်တာ" as a place of Noun tag. Therefore, the above example sentence is parsed and segmented as သူတို့ <PRON_PERSON >+ ဒဂုံစင်တာ< NAME_PLACE > + သို့ < PREP_ARRIVAL > +သွားသည် < verb.stative >.

5.2. Unknown Word Verification

After the name entity is recognizing , the tagger still does not agree with grammar rule, the possible unknown word list are generate and verify by user and add to the Myanmar Language Resources (MLRs). As the MLRs are depend on existing structure of WordNet, the word may need to enter only of equivalent English corresponding the detection of unknown Myanmar word. The synset definition and their POS are automatically assigned to this word. The user needs to verify these synset definitions manually. As a synset definition of each word, this system fine the relations WordNet that applied to the noun, adjective and verb concepts are synonymy, antonymy, hypernymy, holonymy, entailment, cause and etc. This can improve the coverage of MLRs and more efficient for further NLP application.

When the input sentence is analyzed and tagged as their function again and the function tagged tree is stored in corpus for further Myanmar NLP applications. Therefore, this system can provide the improvement of training data in corpus.

6. EXPERIMENTAL RESULTS

6.1. Experimental Study

We test the system in general domain. Sentence types in testing case are simple and compound. We tested with 100 sentences and the length of source sentences consists of word between 5 and 15. At present, Myanmar WordNet covered the 20532 words and bilingual computational lexicon covered the 25378 translation words and all are stem form. Morphocon of ML2KR handled the inflectional case of word.

We now present the results of our experiments in recall, precision and F-measure, as usual in such experiments for unknown word detection and accuracy of syntax analyzer which is depend on syntax analyzer and unknown word detector.

$$precision = \frac{\text{No: of Correct function Tags}}{\text{Total Function Tags}} * 100 \quad (1)$$

$$recall = \frac{\text{No: of Correct function Tags}}{\text{No:of Actural Existiong Function Tags}} * 100 \quad (2)$$

$$F - measure = \frac{2*(\text{precision}*\text{recall})}{(\text{precision}+\text{recall})} \quad (3)$$

6.2. Experimental Result of Name Entity Recognition

Table 5 shows the results of name extraction. The accuracy for retrieving person names was quite satisfiable. If the particle of name does not include and this may confuse with noun word, we could not extract as names. For example, "လမင်း", "နေမင်း", "စံပယ်"etc.

Table 5. Evaluation Results for person name extraction

	Recall	Precision	F-measure
With Particle (ForP)	94.52	95.89	95.20
Without Particle (BackP)	68.88	78.88	75.88

6.3. Experimental Result of Syntax Analyzer

In order to measure the performance of the system, we have tested many experiments using our approach on different types of sentences till we get the best accuracy. We can evaluate the result how many wrong chunks are tagged and how many chunks can be correctly tagged as shown in table 6. The grammar-based systems have limitations because natural language often does not conform to the rules of the grammar. Unusual constructions, casual speech, innovative expressions, mistakes, noise, and interruptions can all result in sentences that are quite understandable to a human reader or listener, but utterly confusing to a rule-based system. It is hard to write a complete and tight grammar.

Therefore, the performance of our parser is evaluated in terms of problems that can be encountered in Myanmar sentences because of some peculiar patterns. The sentences that have peculiar patterns are entered into the system and check the accuracy of our parser. For function tagging, our evaluation result is depended upon all of processing steps result.

Table 6. Evaluation Results Syntax Analyzer

	Recall	Precision	F-measure
Known	96.86	98.3	97.57
Unknown+Known	94.32	96.94	95.61
Total	96	97.96	96.97

7. CONCLUSION

We proposed an "all-purpose" method for Myanmar unknown word detection and identification. Our method is based on an morphological analysis that generates segmentations and POS tags using CFG. Our experiments showed that the proposed method is able to detect person names and organization names quite accurately and is also quite satisfactory. This paper proposed the approach for improvement the coverage of Myanmar WordNet, bilingual computational lexicon and function tagged corpus using rule based syntax analyzer. Lexical rules have to be applied to defined POS and name.

The initial system dictionary was too small, only contains 25,378 entries. Therefore, we looked for some ways to enlarge our system dictionary using unknown word detection methods. Currently, our dictionary contains 25,769 entries which is quite compatible with other systems. Our results showed that by increasing the number of entries in the dictionary, the accuracy of word segmentation and POS tagging is also improved.

REFERENCES

[1] Z Wang and G Fu. Chinese Unknown Word Identification as Known Word Tagging. *Journal of Chinese Language and Computing*. 17(1): 57-66

[2] SL Phyue. *Building Myanmar Language Resources*. In proceeding of Applied Information and Computation Technology, Myanmar. 2011.

[3] SL Phyue and A Thida. Morphological Processor for Inflectional Case of Multipurpose Lexico-Conceptual Knowledge Resource. *International Journal of Advanced Research in Computer Engineering & Technology*. 2012; 1(7).

[4] K Zhang, Q Liu, H Zhang and Xue-Qi Cheng. *Automatic Regcognition of Chinese Unknown Words Based on Roles Tagging*. In Proceedings of 1st SIGHAN Workshop on Chinese Language Processing. 2002

[5] KJ Chen and WY Ma. *Unknown Word Extraction for Chinese Documents*. The 19th International Conference on Computational Linguistics. 2002; 1: 169-175.

[6] WY Ma and KJ Chen. *A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction*. In Proceedings of the second SIGHAN workshop on Chinese language processing. 2003; 17: 31-38.

[7] GC Ling, M Asahara, Y Matsumoto. *Chinese Unknown Word Identification Using Character-based Tagging and Chunking*. In Proceeding ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. 2: 197-200

[8] M Asahara and Y Matsumoto. *Japanese Unknown Word Identification by Character-based Chunking*. In proceedings International Conference on Computational Linguistics. 2004: 459-465

[9] D Shen, M Sun and C Huang. *The application & implementation of local statistics in Chinese unknown word identification*. In *COLIPS*, (in Chinese). 1997; 8.

[10] SL Phyue. *Lexical Analyzer for Myanmar Language*. In Proceedings of the International Conference on Computer Applications, Yangon. 2012: 95-101

[11] SL Phyue. *Construction of Myanmar WordNet Lexical Database*. IEEE student Conference on Research and Development, Malaysia. 2011.

BIOGRAPHIES OF AUTHORS

SL Phyue has received her B.C.Sc (Bechelor of Computer Science) and M.C.Sc (Master of Computer Science) degrees from University of Computer Studies, Yangon in 2004 and 2008, respectively. Currently, she is a candidate for the degree of Ph.D of Information Technology in University of Computer Studies, Mandalay in Myanmar. She research interests is Natural Language Processing, Information Retrieval and semantic technology.

Dr. Aye Thida is an Associate Professor of Research and Development Department (1) at University of Computer Studies, Mandalay (UCSM), Myanmar. She was one of the leaders of Natural Language Processing Project. Her team has developed Myanmar to English Translation System in 2011. Her research interests include Distributed Processing, Queuing and Natural Language Processing. She is currently working Myanmar to English Translation System Project. Dr. Aye Thida received B.Sc (Hons) Maths degree from the Mandalay University, Myanmar and her M.I.Sc and Ph.D degrees in Computer Science from the University of Computer Studies, Yangon (UCSY), Myanmar.