

Efficient Data Preparation for Manipuri Language Processing: Preprocessing Strategies for Word Sense Disambiguation

Chingakham Ponykumar Singh¹, Dr. H. Mamata Devi²

^{1,2}Department of Computer Science, Manipur University, Indo Myanmar Road, Canchipur, Imphal, Manipur – 795003

ABSTRACT

The data stored within data centers often arrives in a state that is not immediately conducive to experimental endeavors. To harness its full potential, data must undergo a refinement process, transforming it into a format that computer systems can readily comprehend and utilize to execute the required actions. This paper focuses on the preprocessing of Manipuri Synset and Manipuri Corpus data, sourced from the TDIL data center, along with electronic dictionary data. The preprocessing tasks encompass the conversion of non-Unicode data to Unicode, spelling correction, tokenization for text segmentation, removal of stop words and stemming to reduce words to their root form. The primary objective of this paper is to prepare data for immediate use in word sense disambiguation for the Manipuri language using the Meitei/Meetei mayek script. This is pivotal for precise language understanding and semantic interpretation. Importantly, the processed data extends its utility beyond word sense disambiguation. It can be applied in various natural language processing (NLP) research areas including Machine Translation, Information Retrieval and Question Answering, where language comprehension is paramount. These preprocessing efforts offer a versatile tool for advancing language technology and facilitating in-depth research in the field of NLP.

Keywords: Preprocessing, Data, IndoWordNet, Manipuri, WSD

1. INTRODUCTION

Meiteilon, also known as Manipuri language is a Tibeto-Burman language spoken primarily in the Indian state of Manipur and some neighbouring regions of India and Myanmar. It has a unique script known as Meitei Mayek, which has been in use since many years. The language has a complex grammatical structure and its verbs are inflected for tense, mood, aspect and person. Meiteilon has been influenced by Sanskrit, Assamese, and Bengali languages and has borrowed many words from these languages. Meitei Mayek script details are shown in the following tables[14,15].

Table 1(a): Meitei Mayek 27 consonants

ꯀ	ꯁ	ꯂ	ꯃ	ꯄ	ꯅ	ꯆ	ꯇ	ꯈ
Kok	Sam	Lai	Mit	Pa	Na	Chil	Til	Khou
ꯉ	ꯊ	ꯋ	ꯌ	ꯍ	ꯎ	ꯏ	ꯐ	ꯑ
Ngou	Thou	Wai	Yang	Huk	Uoon	Ee	Pham	Atiya
ꯒ	ꯓ	ꯔ	ꯕ	ꯖ	ꯗ	ꯘ	ꯙ	ꯚ
Gok	Jham	Rai	Ba	Jil	Dil	Ghou	Dhou	Bham

Table 1(b): Meitei Mayek 8 half consonants

ꯛ	ꯜ	ꯝ	ꯞ	ꯟ	ꯠ	ꯡ	ꯢ
Kok Lonsum	Lai Lonsum	Mit Lonsum	Pa Lonsum	Na Lonsum	Til Lonsum	Ngou Lonsum	EE Lonsum

Table 1(c): Meitei Mayek 8 vowels, Cheikhei, Lum Iyek and Apun Iyek





















										
Onap	Inap	Anap	Yenap	Sounap	Uunap	Cheinap	Nung	Cheikhei	Lum Iyek	Apun Iyek

Table 1(d): Meitei Mayek 10 numerals

Phun	Ama	Ani	Ahum	Mari	Manga	Taruk	Taret	Nipal	Mapal
									

Words may have different sense or meaning based on the context of its usage in a particular area or topic. Words having the same spelling with the same meaning have the ability to expressed or represent different situation or meaning based on the area where they are being used. This leads to a serious problem of ambiguity to the layman while categorizing different words. Even our human language has so many ambiguity as different words can be represented in different ways. This ambiguity creates lots of problem while expressing a statement. WSD provide an effective mechanism to resolve these problems. Not only in humans, in machine translation also these problems of ambiguity still remain a mystery which many researchers are trying to solve[7]. The unique characteristic that sets language processing is its reliance on language knowledge[6]. Word Sense Disambiguation is a sub branch of NLP (Natural Language Processing) which has the ability to determine, which meaning of word is activated by the use of word in a particular context and also deals with determining the intended meaning of a word in a given context. In other word, it is the process of identifying the correct sense of a word from a set of possible senses based on the context in which the word approaches.

The structure of the paper is as follows: Section 2 provides a detailed examination of the work, including insights and preprocessing procedures. Section 3 is dedicated to the discussion and Section 4 offers the paper's concluding remarks.

2. PROPOSED METHODOLOGY

In this paper, we focus on how the Manipuri related data to WSD are collected which are either directly readable by the computer or not. All these data are converted into Meitei/Meetei mayek machine readable format and the preprocessing task are then performed on these machine readable Manipuri Meitei/Meetei mayek data. The detailed explanation is discussed in this paper.

2.1 DATA COLLECTION

Manipuri language is a language that has very limited electronic data specially in the Meitei Mayek script. Through proper channel, the data from the TDIL are collected. These collected data comprise of many fields such as science, arts, literature and media. Art field contains data from economics, history, law, linguistics, philosophy, politics, psychology, religion and sociology while science data contains biology, botany, chemistry, geography, mathematics, medicine, physics, wild life, zoology and other related data. In literature, it contains the data of Arts and Crafts, criticism, culture, didactic, novel, short fictions, theatre and trivia. Lastly, media data contains magazine and newspaper.

Table 2: Description of data collection

Name of the resource	Script	Format	Remark
Manipuri Synset (IndoWordNet data)	Meitei Mayek	Non – Unicode	Need to be convert in Unicode format
Manipuri Corpus (Monolingual corpus)	Bengali	Unicode	Need to be converted to Meitei Mayek Unicode format
Electronic Dictionary	Bengali	Unicode	Need to be converted to Meitei Mayek Unicode format

2.2 TASKS IN PREPROCESSING

2.2.1 NON – UNICODE TO UNICODE CONVERSION

Data are available in various formats as per the suitability of data. But when it comes to processing of data by the machine, the data in Unicode format works the best. Only few Indian regional languages data are available in Unicode format. Manipuri language is one of the Indian regional language which has limited electronic data and among the available data, most of them are in non-unicode format or in Bengali script. The necessity to Manipuri language data to be processed by the machine forces the non-unicode Manipuri data to be converted in Unicode Manipuri data[15]. Initially, 20% of data were converted into unicode format by manually typing or extracting the data from the TDIL IndoWordNet website. The non – Unicode data format is programmatically converted into Unicode data format by mapping of the non – Unicode character code with the Unicode character code of each character of Manipuri language(Meitei Mayek script) and it is stored in machine readable file format.

Table 3: List of Meitei Mayek Unicode values

Character	Name	Decimal	Hex	Character	Name	Decimal	Hex
ꯀ	KOK	43968	ABC0	ꯁ	LAI	43996	ABDC
ꯂ	SAM	43969	ABC1	ꯃ	MIT	43997	ABDD
ꯄ	LAI	43970	ABC2	ꯅ	PA	43998	ABDE
ꯆ	MIT	43971	ABC3	ꯇ	NA	43999	ABDF
ꯈ	PA	43972	ABC4	ꯉ	TIL	44000	ABE0
ꯊ	NA	43973	ABC5	ꯋ	NGOU	44001	ABE1
ꯌ	CHIL	43974	ABC6	ꯍ	I	44002	ABE2
ꯎ	TIL	43975	ABC7	ꯏ	ONAP	44003	ABE3
ꯐ	KHOU	43976	ABC8	ꯑ	INAP	44004	ABE4
ꯒ	NGOU	43977	ABC9	ꯓ	ANAP	44005	ABE5
ꯔ	THOU	43978	ABCA	ꯕ	YENAP	44006	ABE6
ꯖ	WAI	43979	ABCB	ꯗ	SOUNAP	44007	ABE7
ꯘ	YANG	43980	ABCC	ꯙ	UNAP	44008	ABE8
ꯚ	HUK	43981	ABCD	ꯛ	CHEINAP	44009	ABE9
ꯜ	UN	43982	ABCE	ꯝ	NUNG	44010	ABEA
ꯞ	I	43983	ABCF	ꯟ	CHEIKHEI	44011	ABEB
ꯠ	PHAM	43984	ABD0	ꯠ	LUM IYEK	44012	ABEC
ꯡ	ATIYA	43985	ABD1	ꯡ	APUN	44013	ABED
ꯣ	GOK	43986	ABD2	ꯢ	PHUN	44016	ABF0
ꯤ	JHAM	43987	ABD3	ꯣ	AMA	44017	ABF1
ꯥ	RAI	43988	ABD4	ꯤ	ANI	44018	ABF2
ꯦ	BA	43989	ABD5	ꯥ	AHUM	44019	ABF3
ꯧ	JIL	43990	ABD6	ꯦ	MARI	44020	ABF4
ꯨ	DIL	43991	ABD7	ꯧ	MANGA	44021	ABF5
ꯩ	GHOU	43992	ABD8	ꯨ	TARUK	44022	ABF6
ꯪ	DHOU	43993	ABD9	ꯩ	TARET	44023	ABF7
꯫	BHAM	43994	ABDA	ꯪ	NIPAL	44024	ABF8
꯬	KOK	43995	ABDB	꯫	MAPAL	44025	ABF9

The main advantage of keeping the Meitei/Meetei mayek data into unicode format is that any unicode data can be processed by the machine for any kind of research works. It also helps in converting one file format to another file format instantly with minimal efforts. For instance, conversion of .txt file format to .xlsx or .rtf or .csv can be done easily.

Table 4: Unicode Converted data

Sample of Non – Unicode data	Unicode converted sample data
mnuQd czlCtr_ib atiTisiQ adu yAMn TunmC mnuQd czhLlC1o	ꯏꯛꯑꯃ ꯏꯂꯛꯇꯣꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛꯏ (Manungda changlaktriba atithising adu yamna thunamak manungda changhanlako)
aYKoin yuMdgi heC TorCpg KzhOdn noQ cuTrClMmi	ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ (Eikhoina yumdage hek thorakpaga khanghoudana nong chutharaklammi)
akib UxtuN puCniQ soNThNb	ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ ꯏꯛꯑꯃꯛ (Akiba uttuna pukning sonthahanba)

Manipuri Synset data, which are not in unicode data format are successfully converted using the above mentioned method. This method can be used of converting any other data which are not in unicode data format to the machine readable unicode data format.

2.2.2 SPELLING CORRECTION

Most of the data entry works are either done automatically or manually. If the data are entered automatically it may wrongly interpret in some special cases like non-ordinary grammatical rule, unsupported special character or symbols by the machine. Manual data entry may contain typo mistakes. Thus, it has become a necessity to recheck the collected data for correctness. In this paper, the Unicode converted data from the above preprocessing step contains spelling mistakes. The “atap” mayeks are consecutively contains in some words, which are not allowed in the Manipuri language(in Meitei Mayek script). This special case happened due to the fact that the rule applied for representing “atap” mayeks to store data in the non-unicode format changes when the “atap” mayeks are actually represented in the Unicode format. The converted data of Manipuri synset and Manipuri corpus contain the above mentioned mistakes. To rectify this error, every word is manually checked by human experts. Human experts not only check the spelling of the words but it also checks whether the used words in the sentence are also appropriate within the sentence. During this phase of preprocessing, some wrong entry of “apun” mayek and “ba” are also encountered, which are manually corrected.

This preprocessing step is very time consuming yet very much needed to avoid misleading meaning of word or sentence, which play a very vital role in NLP applications like Word Sense Disambiguation, Question and Answering and Machine Translation to name a few.

2.2.3 TOKENIZATION

Word Sense Disambiguation can be performed on word level or sentence level[7]. Hence, the whole data contains in the corpus needs to be brought down into smaller unit. The process of bringing down into an individual smaller unit from the whole data in the corpus is termed as tokenization[4]. The individual word so obtained after tokenization are term as tokens. These tokens can be words, subwords, or characters, depending on the granularity of the tokenization approach. Tokenization can be performed using various delimiter such as spaces, comma(,), full stop(.), enter etc[11,12].

Tokenization can be performed on sentence level or word level. Sentence level tokenization is the splitting of the given whole content into individual sentence while word level tokenization is the splitting the whole content of a corpus into a single individual word[12]. In this paper, tokenization is performed by using spaces as a delimiter. Manipuri synset and Manipuri corpus data are broken down in individual word and hence word level tokenization is performed in this paper.

Table 6: List of Manipuri Suffixes

ꯀꯁꯃꯤ (Gidana)	ꯃꯩꯂꯤ (Dubuna)	ꯀꯁꯃꯤꯂꯩꯂꯩ (Gidamaktne)	ꯀꯃ (Sina)	ꯃꯩꯂꯩ (Dubude)	ꯀꯁꯃꯤꯂꯩ (Sidage)	ꯃꯩꯂꯩꯂꯩ (Dudage)
ꯀꯁꯃꯤꯂꯩ (Gidade)	ꯃꯩꯂꯤ (Dana)	ꯀꯁꯃꯤꯂꯩ (Gidade)	ꯃꯩꯂꯤ (Duna)	ꯀꯃ (Gi)	ꯀꯁꯃꯤꯂꯩ (Dagede)	ꯀꯁꯃꯤꯂꯩ (Sigeda)
ꯀꯁꯃꯤꯂꯩꯂꯩ (Sigedade)	ꯀꯁꯃꯤꯂꯩ (Sidana)	ꯀꯁꯃꯤꯂꯩꯂꯩ (Sigede)	ꯃꯩꯂꯩ (Nade)	ꯃꯩꯂꯩ (Duge)	ꯀꯁꯃꯤꯂꯩꯂꯩ (Sidagede)	ꯀꯁꯃꯤ (Gida)
ꯃꯩꯂꯩꯂꯩꯂꯩ (Dugidadi)	ꯃꯩꯂꯩꯂꯤ (Dudana)	ꯃꯩꯂꯩꯂꯩꯂꯩ (Dugede)	ꯀꯃꯩꯂꯩ (Sinade)	ꯀꯃꯩꯂꯩ (Side)	ꯃꯩꯂꯩꯂꯩꯂꯩ (Dugedade)	ꯃꯩꯂꯩꯂꯩ (Dunade)
ꯃꯩꯂꯩꯂꯩꯂꯩ (Dugiga)	ꯃꯩꯂꯩꯂꯩꯂꯩ (Dagena)	ꯃꯩꯂꯩꯂꯩꯂꯩ (Douna)	ꯃꯩꯂꯩꯂꯩ (Dunade)	ꯀꯃꯩꯂꯩ (Sige)	ꯀꯃ (Ga)	ꯃꯩ (Na)
ꯀꯁꯃꯤꯂꯩꯂꯩ (Gigade)	ꯀꯁꯃꯤꯂꯩꯂꯩꯂꯩ (Sidagena)	ꯃꯩꯂꯩꯂꯩ (Su)	ꯀꯃꯩꯂꯩꯂꯩ (Sinabude)	ꯀꯃꯩꯂꯩ (Gide)	ꯀꯃꯩꯂꯩ (Siga)	ꯀꯃꯩꯂꯩ (Sina)
ꯀꯃꯩꯂꯩ (Gina)	ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ (Dudagena)	ꯀꯃꯩꯂꯩꯂꯩ (Gisu)	ꯃꯩꯂꯩꯂꯩꯂꯩ (Dine)	ꯃꯩꯂꯩꯂꯩ (Dage)	ꯃꯩꯂꯩꯂꯩ (Duga)	ꯃꯩꯂꯩꯂꯩ (Duna)
ꯀꯃꯩꯂꯩꯂꯩꯂꯩ (igina)	ꯀꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ (Sigedamak)	ꯀꯃꯩꯂꯩꯂꯩꯂꯩ (Gidasu)	ꯀꯃꯩꯂꯩꯂꯩꯂꯩ (Sidene)	ꯃꯩ (Da)	ꯀꯃꯩꯂꯩꯂꯩ (Gede)	ꯃꯩꯂꯩꯂꯩ (Nade)
ꯃꯩꯂꯩꯂꯩꯂꯩ (Dugina)	ꯀꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ (Gidamak)	ꯃꯩꯂꯩꯂꯩꯂꯩ (Busu)	ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ (Dudene)	ꯀꯃꯩꯂꯩ (Sida)	ꯀꯃꯩꯂꯩꯂꯩ (Sigade)	ꯀꯃꯩꯂꯩꯂꯩ (Sinadi)
ꯀꯃꯩꯂꯩꯂꯩ (Ginade)	ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ (Dugeda)	ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ (Dounade)	ꯃꯩꯂꯩꯂꯩ (Bu)	ꯃꯩꯂꯩꯂꯩ (Duda)	ꯃꯩꯂꯩꯂꯩꯂꯩ (Dugedi)	ꯀꯃꯩꯂꯩꯂꯩꯂꯩ (Sigenade)
ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ (Dugidamak)	ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩꯂꯩ (Dhounabu)	ꯀꯃꯩꯂꯩꯂꯩ (Sibu)	ꯃꯩꯂꯩꯂꯩ (Dade)	ꯀꯃꯩꯂꯩ (Gina)	ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ (Dugenade)	ꯀꯃ (Gi)
ꯀꯃꯩꯂꯩꯂꯩꯂꯩ (Sidouna)	ꯃꯩꯂꯩꯂꯩ (Dubu)	ꯀꯃꯩꯂꯩꯂꯩ (Sidade)	ꯀꯃꯩꯂꯩꯂꯩꯂꯩ (Sigena)	ꯃꯩꯂꯩꯂꯩ (Buna)	ꯀꯃꯩꯂꯩꯂꯩ (Sige)	ꯀꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ (Sidounabu)
ꯃꯩꯂꯩꯂꯩ (Bude)	ꯃꯩꯂꯩꯂꯩꯂꯩ (Dudade)	ꯃꯩꯂꯩꯂꯩꯂꯩ (Dugena)	ꯀꯃꯩꯂꯩꯂꯩꯂꯩ (Sibuna)	ꯃꯩꯂꯩꯂꯩ (Duge)	ꯃꯩ (Na)	ꯀꯃꯩꯂꯩꯂꯩ (Sibude)

Table 7: List of Manipuri Prefixes

ꯀ	ꯃ	ꯃꯩ	ꯃꯩꯂꯩ	ꯃꯩꯂꯩꯂꯩ	ꯃꯩꯂꯩꯂꯩꯂꯩ	ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ	ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩꯂꯩ	ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩꯂꯩꯂꯩ
---	---	----	------	--------	----------	------------	--------------	----------------

Sample output:

ꯀꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ to ꯀꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩꯂꯩ, ꯃꯩꯂꯩꯂꯩꯂꯩ to ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ, ꯃꯩꯂꯩꯂꯩꯂꯩ to ꯃꯩꯂꯩꯂꯩꯂꯩꯂꯩ etc.

2.2.6 Preparation of ambiguous words

Initially Manipuri ambiguous words are prepared from the IndoWordNet data. Those words in IndoWordNet which has more than one sense are ambiguous words. But in this paper, only those word which has the same spelling are considered. Also, from Manipuri Electronic dictionary, the ambiguous data are selected manually by human experts.

A word is said to be ambiguous if one of the following conditions is satisfied:

- a) If a word in the IndoWordNet contains the more than one synonym.
- b) If a word is used in different parts of speech.
- c) If a word has more than on meaning in a dictionary.

Table 8: List of Manipuri ambiguous words

ᱵᱚᱠᱮᱴᱷ	ᱵᱟᱱ	ᱵᱚᱠᱟᱰ	ᱠᱤᱨᱢᱤ	ᱵᱚᱰᱤ	ᱵᱚᱠᱚ
(Laibak)	(Thong)	(Kolom)	(Anganba)	(Epham)	(Guli)
ᱠᱤᱨᱢᱤ	ᱵᱟᱠᱤᱰ	ᱵᱚᱠᱚᱰ	ᱠᱤᱨ	ᱵᱚᱠᱤᱨᱚ	ᱵᱚᱠᱚᱰᱚ
(Angangba)	(Chumlaba)	(Koubru)	(Ahan)	(Ibungo)	(Chakri)
ᱠᱚᱰᱚᱴ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱠᱤᱨᱚᱰᱚ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ
(Asangba)	(Onba)	(Okpa)	(Ayetpa)	(Emai)	(Ritu)
ᱵᱚᱠᱚᱰᱚ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰᱚᱰ	ᱵᱚᱠᱚᱰ
(Purnima)	(Otpa)	(Kei)	(Ennaba)	(Onnateinaba)	(Chaba)
ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰᱚᱰᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ
(Khoi)	(Khong-hamba)	(Mi)	(Enba)	(Erei)	(Epa)
ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ
(Mayon)	(Khong)	(Kot)	(Ening)	(Esen)	(Akiba)
ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ
(Phibam)	(Khon)	(Kon)	(Marup)	(Sanaba)	(Khuya)
ᱵᱚᱠᱚᱰᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ
(Loisinba)	(Kanba)	(Pi)	(Epan)	(Ghee)	(Kheiroi)
ᱵᱚᱠᱚᱰᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ
(Khalliba)	(Kwak)	(Konung)	(Epot)	(E)	(Aengba)
ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	ᱵᱚᱠᱚᱰ	
(Ahoba)	(Kang)	(Angka)	(Epom)	(Gadha)	

2.2.7 SENSE INVENTORY

For Manipuri language, there is almost negligible amount of e-resources. Development of dataset of printed script of Manipuri language have already started, which has become a necessity for such least researched language[8]. Sense Inventory is the database of the Manipuri words that contains the meaning of the word and an example sentence that contain that word. The key principles encompassing sense inventories are clarity, coherence and comprehensive inclusion of the entire spectrum of significant meaning differentiations[1]. The entries in this database are mainly taken from the Manipuri IndoWordNet data available at TDIL data center and the Manipuri Electronic data, which are taken through proper channel. Here, data with single meaning as well as multiple meanings are stored together.

The structure of IndoWordNet’s Manipuri data is shown below:

ID: 8

CAT: ADJECTIVE

CONCEPT: ᱠᱤᱨᱢᱤ ᱵᱚᱠᱚᱰᱚᱰ (Afaba oidaba)

EXAMPLE: ᱵᱚᱠᱚᱰᱚᱰ ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ
(Houdongna lambi lamlanba asi mangol oidaba thoudokni haina lounei)

MANIPURI-SYNSET: ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ ᱵᱚᱠᱚᱰ (Mangol oidaba douyadaba)

Here, MANIPURI-SYNSET is the word(sometimes also contains the synonymous words) for which the details are represented in the above format, ID represent the unique identification number that has been assigned to a word, CAT is the part-of-speech of the word, CONCEPT represent the meaning/sense of the given word and EXAMPLE is the sentence that contains the word.

The database is stored in the following structure for every word:

- a) **ID** – Identification number of a word.
- b) **WORD** – A word.
- c) **POS** – Associated part-of-speech of the word.
- d) **Sense1, Sense2, Sense3** etc – Associated sense(s) of the word
- e) **Example Sentence1, Example Sentence2, Example Sentence3** etc – Associated sentences that contains the word.

Table 9: Sample look of the Sense Inventory

ID	WORD	POS	SENSE1	EXAMPLE SENTENCE1	POS	SENSE2	EXAMPLE SENTENCE2	POS	SENSE3	EXAMPLE SENTENCES3
1	നോ	NOUN	പ്രകൃതി	നോ	NOUN	പ്രകൃതി	നോ	NOUN	പ്രകൃതി	നോ

The Sense Inventory consists of 16351 words, in which 10185, 2024, 332 and 3810 are noun, verb, adverb and adjective respectively.

With respect to Word Sense Disambiguation, we can accumulate information related to the word’s senses from WordNet, including synonyms, glosses, example sentences, hypernyms and meronyms to measures the overlap between the context and the sense bag using intersection similarity, allowing the most probable sense to be determined based on maximum overlap. The efficiency of knowledge-based contextual overlap WSD algorithms using WordNet/IndoWordNet can be increased by the use of diverse glosses, longer glosses, proper nouns, enriched synset structures, frequently used terms, and distributional constraints[5].

3. Results and Discussion

The above preprocessed data can play a vital role in performing many Natural Language Processing tasks such as Machine Translation, Information Retrieval, Word Sense Disambiguation, Sematic Analysis, Question and Answering, Word Sense Induction, Text Classification etc.

Tokenization is the basic preprocessing step in every NLP application. The tokenized data can be used to further study an individual word or a sentence separately. In other word, it will be helpful in performing Morphological study of the Manipuri Language.

When applied to WSD tasks, tokenization gave advantages in terms of granularity, context preservation, simplification of feature selection, computability with language models and evaluation and reproducibility[11,12]. All these benefits gave a path in solving problems of WSD. The tokenized data obtained in this paper will be mainly beneficial to the word level Word Sense Disambiguation. By incorporating spelling correction in the WSD process advantages like enhancing word recognition, expansion of vocabulary coverage, contextual coherence and robustness to a noisy data are achieved.

Performing stemming in Natural Language Processing (NLP) can have advantages for word sense disambiguation tasks. The advantages of incorporating stemming into the word sense disambiguation process includes reduction of lexical variations, improving coverage and recall, dimensionality reduction and improve efficiency[14].

However, it is important to note that stemming is a simplification technique that can lead to loss of information. Stemming may result in the merging of different word senses or the creation of false stems that do not accurately represent the intended meaning[12, 4]. Consequently, stemming should

be applied judiciously and in combination with other techniques to enhance word sense disambiguation accuracy.

The unicode conversion program can be very useful in creating the large corpus of Manipuri Meitei Mayek data, which has very limited electronic data. Bengali script data are available in plenty. The local newspaper data are now the rich source of Bengali script Manipuri data. This program code can convert these Bengali script Manipuri data into Meitei Mayek script data.

Also, Sense Inventory data can correctly translate word(s) from other language to Manipuri language. The same Sense Inventory will act as a sole repository to carry out Word Sense Inventory for Manipuri language. Since, this repository contains all the senses of the Manipuri words, strong and heuristic search can be performed easily and results can be obtained instantly. The same Sense Inventory can be a useful aid in performing the Semantic Analysis, as this repository contains the relationships of word with other words. The relationship includes Hypernymy, Hyponymy, Meronymy, Synonym and Antonyms of the Manipuri Meitei Mayek words.

4. CONCLUSION AND FUTURE WORK

Due to the lack of e-resource of the Manipuri Meitei/Meetei mayek data, lots of processing work was carried out in the available data to bring mentioned data in the machine readable format. Further, to make these machine readable data into a WSD usable data various NLP preprocessing steps like spelling correction, tokenization, stop word removal etc were also carried out. These NLP preprocessing tasks was of very lengthy and time consuming process, which cannot be also omitted. These preprocessing steps are required to yield accurate results and build a promising WSD system for Manipuri language. For research purpose, bigger the size of the corpus better will be the performance of the developing system. Hence, above to these preprocessed data, more and more data can be collected by the future researchers and store into this corpus so as to increase the corpus size and make this corpus the gold standard corpus for the Manipuri language. This work being the first of its own kind for Manipuri language, surely further works can be carried to convert Manipuri language as a limited e-resource to a plentifully available data resource language. The main benefit of this paper is that the Manipuri NLP research team can hand pick up this processed data and use at will to achieve their desired task.

References

- [1] Agirre, E., & Edmonds, P. (2007), Word Sense Disambiguation: Algorithms and Applications, Springer, ISBN 978-1-4020-4808-4
- [2] Yarowsky, D. (1995), Unsupervised word sense disambiguation rivaling supervised methods, Association for Computational Linguistics (ACL).
- [3] Manning, C. D., & Schütze, H.(1999), Foundations of Statistical Natural Language Processing, The MIT Press, ISBN: 0-262-13360-1
- [4] Andrei Mincă and Ștefan Diaconescu(2011), An Approach to Knowledge-Based Word Sense Disambiguation Using Semantic Trees Built on a WordNet Lexicon Network, 978-1-4577-0441-3/11/\$26.00 ©2011 IEEE
- [5] Alok Chakrabarty, Bipul Syam Purkayastha, and Lavya Gavshinde(2010), Knowledge-Based Contextual Overlap keen Ideas for Word Sense Disambiguation using Wordnet, 3rd Indowordnet workshop under the aegis of the 8th International Conference on Natural Language Processing (ICON 2010), 8-11 December 2010, IIT Kharagpur
- [6] Daniel Jurafsky, James H. Martin(2009), Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition, Second Edition, Prentice-Hall, Inc, ISBN: 0131873210



- [7] Chingakham Ponykumar Singh, Dr. H. Mamata Devi, Dr. Nongmaithem Ajith Singh(2021), A Review on Word Sense Disambiguation of different Indian and Foreign languages, IT in Industry, Vol. 9, No.2
- [8] Y. Loijing Khomba Khuman, Dickeeta Salam, Ch. Ponykumar Singh, Dr. H. Mamata Devi, Dr. Nongmaithem Ajith Singh(2022), A benchmark dataset for printed Meitei/Meetei script Character Recognition, Data in Brief, 45 doi: 10.1016/j.dib/2022.108585
- [9] Kishorjit Nongmeikapam, Bishworjit Salam, Makakmayum Romina, Ngariyanbam Mayekleima Chanu, Sivaji Bandyopadhyay(2011), A Light Weight Manipuri Stemmer, Proceedings of NICLC 19-20 February, Cochin, India
- [10] Lesk, M(1986), Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, In Proceedings of the 5th SIGDOC; ACM: New York, NY, USA, 1986; pp. 24–26.
- [11] Steven Bird, Ewan Klein, and Edward Loper, Natural Language Processing with Python, Page 121, ISBN: 978-0-596-51649-9
- [12] Dipanjan Sarkar and Raghav Bali, Text Analytics with Python, pp. 108-112, ISBN-13(pbk): 978-1-4842-2387-1, ISBN-13(electronic): 978-1-4842-2388-8, DOI 10.1007/978-1-4842-2388-8.
- [13] S. Poireiton Meitei, Bipul Syam Purkayastha, H. Mamata Devi(2015), Development of a Manipuri stemmer: A hybrid approach, International Symposium on Advanced Computing and Communication (ISACC)
- [14] Ch. Yashwanta Singh(2000), Manipuri Grammar, Rajesh Publications
- [15] https://en.wikipedia.org/wiki/Meitei_script