ISSN 2007-9737

Little Wins: Collecting, Preparing and Publishing Resources for Assamese Word Sense Disambiguation

Jumi Sarmah[1,*], Anup Kumar Barman[2], Shikhar Kumar Sarma[3]

[1] Faculty of Computer Technology,
Assam down town University, Guwahati,
India
[2] Department of Computer Science and Engineering,
Central Institute of Technology, Kokrajhar,
India
[3] Department of Information Technology,
Gauhati University, Guwahati,
India

jumi.sarmah@adtu.in, ak.barman@cit.ac.in, sks001@gmail.com

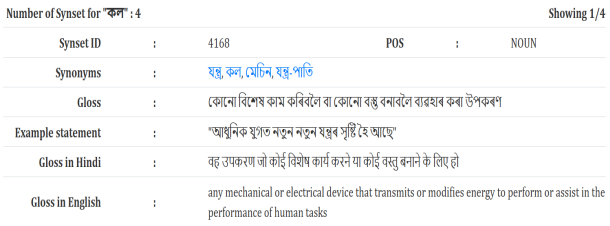
Abstract. This paper presents the creation of the Assamese Ambiguous Sense Inventory (ASI) and Sense Annotated Data Set (SeAnDa) for the Assamese Word Sense Disambiguation (WSD) task. WSD is a computational process that identifies the appropriate sense of an ambiguous term relevant to the context. In this paper, we describe the process of creating ASI and SeAnDa for the implementation of the Assamese Supervised WSD task. The ASI consists of a database of ambiguous terms with their multiple senses, and based on the ASI, a sense-annotated dataset was prepared from the Assamese raw Corpus. The ambiguous terms are extracted from the Assamese WordNet and Corpus. Currently, we have an inventory of 100 ambiguous terms with their various glosses in both Assamese and English, and a sense annotated dataset of minimal size 2K sentences. The authors have analyzed the ambiguous words considering the parameters- Parts of speech and the number of senses. It is reported that most of the ambiguous terms in the inventory are nouns, and most of the terms have binary senses. The ASI and SeAnDa acts as the preliminary resources for implementing the Assamese Supervised WSD task with Iterative learning and Hold-out evaluation strategy. We here adopted and applied the Naïve Bayes Classifier achieving an accuracy of 71%. As Assamese is a computationally low-resourced language, these resources will assist researchers and developers in their future research purpose.

Keywords. Assamese, corpus, assamese sense inventory (ASI), sense annotated data (SeAnDa), low resource, parts of speech, wordNet, word sense disambiguation (WSD), sense annotation, supervised learning

1 Introduction

The Word Sense Disambiguation (WSD) process is identifying the relevant sense of an ambiguous term. It aims is to find the accurate sense si among the set of senses s1, s2, s3……, sn. This intermediate task is important in various NLP applications, such as Information Retrieval, Machine Translation, Question Answering, Speech Processing and Document Classification. Given the high importance of WSD in these NLP applications, we will attempt in implementing the Assamese WSD system through supervised approaches. In this paper, we have mentioned the various supervised models for implementing WSD system for low resource languages.

To develop a supervised WSD system, we first need to identify the ambiguous terms and their multiple senses, so that the appropriate sense can be assigned to the term for further processing and



Number of Synset for "কল" : 4					Showing 1/4
Synset ID	:	4168	POS	:	NOUN
Synonyms	:	যন্ত্র, কল, মেচিন, যন্ত্র-পাতি			
Gloss	:	কোনো বিশেষ কাম কৰিবলৈ বা কোনো বস্তু বনাবলৈ ব্যৱহাৰ কৰা উপকৰণ			
Example statement	:	"আধুনিক যুগত নতুন নতুন যন্ত্ৰৰ সৃষ্টি হৈ আছে"			
Gloss in Hindi	:	वह उपकरण जो कोई विशेष कार्य करने या कोई वस्तु बनाने के लिए हो			
Gloss in English	:	any mechanical or electrical device that transmits or modifies energy to perform or assist in the performance of human tasks			

Fig. 1. A sample view of Assamese WordNet

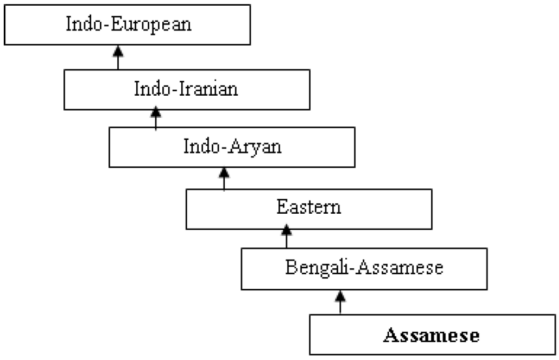


Fig. 2. Language Family of Assamese Language

to prepare training, testing and validation data sets. The sense inventory helps the authors to create SeAnDa. The ASI and SeAnDa act as the basic linguistic resources for developing a supervised WSD system.

Firstly, the authors attempt to extract the ambiguous terms (create the ASI) from the Assamese WordNet and Corpus where the extraction process follows a systematic, step-by-step procedure. Secondly, to create the SeAnDa, we map the Sense Inventory to the raw Corpus. Thirdly, to mark the appropriate sense to the ambiguous term from the set of various senses, the native speaking lexicographers tagged the appropriate sense in the specific context.

'Corpus' refers to a collection of data, either spoken or written, in any language that is stored, managed, and analyzed in digital format. A Corpus is a collection of machine-readable texts represented in a particular natural language and provides a testbed for developing NLP systems. The Assamese Corpus was developed by Gauhati University [15]. It is a collection of 1.5 million words comprising various genres mainly Media, Learned Material and, Literature. For creating the sense-tagged corpus, we took the raw corpus and map with the sense inventory (ASI) to generate the sense-tagged corpus.

The Assamese WordNet [16] is an important resource developed for Assamese NLP. It is a lexical database that has four main components- ID, CAT, Synset, and Gloss. The ID indicates the primary key, consisting of a unique identification number. CAT denotes the Parts of Speech category, Synsets are an array of Synonymous terms that are placed according to the frequency of their use in their respective language and are the basic building blocks of WordNet. The Gloss states the various meanings/(s) of the ambiguous term. Assamese WordNet was constructed based on Hindi WordNet which was a part of the Indo WordNet project.

The statistics of Assamese WordNet show 14,958 entries to date. A pictorial view is shown in Figure 1. Table 1 displays the entry of an ambiguous term- "কল/kol" in Assamese WordNet[1]. The Assamese language (language family shown in Figure 2) is the official language of the Northeastern state of India. It shares a border with Bhutan and Bangladesh and has culture and climate similar to Southeast Asia.

However, some recent development from a technological perspective has ventured into this language. Assamese Information Retrieval (IR) system- CLIA [3], Assamese WordNet, and Corpus (Monolingual, Bi-lingual, POS annotated) are among the important resources developed in the Assamese NLP domain.

The contributions of the article are as follows: i) Contributed an ambiguous sense inventory derived from Assamese WordNet and Corpus consisting of ambiguous terms that serve as an important linguistic resource for the supervised WSD task. It marks the initial resource required for the Supervised Word Sense Disambiguation. ii) Contributed a Sense annotated dataset for training, testing and validation of implementing Assamese WSD. For training, incremental learning

[1] www.cfilt.iitb.ac.in/indowordnet/wordnet?langno=2&query=%E0%A6%95%E0%A6%A6%E0%A6%B2%E0%A7%80

strategy is applied, as the sense annotated data is minimal.

The paper is organized as follows.

Section 2 gives a survey of the various methodologies used to implement WSD, including information about sense inventories, training and testing data.

Section 3 describes the detailed process of extracting ambiguous terms from WordNet and the Assamese Corpus through an algorithm. Additionally, this section provides an analysis of parameters such as the number of senses and parts of speech (POS).

Section 4 explains the process of generating a sense-annotated dataset by mapping/linking it with the sense inventory and the raw corpus. It gives us a glimpse of the Assamese sense annotated dataset.

Section 5 mentions the process pipeline of Supervised Assamese Word Sense Disambiguation (WSD) model, mentioning its evaluation.

Finally, the last, Section 6 concludes the paper by highlighting the published sense inventory and sense-tagged data, which is made available on the GitHub platform for future research purposes through this research work.

2 Related Work

This work is inspired by various studies surveying supervised approaches for developing WSD systems for the Assamese Language. AI and NLP have witnessed a drastic change, and one of the most important NLP applications is sense disambiguation. Several works and surveys on WSD are mentioned in [1, 2, 4, 6, 8].

Important resources for WSD include sense inventories and sense-annotated data, as discussed by the authors of [5]. Sense Inventories like Princeton WordNet [9], BabelNet [11], and Wikidictionary1 are essential prerequisites for WSD tasks. In case of Sense Annotated Data sets, SemCor [10] is the largest manually annotated dataset, comprising 200,000 sense annotations using the WordNet sense inventory.

2.1 Data for Training

While English training data is widely available, the same does not hold for languages like Indo Aryan Language Assamese. Although hand-labeled data are difficult to obtain on a large scale for many languages, some efforts in the past were directed towards creating manually translated versions of SemCor [13], but many of these are no longer available. Therefore, several subsequent works proposed automatic methods for producing high-quality sense annotated data both in English [19, 7] and other languages by leveraging: information from Wikipedia [18]. We have developed an Assamese ambiguous word sense inventory (ASI) extracted from the Assamese WordNet and Corpus.

To retrieve ambiguous terms from the Corpus, manual intervention was required, as mentioned in section 3.

2.2 Data for Testing

Evaluation in WSD is generally carried out using manually annotated datasets from the Senseval and SemEval evaluation campaigns. English WSD benefits from the evaluation suite of [14]. Only very recently, a comprehensive benchmark has been put forward to standardize the evaluation in this setting mentioned by the author [12]. A WSD model has been developed for the Assamese Language, and the results with state of art accuracy is achieved and mentioned by the authors [17]. However, this paper mainly focusses on the detailed extraction process of ambiguos terms from Assamese WordNet, Corpus and the resource availability- both ASI and SeAnDa in the Github platform for further research work in this NLP application.

3 Ambiguity Extraction from WordNet and Corpus

3.1 Extraction from Assamese WordNet

We applied a small algorithm to extract the ambiguous terms, i.e., those with more than one sense. A current list of 100 ambiguous words was extracted and validated. A simple

Table 1. "কল"/kol entry in Assamese WordNet

Word: কল **(kol)**	
Synset Id: 4168	**Synset Id: 4505**
POS: Noun	**POS: Noun**
Gloss: কোনো নির্দিষ্ট কার্য কৰিবৰ বাবে কোনো বস্তুৰ ব্যৱহাৰ কৰা উপকৰণ	**Gloss:** এবিধ দীঘলীয়া মঙহাল মিঠা ফল
Example sentence: আধুনিক যুগত নতুন নতুন যন্ত্রৰ সৃষ্টি হৈ আছে	**Example sentence:** তেওঁ কল খাই আছে

computer program was written in Java to extract the ambiguous terms from the Assamese WordNet universal file, namely –"unv.syns".

We use the following algorithm.

— Step.1: The program checks whether each term in an array of synonyms also appears in other synonym arrays.

— Step.2: If yes, a database is created with those terms and their corresponding Assamese and English senses are stored.

— Step.3: Assamese native speakers performed manual validation to verify if all the terms and their senses were distinct and accurate. Currently, the homonyms (having distinct senses) were considered in the Assamese sense inventory.

— Step.4: If the terms has no distinct senses, the term is excluded, and the database is updated. Finally, a repository consisting of the ambiguous terms and their senses was created from Assamese WordNet. The system flow diagram is shown in Figure 3. Below. A sample of ASI and a database of ambiguous terms with POS, is shown in Table 2 and Table 3 respectively.

For processing derivations we use the following ideas.

— It is found that most of ambiguous terms have the noun POS. Thus, it is observed that nouns dominate ambiguous words, as they represent the most tangible and abstract entities, shown in Table 4.

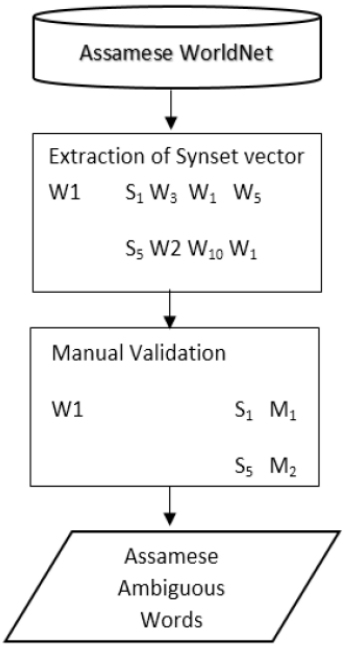


Fig. 3. Flowchart for extracting ambiguous terms from WordNet

— Most of the ambiguous terms have binary senses. Binary senses indicate that most of the terms have two primary meanings. In Computational linguistics research, it is found that most of the WSD algorithms work well when disambiguating terms having binary sense as compared to having multiple senses. It is found that the highest number of senses is seven for the ambiguous term "জাল(jal)" out of the 100 listed ambiguous terms.

3.2 Extraction from Assamese Corpus

The below diagram (Figure 4) shows how ambiguous words were mined from the corpus. Some preprocessing steps were applied before ambiguous words were derived from the corpus. They are mentioned step-by-step below.

The Pre-processing Steps are as follows. The corpus underwent a cleaning process to remove unwanted characters, junk values, and extraneous symbols such as punctuation marks (e.g., ’, “”, ? , , [], !, -) which could otherwise introduce errors during post-processing. Short sentences comprising only two or three words were filtered out, as they convey minimal semantic information and are of limited utility for subsequent analysis. Inconsistent data, including typographical errors, were manually corrected to enhance the quality and reliability of the dataset.

Table 2. ASI for "অর্থ/Artho"

Term	Assamese senses	English senses
অর্থ	টকা পইচা	Money
	শব্দ বা বাক্য এটা বুজায়	Meaning of a Word

Table 3. POS-wise ambiguous term

Ambiguous Terms	POS
অংক	Noun
অজান	Adjective
অঞ্চল	Noun
অথর্ব	Adjective
অনা	Verb
অনুচ্ছেদ	Noun
অন্তৰ	Noun/Adjective
অপ্রস্তুত	Adjective
অর্থ	Noun
আই	Pronoun
আচল	Noun/Adjective
চাৰ	Noun
আজান	Noun

From the cleaned corpus, individual sentences were extracted and assigned unique numeric identifiers. The sentences were tokenized using the punctuation mark ’|’, which denotes sentence

Table 4. POS-wise sense distribution table

POS	NO OF SENSES
Noun	56
Adjective	15
Noun/Adjective	7
Pronoun	1
Verb/Noun	2
Verb/Interjection	1
Verb	8
Adverb	0
Adverb/Verb	1
Noun/Proper Noun	1
Proper Noun	5
Noun/Adverb	1
Pronoun/Adjective	1
Interjection	1

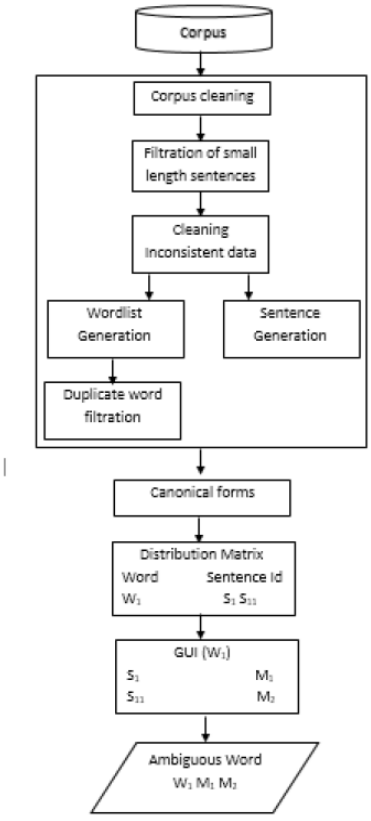


Fig. 4. Flowchart for deriving ambiguous terms from the corpus

boundaries in the Assamese language, and redundant copies of the generated sentence files were prepared for use in later stages. Subsequently, a word list was automatically generated by tokenizing one of the sentence files at spaces (‘ ’), which separate words in Assamese.

Duplicate entries in the word list were removed to eliminate redundant computations, resulting in a unique set of words. To further refine the dataset, canonicalization techniques were applied. This included stop-word removal, wherein 107 commonly occurring Assamese words such as aaru- আৰু (and) and baa- বা (or) were filtered out using a custom Java program, and stemming, which reduced inflected words to their root forms. Finally, morphological analysis was performed on both the sentence and word list files to analyze and process the structural properties of words effectively.

3.2.1 Processing Steps

The processing steps are as follows.

Distribution Matrix formation: A distribution matrix shown in Figure 5 is formed by considering the word list and mapping with the words of the sentences generated from the corpus. Those words which are matched to the words of the Sentence-Ids are denoted by a true value denoted by 1 in that row form. Multiple true values in the column of sentence-Id means the root word matches are present in a particular sentence. The column indicates the occurrence of words in multiple sentences. Row indicates the occurrence of multiple words in different sentences.

$$\begin{bmatrix} & W_1 & W_2 & W_3 & ..... & W_m \\ S_1 & 1 & 0 & 1 & & 1 \\ S_2 & 0 & 1 & 0 & & 0 \\ S_3 & 1 & 0 & 0 & & 1 \\ .. & .. & .. & .. & & .. \\ .. & .. & .. & .. & & .. \\ S_n & 1 & 0 & 0 & & 1 \end{bmatrix}$$

Fig. 5. The distribution matrix

Assamese Sense-tagging GUI: A GUI (Figure 6) below is developed for annotating sense to the words respective to their occurrence in sentences. Manual annotation with the help of a GUI (Graphical User Interface) helps in implementing this task. Those words which have unique distinct meanings are only considered as ambiguous corresponding to their use in sentences.

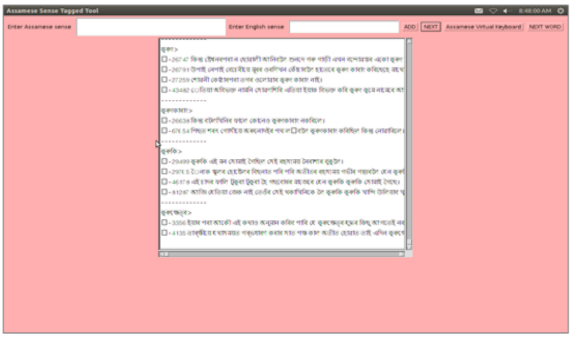


Fig. 6. Snapshot view of manual annotation UI

4 Sense Annotated Data Set: SeAnDa

We tried mapping raw corpus with sense inventory (general form) as shown in Figure 7. A sample of the mapping raw corpus with inventory is shown below in Figure 8. After mapping with the raw corpus the Lexicographers tagged the appropriate sense to the ambiguous term.

$$\begin{bmatrix} w1 & s1 & s2 & s3 \\ w2 & s1 & s2 & \square \\ w3 & s1 & s2 & \square \end{bmatrix} \begin{bmatrix} t1 & w1\{s1,s2,s3\} & t2 & t3 \\ t4 & t5 & t6 & t7 \\ t8 & t9 & t10 & w2\{s1,s2\} \\ t11 & t12 & w3\{s1,s2\} & t13 \end{bmatrix}$$

Fig. 7. Mapping of words with senses

The sample of Assamese SeAnDa and ASI can be found in the below platform: https://github.com/jumi123/Assamese-Sense-Inventory.git.

The sense annotated data set serves as training data in the Supervised ML task of Assamese WSD. As we are at the starting level of developing the WSD system and the size of SeAnDa is minimal,

we followed incremental learning strategy in the process of learning the Assamese WSD system.

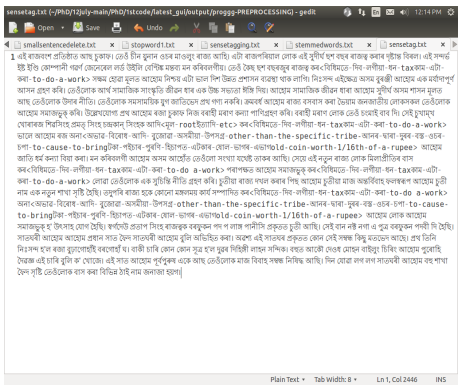


Fig. 8. Linking sense inventory with corpus

5 Assamese WSD model and Evaluation

This section demonstrates the process of the supervised approach to solve the WSD task for the Assamese language. Assigning proper sense to an ambiguous word is the aim of the WSD task. This task involves selecting a true sense from a set of senses say S = s1, s2, s3 of an ambiguous word depending on the particular context where it is applied. This is the generic concept behind WSD as already mentioned in this research work. The Hold-out evaluation strategy is followed to evaluate the accuracy of the system. An accuracy of 71% is achieved while implementing the WSD task. The methodology is mentioned and the system architecture is shown in Figure 9.

5.1 Methodology

5.1.1 Data Set Preparation: Trained & Test Data Set

A supervised approach needs a training data set for the model learning purpose and sense tagging the unseen input. Raw data were collected from various sources- Corpus, Web data, WordNet example sentences. Raw here denotes un-annotated data. The only purpose of collecting data was preparing sense annotated data set as no sense annotated data is available digitally for this Indo-Aryan language. From Corpus, WebData using Nutch1.4 and WorNet example sentences we collected the raw data content in Assamese language.

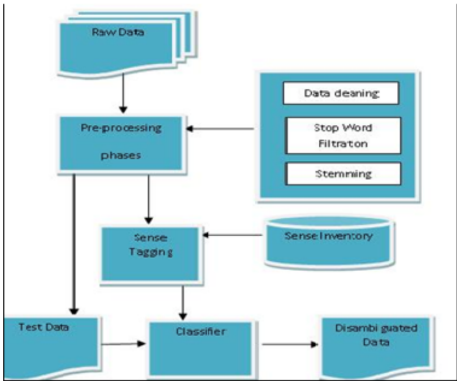


Fig. 9. System Architecture of Assamese WSD

5.1.2 Pre-processing phases-Stop-word filtration and Stemming

Stop-words were filtered and stemming was performed on the terms in the context of both trained and test data set. Also, filtration of small length sentences (size<=3) are removed.

5.1.3 Sense Tagging

As for example-sentence: The raw data contains the ambiguous words along-with their sense definition by mapping to the Ambiguous word sense inventory through a program written in JAVA "সংগীত, নাটক, চিত্রকলা আদি [১. মূল (root), ২. ইত্যাদি (etc.)] হৈছে কলা [১. শিল্প (arts, culture), ২.কাণেৰে নুশুনা (deaf), ৩. কোষ (cell)]" The sense inventory helps the manual annotator in assigning the appropriate sense to the ambiguous word. After manual tagging the sentence above becomes: "সংগীত, নাটক, চিত্রকলা আদি [ইত্যাদি (etc.)] হৈছে কলা [শিল্প (arts, culture)]" A total of 50 ambiguous words were retrieved from WordNet out of a 15K number of synsets and a number of 60 ambiguous words are formed from total words of 18K from the Corpus. We found the ambiguous word having Noun POS occurs 78% in the total trained data. As

ambiguous word like আদি(Adi) with sense ইত্যাদি (Ityadi) occurred many more times in the tagging corpus by which the model would be over learned with the ambiguous word so it was manually removed. Currently, 2K sentences containing ambiguous words are sense tagged. The size of the training data is small in our case for the reason found- As the sense inventory is small it may not cover the ambiguous words.

5.1.4 Selection of Features

Features provide significant importance in detecting the sense of ambiguous words with respect to the context. They are selected from the filtered sense tagged data so that the classifier is trained with those features. The un-ordered semantic feature surrounding the ambiguous word at a specific position is considered for our work. The features consist of the target word (w) and three words to left position and three words to the right position with respect to the ambiguous word. These neighboring features give a strong clue for disambiguation task.

However, it is found that the target word or ambiguous word occurring at the beginning and at the end position in a sentence gets lesser clues for disambiguation as the features considered in that cases are only the features that occurred in left or right position respectively.

Say, such as in the Assamese sentence:জোতা পিন্ধা, দোলাত উঠা আদি [ইত্যাদি – etc.] সকলোতে নিম্নশ্রেণীৰ লোকৰ বাধা আৰোপ কৰা হৈছিল।" (Eng form: poor people's were restricted from tasks like wearing shoes etc). The word আদি(Adi) is an ambiguous word in Assamese language. The features of this word by considering the offset (-2, -1, 0, 1, 2) are: -2(দোলা),-1(উঠা), আদ<ইত্যাদি -etc>+1(সকলোতে), +2(নিম্নশ্রেণী) With these features, the final sense-tagged data/ trained data set is prepared

5.1.5 Smoothing

It refers to the technique of assigning a value to the unseen event by the Likelihood generator while estimating posterior probability. The purpose is to assign a value to improve the accuracy of the probability estimator. If a feature corresponding to a sense is not available in training data than zero probability will hamper in the classification task. A simple smoothing technique is used by assigning a low value .001 to the probability of an unseen event.

After smoothing, a value to the unseen event is assigned and the probability of getting zero value can be avoided and test vector will be assigned a sense.

5.1.6 Feeding the model and learning

At first, the features are selected accordingly to be fed to the algorithm selected for disambiguation task. They are fed so that the machine gets learned or trained with those features and help in disambiguating sense of the words having more than one meaning.

As a child or a student learns only when sufficient training is provided to them with proper illustrations and examples, in the same way, a machine gets to learn when it is trained with appropriate features. The input provided or fed to the model must meet the need of the algorithm.

5.1.7 Testing Model

A set of total of 800 sentences were also collected from resources like Corpus, Web content, Text materials are taken for testing the classifier. The preprocessing phases mentioned above were also sequentially performed on the test set.

5.1.8 Hold-out Evaluation and Iterative Learning

This phase is an essential part of any model development process. Hold-out evaluation strategy follows the principle- we divide the entire data set into two independent samples- Training and testing data. More the training data better is the derived model and more the test data better is the accuracy of error estimation of the system. Basically, 80% is treated as the trained set and 20% as the test set of the whole data set.

Iterative Learning: This learning paradigm literally means to adopt the hold-out evaluation strategy in an iterative method way. It is the process of repetition where the new model's output

is influenced by the performance of the previous built.

5.1.9 Implementation method

The system was learned with an initial set of training data at the first iteration, then increasing the training data subsequently along with the learned model generated in the previous iteration. It is found that as the training data size increases gradually there is a change in the result accuracy. The training set of n instance is sampled and is trained n times. The testing set is sampled to n different instances that don't occur in the training set to avoid overlapping.

The training and testing set was partitioned into four sets. Starting with the trained set of 1200 sentences and a test set of 100 sentences 57% correct results were retrieved. A change in the classifier output was observed when the trained data set was increased to 2200 and 250 test sentences. Satisfactory result outcomes with a percentage of 71% on testing the classifier with 300 test sentences on sense trained data of size 2.7K. The changes in the accuracy results are shown in Figure 10. It basically relies on the conditional probability of each sense si of an ambiguous word given the features (V) in the given context. The sense si which maximizes the below formula is considered as the appropriate sense of the ambiguous word:

$$s_i = \arg\max_{s_i \in S} P(s_i \mid V). \tag{1}$$

In Equation (1), S = the set of senses for the target word, si = a sense, V = the vector representation of the features, P(si|V) = the posterior probability of class (s, target) given predictor (V, attributes). It is the probability of instance V being in class si which has to be computed.

Applying the Bayes Rule on above equation we get

$$s_i = \arg\max_{s_i \in S} \frac{P(V \mid s_i) * P(s_i)}{P(V)}. \tag{2}$$

In Equation (2), the denominator is constant so it can be left out from the equation. P(si) denotes the prior probability of occurrence of sense si and it is estimated by counting the number of occurrence of the sense in the context and dividing by the total number of instances and shown in Equation (3). P(V|si) is the likelihood which is the probability of generating instance V given the class si:

$$P(s_i) = \frac{C(s_i)}{C(w)}. \tag{3}$$

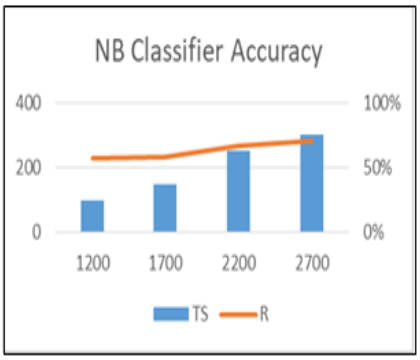


Fig. 10. System Architecture of Assamese WSD

6 Conclusion and Future Work

This paper concludes by presenting an archive of ambiguous terms systematically extracted/collected from Assamese WordNet and the Corpus and published in Github platform to assist the researchers of NLP and Computational Linguistics. These terms are accompanied by their senses/meanings forming a detailed sense Ambiguous sense Inventory (ASI).

The sense inventory serves as a crucial tool for developing a sense-annotated dataset. The annotated dataset is created using the sense inventory and the raw corpus. The raw corpus was joined with the SI where the ambiguous term is detected and the native speaker map the appropriate sense to the term.

This sense annotated dataset was then subsequently utilized to build a supervised WSD model trained with Naïve Bayes algorithm achieving state-of-the-art accuracy.

In conclusion, the development of ASI and the SeAnDa dataset significantly advances the field of WSD offering valuable resources and scopes for

developers and researchers. These initiatives aim to enhance semantic understanding and support the creation of resources for various Assamese NLP tasks.

Acknowledgments

We gratefully acknowledge the Visvesvaraya PhD Fellowship funded by MeitY for supporting the research work, and the research members of the NLP Lab, Department of IT, Gauhati University, for providing the data resources and their valuable suggestions and feedback.

References

1. **Barba, E., Pasini, T., Navigli, R. (2021).** Esc: Redesigning wsd with extractive sense comprehension. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4661–4672.
2. **Barba, E., Procopio, L., Navigli, R. (2021).** Consec: Word sense disambiguation as continuous sense comprehension. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1492–1503.
3. **Barman, A. K., Sarmah, J., Sarma, S. K. (2019).** Developing assamese information retrieval system considering nlp techniques: An attempt for a low-resourced language. ADBU-Journal of Engineering Technology, Vol. 8, No. 2.
4. **Bevilacqua, M., Navigli, R. (2020).** Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2854–2864.
5. **Bevilacqua, M., Pasini, T., Raganato, A., Navigli, R. (2021).** Recent trends in word sense disambiguation: A survey. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21) Survey Track.
6. **Conia, S., Navigli, R. (2021).** Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 3269–3275.
7. **Loureiro, D., Camacho-Collados, J. (2020).** Don't neglect the obvious: On the role of unambiguous words in word sense disambiguation. Proceedings of EMNLP, pp. 3514–3520.
8. **Maru, M., Conia, S., Bevilacqua, M., Navigli, R. (2022).** Nibbling at the hard core of word sense disambiguation. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4724–4737.
9. **Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J. (1990).** Introduction to wordnet: An online lexical database. International Journal of Lexicography, pp. 235–244.
10. **Miller, G. A., Leacock, C., Tengi, R., Bunker, R. T. (1993).** A semantic concordance. Human Language Technology.
11. **Navigli, R., Ponzetto, S. P. (2012).** Babelnet: The automatic construction, evaluation, and application of a wide-coverage multilingual semantic network. Artificial Intelligence, pp. 217–250.
12. **Pasini, T., Raganato, A., Navigli, R. (2021).** Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. Proceedings of AAAI.
13. **Petrolito, T., Bond, F. (2014).** A survey of wordnet annotated corpora. Proceedings of GWNC.
14. **Raganato, A., Camacho-Collados, J., Navigli, R. (2017).** Word sense disambiguation: A unified evaluation framework and empirical comparison. Proceedings of EACL, pp. 99–110.
15. **Sarma, S. K., Bharali, H. (2012).** A structured approach for building assamese corpus: Insights, applications, and challenges. Proceed-

ings of the 10th Workshop on Asian Language Resources, The COLING 2012 Organizing Committee, Mumbai, India, pp. 21–28.

16. **Sarma, S. K., Gogoi, M. (2010).** Foundation and structure of developing assamese wordnet. Proceedings of the 5th International Conference of the Global WordNet Association.
17. **Sarmah, J., Sarma, S. K. (2016).** Word sense disambiguation for assamese. Proceedings of IEEE 6th International Conference on Advanced Computing (IACC).
18. **Scarlini, B., Pasini, T., Navigli, R. (2020).** With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. Proceedings of EMNLP, pp. 3528–3539.
19. **Taghipour, K., Ng, H. T. (2015).** One million sense-tagged instances for word sense disambiguation and induction. Proceedings of CoNLL, pp. 338–344.

Article received on 27/05/2025; accepted on 05/09/2025.
**Corresponding author is Jumi Sarmah.*