First Classified Annotated Bibliography of NLP Tasks in the Burmese Language of Myanmar

JATINDERKUMAR R. SAINI

Narmada College of Computer Application
Zadeshwar, Bharuch, Gujarat - 392 011, India
[1]saini_expert@yahoo.com

Abstract. Natural Language Processing (NLP) has emerged with a wide scope of research in the area. The Burmese language, also called the Myanmar Language is a resource scarce, tonal, analytical, syllable-timed and principally monosyllabic language with Subject-Object-Verb (SOV) ordering. NLP of Burmese language is also challenged by the fact that it has no white spaces and word boundaries. Keeping these facts in view, the current paper is a first formal attempt to present a bibliography of research works pertinent to NLP tasks in Burmese language. Instead of presenting mere catalogue, the current work is also specifically elaborated by annotations as well as classifications of NLP task research works in NLP related categories. The paper presents the state-of-the-art of Burmese NLP tasks. Both annotations and classifications of NLP tasks of Burmese language are useful to the scientific community as it shows where the field of research in Burmese NLP is going. In fact, to the best of author's knowledge, this is first work of its kind worldwide for any language. For a period spanning more than 25 years, the paper discusses Burmese language Word Identification, Segmentation, Disambiguation, Collation, Semantic Parsing and Tokenization followed by Part-Of-Speech (POS) Tagging, Machine Translation Systems (MTS), Text Keying/Input, Recognition and Text Display Methods. Burmese language WordNet, Search Engine and influence of other languages on Burmese language are also discussed.

Keywords: Bibliography, Burmese, Language, Myanmar, Natural Language Processing (NLP)

(Received February 11th, 2016 / Accepted May 30th, 2016)

1 Introduction and Related Works

Natural Language Processing (NPL) deals with the processing of naturally human spoken languages by computers. This area lies on the merger boundaries of various disciplines of Computer Science including Computational Linguistics, Artificial Intelligence and Text Analysis, to name a few. The processing of written natural language is often termed as script-processing. Burmese is an officially accepted language by the constitution of Myanmar and is termed Myanmar language. According to Wikipedia [35], Burmese is a tonal, pitch-register, syllable-timed, largely monosyllabic and analytic language with a Subject-Object-Verb (SOV) word order. It is a member of the Lolo-Burmese grouping of the Sino-Tibetan language family. This language has no white spaces and hence no word boundaries. To the best of the author's knowledge, this is the first formal attempt worldwide to catalogue all the research works of NLP of some language. Naturally, then, it is more so for the Burmese language which is spoken by nearly 33 million people as the first language [35] further added by 10 million people speaking it as a second language [35], worldwide.

It is not so that the entire focus of research community is on NLP only, even discussing in context of Burmese language. There are several instances of research where the authors have dialogued on the Burmese language in context of politics, languages of families, etc. Aye and Sercombe [2] have authored a chapter a book volume that tracks the complex relationships between language, education and nation-building in Southeast Asia. Though the book focuses on role of language policies used by respective nation as instruments of control, assimilation and empowerment, their specific focus is on Burmese language in context of Burmese nation-building. A study of various Tibeto-Burman languages in addition to other languages has been presented through analysis of opinion mining research by Kaur and Saini [6]. Similarly, Fu [3] has presented the study of classifiers in entire family of Tibeto-Burman languages. It is noteworthy that Burmese languages of this family. Importantly, the researcher has proposed that the unit they form with the numeral acts as a head of the noun phrase.

2 Methodology

For a period spanning over 25 years, from 1990 to 2016, the research instances pertinent to NLP tasks of the Burmese language were mined from the web. Technically, this is an application area of Web Content Mining. The web locations used for the current work were sourced from online publications by various press and publishing houses including IEEE, Springer, ACM, Northern Illinois University (NIU)-Center for Burma Studies, Veer Narmad South Gujarat University (VNSGU), Foundation of Computer Science and IACSIT Press. As the main aim of current work was to create an annotated bibliography of research works, attention was more focused on the abstracts of research works. The research instances were also sourced from indexing and social networking sites like researchgate.net and academic.edu for initial search of pertinent research works on NLP in Burmese language.

Creating a bibliography does not pose many challenges in general but doing so for an un-common and very specific job of NLP tasks in Burmese language was definitely a challenge. The biggest hurdle was to find the scattered research material and gather them at one location. This was followed by removal of duplicate entries of research records. This task was also challenged by the fact that the same researcher may author different papers at different times in addition to doing so with another set of co-authors. Their titles may vary slightly or they may present an extension of their or others' existing algorithms. The parallel development of research works by different authors, at same or different times, was also a challenge. This was followed by the hefty task of annotating the research works, followed by identification of possible classes for the research works and then classifying relevant research works to a specific class. For instance, the category dealing with Burmese language text keying/input has been generalized to include text input for mobile phones, smart phones, laptops, Personal Computers (PCs), tablet PCs, Personal Digital Assistants (PDAs) and portable game players, as well. This task became, further, so demanding as the author of current work went through each research work thoroughly in order to annotate it as well as first identify the class for it and then categorize it. The final step was to integrate all research papers in a class in a flow, followed by maintaining a similar flow for the classes as well. These steps do involve the obvious creations, deletions, mergers and splitting of classes. Lastly, the author of current research work settled down with 6 categories of research works of NLP tasks for Burmese language. Based on the notion that first Burmese language word identification and related tasks are required, followed by POS tagging and MTS, followed by text input and display methods, the categories have been sorted in this order. This is followed by discussion on Burmese language WordNet, search engine and the influence of other languages on the Burmese language. It is needless to mention that an absolute non-overlapping class definition is not possible, given the fact that often NLP tasks are dependent on each other and assigning a sequential order to the activities is not possible. Still, a best attempt creation of catalogue, identification of classes, assignment of classes to research works and annotation is presented here. For instance, it is not possible to develop Machine Translation System (MTS) without Word Sense Disambiguation (WSD) and WSD cannot be thought of until Tokenization has been accomplished.

3 Results and Findings

This section presents the annotated classification of research works in Burmese language. The section is divided into seven sub-sections. Each one of the first six sub-sections focuses on a specific sub-domain while the seventh sub-section includes the summarized presentation of the state-of-the-art of the developed NLP tasks of the Burmese language. The first sub-section focuses on the identification, segmentation, disambiguation, collation, semantic parsing and tokenization for Burmese language. The second and third sub-sections focus respectively on Part-Of-Speech (POS) tagging and Machine Translation System (MTS). The fourth sub-section presents a discourse on the Burmese language text keying/input, recognition and display methods. Development of Burmese language WordNet and search engine followed by influence of other languages on the Burmese language are discussed respectively in the fifth and sixth sub-sections.

3.1 On Word Identification, Segmentation, Disambiguation, Collation, Semantic Parsing and Tokenization for Burmese Language

This section presenting a total of 9 approaches is mostly dominated by research approaches dealing with Burmese language word boundary identification, segmentation and sense-disambiguation. The later part of this section also discusses approaches for Burmese language semantic role labeling, word collation and sorting through tokenization.

Mon *et al.* [12] have used corpus based dictionary to propose a unified approach for the Burmese language word analysis. They have also made use of Finite State Automata (FSA), Rule Based Heuristic (RBH) Approach and Statistical Approach. Their contribution is analysis of the boundary identification and segmentation of the words written in Burmese language. In a similar work, Pa *et al.* [18] have used Conditional Random Fields (CRF) and presented a method for word boundary identification for Burmese text. Khaing and Aung [7] have discussed supervised, semi-supervised, unsupervised and knowledge-based approaches for Word Sense Disambiguation (WSD). WSD is required to be done in NLP tasks when same word may convey different meanings contextually. They have discussed WSD techniques for nouns of Burmese language. Aung and Thein [1] have presented a technique based on Nearest Neighbor Cosine classifier to disambiguate the ambiguous words of Burmese language. Basically, working towards WSD but with specific POS tags viz., 'noun' and 'verb', they have presented a WSD system. Their system also uses Myanmar-English parallel corpus as training data.

Using Myanmar Verb Frame (MVF), Naing and Thida [16] have presented a shallow semantic parsing approach commonly known as Semantic Role Labeling (SRL) technique for pre-segmented constituents of datasets in Burmese language. Mon [11] has presented a spell checker for the Burmese language. Yuzana and Tun [41] presented a collation strategy based on heuristics chart for the Burmese Language. Their method was designed to first create slices of the syllables of names and then collate them according to the traditional standard Burmese language spelling book order. They claimed that their method was able to handle simple words as well as words with complex syllabic structure. The same researchers, Yuzana and Tun [40], continuing their already proposed research work, further presented with a modified collation algorithm that performed better. A Burmese word tokenizer for sorting purpose has also been proposed by Thwin *et al.* [30]. They have handled the difficulties and complications of Burmese language's typical structure by using existing LIPIDIPIKAR treatise.

3.2 On Burmese Language Part-Of-Speech (POS) Tagging

This section presents 5 approaches for POS tagging NLP task in Burmese language of Myanmar. Myint [14] has proposed a hybrid approach for POS tagging. The author claims that such an approach could be well used for machine translation task. Myint *et al.* [15] have presented a bigrams based POS tagger for Myanmar language. The bigram tagger developed by them has used two phase approach for development purpose. They have implemented training with Hidden Markov Models (HMM) using Baum-Welch algorithm and decoding with Viterbi algorithm. Thant *et al.* [25] have presented a syntactic analysis for the Burmese language. Their syntactic analysis deploys two steps. Their first step implements function tagging and second step implements grammatical relation. They have used Naïve Bayesian theory to disambiguate the possible function tags of a word. During the second phase, they have applied Context Free Grammar (CFG) to find out the grammatical relations. Thant *et al.* [24] lament the difficulty of development of NLP algorithms for Burmese language owing to its free phrase order and complex morphological system. In absence of readily available text corpus, they have used a small functional annotated tagged corpus. They have used Naïve Bayesian statistics to disambiguate the function tags for each in sentences under consideration. Their core contribution is proposal of a set of function tags for Burmese language by handling the problem of assigning function tags to Burmese words. Zin and Thein [42] have used a machine learning and corpus-based approach for experimenting with manually-tagged and non-manually-tagged corpus. They have aimed at developing a POS tagging for the Burmese language using Hidden Markov Model.

3.3 On Burmese Language Machine Translation Systems (MTS)

This section presents 5 approaches for Burmese language Machine Translation Systems (MTS). Zin *et al.* [43] have used a statistical approach by using bilingual corpus of Burmese and English languages and presented a MTS for Burmese to English translation. They presented the system in two parts, the first part comprising of the source language is based on the n-grams model of the language while the second part comprising of the translation model considers the structural aspects of the language along with the statistical approaches like Naïve Bayesian approach. Thandar and Thein [23] have presented a parallel corpus for Burmese and English languages. They have then used corpus-based and dictionary-lookup-based approaches in addition to morphological features of the Burmese language, in order to provide a word-alignment system. As they have used multiple approaches simultaneously, they term their approach as hybrid and claim that in order to present a good MTS, good word alignment is highly required. Wai and Thein [32] have presented a reordering rule generation and Markov-based reordering model implementation. They claim that this model could be incorporated into English-Burmese translation model. They have used parallel tagged aligned corpus for achieving their results. Wai [31] have also worked on Burmese to English MTS. The author has presented a corpus-based approach to WSD by developing an ensemble of Naïve Bayesian classifiers, each of which is based on lexical features. The researcher has proposed a framework to solve ambiguous verb problems. Win [36] has presented MTS with words to phrase reordering. The researcher's MTS is designed to work for Myanmar to English machine translation by using English grammar rules.

3.4 On Burmese Language Text Keying/Input, Recognition and Display Methods

This section presents 13 approaches for the Burmese language text processing methods. These text processing methods include the approaches for text keying for mobile phone and the like devices, text recognition from scanned images of printed documents in Burmese language and the display of Burmese language numerals. The section also presents approaches used by researchers for keyboard usage and a syllable-based language model for continuous speech recognition system for Burmese language.

Oo and Thein [17] have coined a term "iTextMM" and presented a Burmese syllable prediction text input system for Android touch screen mobile phones. They have claimed that their system is an Intelligent Text Input System and has been provided for public use in Burmese market. The responses received too are promising. Thu and Urano [27] presented an approach for enabling a practical and efficient composing of message text with Myanmar language on a mobile phone. They proposed an idea of key mapping, which they term as "Positional Mapping", for the Burmese language. They have advocated that their idea is applicable for mobile phones based on Burmese language alphabets. Thu [26] has presented a consonant cluster prediction text entry method for Burmese language. The researcher's method is based on positional vowel information. The author was not able to achieve much good results but the results were promising enough for improvement in coming times. The chief advantage of the researcher's proposed predictive text input method was that even first time users can type Burmese sentences and the method was eligible enough to be used for various kinds of mobile devices. The method was further extendable for other similar syllabic languages of neighboring countries. Thu *et al.* [29] have argued that text input of syllabic scripts poses a unique challenge because many syllabic characters are formed by combinations of consonants, dependent vowel signs, tones, subscript consonants, etc. They have stated that input of such texts is not easily accomplished even with keyboards. Hence, they propose a technique consisting of gesture recognition for syllabic scripts text. They further claim that their method could be used for various devices including the hand-held devices like mobile phones. In a similar research work involving keyboards, Thu and Urano [28] have discussed a detailed comparison of various Burmese keyboard layouts like CE, Win Myanmar, Zawgyi Myanmar, MyaZedi and Myanmar3. They have examined keyboard mapping, Keystrokes Per Character (KSPC) and Characters Per Minute (CPM) in their research work.

Extending the existing Hopfield Neural Network method, Swe and Tin [22] have developed a Burmese character recognition algorithm which basically aims to identify the scanned printed Burmese characters. After scanning and recognition, the results could be used for translation of scanned material as well. In a similar approach, Win *et al.* have presented the approach for converting Burmese printed document image into machine understandable text format [37] and then reading it and optimizing it with a segmentation method and hierarchical classification scheme [38]. Sandar [20] has presented a comparison of recognition for Off-line Myanmar Handwriting and Printed Characters. The author has used the discrete hidden Markov models to accomplish the research work. Mar and Thein [10] have presented a Burmese character recognition system useful for handwriting identification as well as author attribution. Using a stylometric technique, they have deployed the usage of Fast Fourier transform (FFT) and Weighted Euclidean Distance (WED) for correct identification of subjects. Karri and Orailoglu [4] have presented an approach using seven segmented display for Burmese language numerals against an eight segmented display proposed for such numerals by Lau and Lwin [9]. The later approach proposed is the modification of the earlier approach. Soe and Theins [21] through their work have presented a syllable-based language model for continuous speech recognition system for the Burmese language.

3.5 On Burmese Language WordNet and Search Engine

This section presents 2 research works, one each for Burmese language WordNet and search engine. Using data extraction, link analyzing and construction phases, Phyue [19] has presented a methodology of constructing Burmese WordNet. The author claims that the created WordNet will act as an important lexical database for NLP tasks involving Burmese language. Mon and Mikami [13] have proposed a Burmese language search engine capable of searching the Web documents coded in multiple encodings of Burmese language. Their search engine is able to search localized content. They have presented the design and the architecture of Burmese language specific search engine.

3.6 On Influence of Other Languages on Burmese Language

This section presents 4 research works wherein the authors have discussed and presented the influence of other languages like Pali, Sanskrit and English on the Burmese language. Waxman and Aung [33] have presented a study and analysis of words borrowed from traditional Indian languages, Sanskrit and Pali, into the Burmese language of Myanmar. They have discussed the adoption and absorption of many such borrowed words, "loan-words" as termed by them, into the Burmese national language. Wheatley and Tun [34] have presented a similar study and analysis of borrowed words but in context of Burmese language and English language. They have discussed the process of phonological and semantic accommodation of words borrowed from English language, in the context of Burmese language. Lammerts [8] has presented a distinctive work on the Burmese language. He has presented the development of manuscript ornamentation in context of Burmese language. The researcher has elaborated on the influences of various languages and countries on the decorative ability of Burmese language. He has discussed this influence since old times. Yanson [39] has discussed the Burmese language more in the context of early Myanmar inscriptions surviving on Pagan donative stones which show the effect of spacial considerations, fluid spellings and the problematics of inscribing stone slabs and palm leaves on orthographic conventions.

3.7 State-of-the-art of the developed NLP tasks of the Burmese language

Given the amount of work involved for the analysis of research works of 25 years on NLP of Burmese language, this sub-section aims to present a visual representation of the state-of-the-art of the developed NLP tasks of the Burmese language. In order to be more representative and informative in terms of depicting evolutionary progression of research on NLP in Burmese language, the analysis is attempted to be presented here in perspective of two dimensions. Firstly, an analysis based on the major contribution of the researchers in a particular domain of NLP tasks in Burmese language is presented through Figure 1.

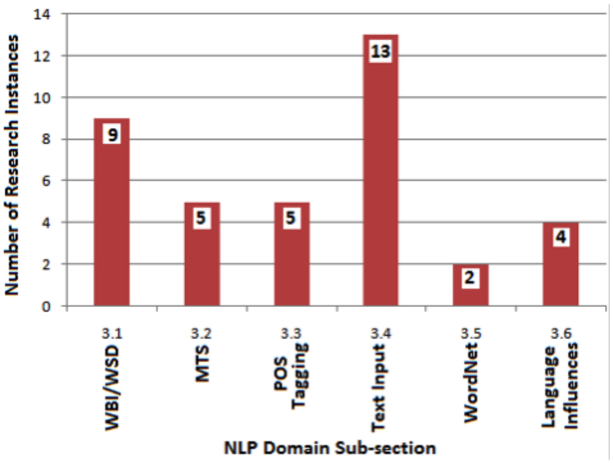

Figure 1: Burmese NLP Domain Based Number of Research Instances

Based on Figure 1, it is clear that the maximum and minimum of Burmese NLP research instances are present in the categories of Text Input/Recognition and development of WordNet respectively. The second largest number of research instances are present in the category of Word Boundary Identification (WBI) and Word Sense Disambiguation (WSD). After NLP domain based analysis of number of research instances, secondly, a timeline based analysis of research works in Burmese NLP was carried out and is presented through Figure 2. From Figure 2, it is notable that after an initial step of Burmese NLP development in early 90s, the progression towards further development was quite stunted. In fact, for the next entire decade, there was hardly any major development. Beginning of the 21st century marked a era of Burmese NLP development and the first decade witnessed publication of 16 research works in total. The maximum number of 12 Burmese NLP research instances in a single year have been found

Table 1: Legend for visual representation of Burmese NLP tasks

Color Code/ Sub-section	Burmese NLP Research Area
3.1	WBI, Word Segmentation, WSD, Collation, Semantic Parsing, Tokenization, Srl, Spell Checker
3.2	POS Tagging, Syntactic Analysis
3.3	MTS
3.4	Text/Speech Recognition, Text Input/Display
3.5	WordNet, Search Engine
3.6	Language Influences

in the year 2011. Even though the development of Burmese NLP tasks is slow in recent years but it has been progressing steadily with a huge scope in the near future.

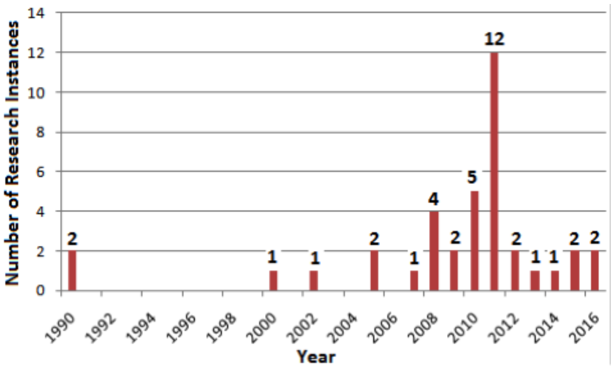

Figure 2: Year-wise Development of Burmese NLP Based Number of Research Instances

Table 1 presents the visual glimpse of year-wise domain based development tasks of Burmese NLP. The last column of Table 1 includes the broad six areas of contribution of researchers in domains of Burmese language NLP.

It is remarkable to notice that these six categories respectively correspond directly to the sub-sections 3.1 through 3.5 of section 3. Table 1 presents the legend used for representing different entities in Table 2. These respective six sub-sections depicted by yellow, blue, pink, green, purple and brown colors respectively are for (a) Word Boundary Identification (WBI), Work Tokenization, Word Sense Disambiguation (WSD), Spell Checker and Semantic Role Labeling (SRL) (b) POS Tagging and Syntactic Analysis (c) Machine Translation System (d) Identification of Text and Speech Input and Display of Text (e) WordNet and Search Engine and (f) Language Influences.

Table 2 has been specifically sorted on the year of publication of the research work to give a meticulous glimpse of the progressive "timeline based" development of the research tasks. Further, the color representation of the six sub-sections is done to clearly depict the Burmese language "NLP domain-based" developmental flow of the research tasks. Focusing on both above two statements simultaneously, it is evident from Table 2 that the initial development of NLP in Burmese language took place during the early 90s and the first decade of the 21^{st} century. It was highly concentrated on methods of text recognition, input and display. The second decade of the 21^{st} century witnessed the second stage of development through research works on Word boundary identification along with word sense disambiguation supported by word tokenization. The next stage was third phase in the evolutionary development and observed the development of POS-tagging and Machine Translation System (MTS). This is the focus area till date in the current decade.

It is very natural to observe that the "timeline based" and "domain based" flow of development of NLP in Burmese language follows a typical, expected and standard sequence. Typically, like any other natural language, it has emerged from text input, text identification, text tokenization to text sense disambiguation followed by POS tagging and now the development of text translation systems. On the sidelines of these observations, the author advocates that this is the first formal attempt worldwide to present the "timeline based" and "domain based" evolutionary development of NLP tasks of any language, and more specifically for Burmese language. Even though an impression of linear flow of development is hinted here, it is nevertheless not the case. Given the better comprehension of capabilities of modern technological breakthroughs, the researchers in last few years have also been found to work on areas otherwise believed to be the initial stages of NLP development of any language. Such research works include usage of machine learning techniques and CRF for word sense disambiguation and word boundary identification, for instance, in 2016. It is further predicted that the times to come may also see a rework on the initial stages of NLP development tasks, given the better understanding of developmental flow, more exposure to similar languages and powerful technological tools.

4 Conclusion

Burmese language NLP is not very widely covered in the scientific literature. Still, based on the analysis of

Table 2: Year-wise Burmese NLP domain-based Research Areas/Approaches and Major Contributions

Sr. No.	**Reference**	**Year**	**Research Area/ Approach**	**Major Contribution**
1	Lau and Lwin	1990	Segmented display approach; 8-segments	Text Display (3.4)
2	Karri and Orailoglu	1990	Modified segmented display approach; 7-segments	Text Display (3.4)
3	Wheatley and Tun	2000	Language comparison of English and Burmese	Language Influences (3.4)
4	Mar and Thein	2005	Handwritten character identification; FFT and WED; Author attribution	Text Recognition (3.4)
5	Sandar	2005	Handwritten and printed character identification; Discrete HMM	Text Recognition (3.4)
6	Swe and Tin	2005	Printed character identification; Hopfield Neural Network	Text Recognition (3.4)
7	Thu and Urano	2007	Typing Burmese on a smart-phone; Key-mapping	Text input (3.4)
8	Thu	2008	Positionoal prediction of consonant cluster	Text Input (3.4)
9	Thu *et al.*	2008	Gesture-based text input	Text Input (3.4)
10	Yuzana and Tun	2008	Collation algorithm	WBI (3.1)
11	Yuzana and Tun	2008	Modified Collation algorithm based on Heuristics Chart	WBI (3.1)
12	Thu and Urano	2009	Comparison of Burmese PC keyboard layouts	Text input (3.4)
13	Zin and Thein	2009	Hidden Markov Model (HMM)	POS Tagging (3.2)
14	Lammerts	2010	Manuscript ornamentation	Language Influences (3.6)
15	Mon and Mikami	2010	Search engine	Search Engine (3.5)
16	Mon *et al.*	2010	RBH and Statistical approach with FSA	WBI (3.1)
17	Thwin *et al.*	2010	Word sorting and tokenization	Word Tokenizer (3.1)
18	Aung and Thein	2011	Nearest Neighbor Cosine classifier	WSD (3.1)
19	Myint C.	2011	Hybrid approach	POS Tagging (3.2)
20	Myint et. al	2011	Bigram-based approach	POS Tagging (3.2)
21	Oo and Thein	2011	Typing Burmese on a smartphone	Text Input (3.4)
22	Phyue	2011	Lexical database	WordNet (3.5)
23	Thandar and Thein	2011	Hybrid approach; English and Burmese parallel corpus	MTS (3.3)
24	Thant *et al.*	2011	Function tagging	POS Tagging (3.2)
25	Thant *et al.*	2011	Naïve Bayes and Context Free Grammar	Syntactic Analysis (3.2)
26	Wai	2011	Naïve Bayes classifier	MTS (3.3)
27	Wai and Thein	2011	Markov-based reordering model	MTS (3.3)
28	Win	2011	Words to phrase reordering; English grammar rules	MTS (3.3)
29	Win *et al.*	2011	Printed character identification	Text Recognition (3.4)
30	Win *et al.*	2011	Printed character identification Heirarchical classification scheme	Text Recognition (3.4)
31	Mon	2012	Spell checker	Spell Checker (3.1)
32	Zin *et al.*	2012	Statistical approach; Bilingual Burmese-English corpus; Naïve Bayes	MTS (3.3)
33	Yanson	2013	Language comparison of stone-age and current Burmese	Language Influences (3.5)
34	Waxman and Aung	2014	Language comparison of Sanskrit, Pali and Burmese	Language Influences (3.5)
35	Naing and Thida	2015	Semantic Role Labeling (SRL)	Role Labeling (3.1)
36	Soe and Theins	2015	Syllable-based model	Speech Recognition (3.4)
37	Khaing and Aung	2016	Supervised, Semi-supervised, Un-supervised and Knowledge-based approaches	WSD (3.1)
38	Pa *et al.*	2016	Conditional Random Fields (CRF)	WBI (3.1)

available research instances of Burmese language NLP tasks, the current paper presented a first bibliography of such research works scattered over a period of more than 25 years from 1990 to 2016 (for research papers accepted and to be published). It was further enriched by descriptive annotations as well as the classification of research instances into pertinent NLP categories. A total of 6 categories were identified, viz. Burmese Language Word Identification, Segmentation, Disambiguation, Collation, Semantic Parsing and Tokenization; Part-Of-Speech (POS) Tagging; Machine Translation Systems (MTS); Text Keying/Input, Recognition and Display Methods; WordNet and Search Engine; and influence of other languages on Burmese Language. The current work is a significant source of input, documented records and contributions to someone willing to take the Burmese NLP research further. Apart from being the starting point for someone to pursue further research, the current work is also an indication of where the field of Burmese NLP is going. The highest numbers of research instances were found for category dealing with Burmese language text keying/input, recognition and text display methods. The minimum numbers of research instances were found for the Burmese language WordNet and search engine. This indicates that there is still a good scope of research in these areas.

It is concluded that the authors have lamented the lack of parallel corpus for Burmese language as well as a formally standardized and dense domain-specific corpus of Burmese language. Being a Subject-Object-Verb (SOV) ordered language with no white spaces and word boundaries poses a major challenge for word identification, segmentation as well as MTS tasks. Like most of other languages, WSD also remains an area of concern for developing NLP algorithms using Burmese language. Finite State Automate (FSA), Rule Based Heuristics (RBH), Conditional Random Fields (CRF) Dictionary-based approach and Statistical approaches have been used by researchers for word identification and segmentation. Supervised, Semi-supervised, Unsupervised, Knowledge-based and Nearest Neighbor Cosine Classifier approaches have been used by researchers for WSD. Nouns and verbs have been main areas of focus for WSD. Myanmar Verb Frame (MVF) has been used for Semantic Role Labeling (SRL) in Burmese language. For POS tagging, the researchers have made use of Bigrams, Hybrid approach, Hidden Markov Model (HMM), Baum-Welch algorithm, Viterbi algorithm, Syntactic Analysis, Function Tagging, Grammatical Relation Analysis with Context Free Grammar (CFG), Naïve Bayesian Approach and Corpus based approach.

For MTS, the concepts of bi-lingual corpus, reordering rule generation and Markov-based reordering model, English grammar rules, word alignment and Naïve Bayes along with other statistical approaches have been pondered upon by researchers. The researchers have also presented text input methods for mobile devices like smart phones using messaging, actual text input by typing as well as through gesture identification. The focus of input through such methods is perceptibly using Burmese language script. A comparison of Burmese keyboard layouts also moves on this line. For processing of scanned images containing Burmese language and author attribution, the researchers have used the concepts of Hopfield Neural Network, Machine understandable text, Hierarchical classification scheme, Discrete HMM, Fast Fourier transform (FFT) and Weighted Euclidean Distance (WED). Researchers have also discussed about influence of Sanskrit, Pali and English languages on the Burmese language.

The author does not claim the pre-eminence or inadequacy of any language when compared with and including the Burmese language. The current work is just an academic research work presenting first results meant to promote NLP of Burmese language and discuss the related discourse. It is believed that the current paper, though not claimed to be an exhaustive one on the subject, will still act as a very good source of starting point for NLP algorithmic development for Burmese language. It is noteworthy that other phonetically similar languages like Khymer, Thai, Hindi and Bangla may also use the concepts presented here for NLP development in respective languages.

REFERENCES

[1] Aung, N. T. T. and Thein, N. L. 2011, Word sense disambiguation system for Myanmar word in support of Myanmar-English machine translation, published in proceedings of SICE Annual Conference (SICE), IEEE, 2835-2840

[2] Aye, K. K. and Sercombe, P. 2014, Language, Education and Nation-building in Myanmar, published in Language, Education and Nation-building, Palgrave Studies in Minority Languages and Communities, ISBN: 978-1-349-54633-6, Springer, 148-164

[3] Fu, J. 2014, The Status of Classifiers in Tibeto-Burman Languages, published in Space and Quantification in Languages of China, ISBN: 978-3-319-10039-5, Springer, 37-54

[4] Karri, R. and Orailoglu, A. 1990, Standard seven segmented display for Burmese numerals, published in proceedings of IEEE Transactions on Consumer Electronics, IEEE, 36(4): 959-961

[5] Kaur, J. and Saini ,J. R. 2014, A Study and Analysis of Opinion Mining Research in Indo-Aryan, Dravidian and Tibeto-Burman Language Families, published in International Journal of Data Mining and Emerging Technologies, ISSN: 2249-3212, Indian Journals, 4(2): 53-60

[6] Kaur, J. and Saini, J. R. 2015, A Study of Text Classification Natural Language Processing Algorithms for Indian Languages, published in VNSGU Journal of Science and Technology, ISSN: 0975-5446, Veer Narmad South Gujarat University, 4(1): 162-167

[7] Khaing, P. P. and Aung, T. N. 2016, Machine Learning Techniques for Myanmar Word-Sense Disambiguation, published in Genetic and Evolutionary Computing, Advances in Intelligent Systems and Computing, ISSN: 2194-5357, Springer, 387: 175-185

[8] Lammerts, C. 2010, Notes on Burmese Manuscripts: Text and Images, published in The Journal of Burma Studies, ISSN: 1094-799X, NIU-Center for Burma Studies, 14: 229-253

[9] Lau, K. T. and Lwin, D.T. 1990, Segmented display for Burmese numerals, published in proceedings of IEEE Transactions on Consumer Electronics, IEEE, 36(2): 84-88

[10] Mar, S. H. and Thein, N. L. 2005, Myanmar Character Identification of Handwriting Between Exhibit and Specimen, published in proceedings of 6th Asia-Pacific Symposium on Information and Telecommunication Technologies, IEEE, 95-98

[11] Mon, A. M. 2012, Spell checker for Myanmar language, published in proceedings of International Conference on Information Retrieval Knowledge Management (CAMP), IEEE, 12-16

[12] Mon, A. M., Thein, M. M., Htay, S. S., Phyue, S. L. and Win T. T. 2010, Analysis of Myanmar Word boundary and segmentation by using Statistical Approach, published in proceedings of 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), IEEE, 5: 233-237

[13] Mon, P. Y. and Mikami, Y. 2010, Myanmar language search engine, published in proceedings of International Conference on Advances in ICT for Emerging Regions (ICTer), IEEE, 69-74

[14] Myint ,C. 2011, A Hybrid Approach for Part-of-Speech Tagging of Burmese Texts, published in proceedings of International Conference on Computer and Management (CAMAN), IEEE, 1-4

[15] Myint, P. H., Htwe, T. M. and Thein, N. L. 2011, Bigram Part-of-Speech Tagger for Myanmar Language, published in proceedings of International Conference on Information Communication and Management, ISSN: 2010-460X, IACSIT Press, 16: 147-152

[16] Naing, M. N. and Thida, A. 2016, A Sematic Role Labeling Approach in Myanmar Text, published in Genetic and Evolutionary Computing, Advances in Intelligent Systems and Computing, ISSN: 2194-5357, Springer, 388: 423-429

[17] Oo, P. W. and Thein, N. L. 2011, iTextMM: Intelligent Text Input System for Myanmar Language on Android Smartphone, published in IT Convergence and Services, Lecture Notes in Electrical Engineering, ISSN: 1876-1100, Springer, 107: 661-670

[18] Pa, W. P., Thu, Y. K., Finch, A. and Sumita E. 2016, Word Boundary Identification for Myanmar Text Using Conditional Random Fields, published in Genetic and Evolutionary Computing, Advances in Intelligent Systems and Computing, ISSN: 2194-5357, Springer, 388: 447-456

[19] Phyue, S. L. 2011, Construction of Myanmar WordNet lexical database, published in proceedings of IEEE Student Conference on Research and Development (SCOReD), IEEE, 327-332

[20] Sandar, K. 2005, A Comparison of Recognition for Off-line Myanmar Handwriting and Printed Characters, published in proceedings of 6th Asia-Pacific Symposium on Information and Telecommunication Technologies, IEEE, 105-110

[21] Soe, W. and Theins, Y. 2015, Syllable-based Myanmar language model for speech recognition, published in proceedings of IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), IEEE, 291-296

[22] Swe, T. and Tin, P. 2005, Recognition and Translation of the Myanmar Printed Text Based on Hopfield Neural Network, published in proceedings of 6th Asia-Pacific Symposium on Information and Telecommunication Technologies, IEEE, 99-104

[23] Thandar, N. K. and Thein, N. L. 2011, Word alignment system based on hybrid approach for Myanmar-English machine translation, published in proceedings of SICE Annual Conference (SICE), IEEE, 2841-2846

[24] Thant, W. W., Htwe, T. M. and Thein, N. L. 2011, Function Tagging for Myanmar Language, published in International Journal of Computer Applications, ISSN: 0975 â8887, Foundation of Computer Science, 26(2): 34-41

[25] Thant, W. W., Htwe, T. M. and Thein, N. L. 2011, Syntactic Analysis of Myanmar Language, published in proceedings of International Conference on Computer Applications (ICCA)

[26] Thu, Y. K. 2008, Positional prediction: consonant cluster prediction text entry method for burmese (myanmar language), published in proceedings of Conference on Human Factors in Computing Systems, ISBN: 978-1-60558-012-8, ACM, 3783-3788

[27] Thu, Y. K. and Urano, Y. 2007, Positional Mapping Multi-tap for Myanmar Language, published in Human-Computer Interaction: Interaction Platforms and Techniques, Lecture Notes in Computer Science, ISSN: 0302-9743, Springer, 4551: 486-495

[28] Thu, Y. K. and Urano, Y. 2009, A comparison of Myanmar PC keyboard layouts, published in proceedings of Eighth International Symposium on Natural Language Processing, IEEE, 15-20

[29] Thu, Y. K., Phavy, O. and Urano, Y. 2008, Positional gesture for advanced smart terminals: Simple gesture text input for syllabic scripts like Myanmar, Khmer and Bangla, published in proceedings of Innovations in NGN: Future Network and Services, 2008. K-INGN 2008, First ITU-T Kaleidoscope Academic Conference, IEEE, 77-84

[30] Thwin, T. T. and Win, A. T., Wai, P. P. and Thwin, M. M. S. 2010, Proposed Myanmar Word Tokenizer based on LIPIDIPIKAR treatise, published in proceedings of 2nd International Conference on Computer Engineering and Technology (ICCET), IEEE, 7: 136-140

[31] Wai, P. P. 2011, Myanmar to English verb translation disambiguation approach based on Naïve Bayesian classifier, published in proceedings of 3rd International Conference on Computer Research and Development (ICCRD), IEEE, 3: 6-9

[32] Wai, T. T. and Thein, N. L. 2011, Markov-based reordering model for English-Myanmar translation, published in proceedings of SICE Annual Conference (SICE), IEEE, 2736-2740

[33] Waxman, N. and Aung S. T. 2014, The Naturalization of Indic Loan-Words into Burmese: Adoption and Lexical Transformation, published in The Journal of Burma Studies, ISSN: 1094-799X, NIU-Center for Burma Studies, 18(2): 259-290

[34] Wheatley, J. and Tun, S. S. H. 2000, Languages in Contact: The Case of English and Burmese, published in The Journal of Burma Studies, ISSN: 1094-799X, NIU-Center for Burma Studies, 4: 61-99

[35] Wikipedia, the free encyclopedia. 2015, Burmese Language, Wikimedia Foundation Inc., Available: https: //en.wikipedia.org/wiki/Burmese_language

[36] Win, A. T. 2011, Words to phrase reordering machine translation system in Myanmar-English using English grammar rules, published in proceedings of 3rd International Conference on Computer Research and Development (ICCRD), IEEE, 3: 50-53

[37] Win, H. P. P., Khine, P. T. T. and Tun, K. N. N. 2011, Converting Myanmar printed document image into machine understandable text format, published in proceedings of Sixth International Conference on Digital Information Management, IEEE, 96-101

[38] Win, H. P. P., Khine ,P. T. T. and Tun, K. N. N. 2011, OCRMPD: OCR system for Myanmar printed document image with a novel segmentation method and hierarchical classification scheme, published in proceedings of IEEE International Conference on Intelligent Computer Communication and Processing, IEEE, 285-291

[39] Yanson, R. A. 2013, Anticipating Computer Language - On Some Conventions in the Burmese In-

scriptions, published in The Journal of Burma Studies, ISSN: 1094-799X, NIU-Center for Burma Studies, 17(2): 391-402

[40] Yuzana, and Tun, K. M. 2008, A comparison of collation algorithm for Myanmar language, published in proceedings of Third International Conference on Digital Information Management, IEEE, 538-543

[41] Yuzana, and Tun, K. M. 2008, Collation Strategy Based on Heuristics Chart for Myanmar Language, published in proceedings of Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, IEEE, 640-645

[42] Zin, K. K. and Thein, N. L. 2009, Part of speech Tagging for Myanmar using Hidden Markov Model, published in proceedings of International Conference on the Current Trends in Information Technology (CTIT), IEEE, 1-6

[43] Zin, T. T., Soe, K. M. and Thein, N. L. 2012, Translation Model of Myanmar Phrases for Statistical Machine Translation, published in Advanced Intelligent Computing Theories and Applications with Aspects of Artificial Intelligence, Lecture Notes in Computer Science, ISSN: 0302-9743, Springer, 6839: 235-242