Word Sense Disambiguation applied to Assamese-Hindi Bilingual Statistical Machine Translation

Anup Kumar Barman
Department of Computer Science & Engineering, Central Institute of Technology Kokrajhar, India
ak.barman@cit.ac.in

Jumi Sarmah
Sarala Birla Gyan Jyoti Guwahati, India
jumis884@gmail.com

Subungshri Basumatary
Department of Computer Science & Engineering, Central Institute of Technology Kokrajhar, India
ph19cse1911@cit.ac.in (corresponding author)

Amitava Nag
Department of Computer Science & Engineering, Central Institute of Technology Kokrajhar, India
amitava.nag@cit.ac.in

Received: 31 August 2023 | Revised: 8 October 2023 | Accepted: 16 October 2023

 DOI: https://doi.org/10.48084/etasr.6342

ABSTRACT

Word Sense Disambiguation (WSD) is concerned with automatically assigning the appropriate sense to an ambiguous word. WSD is an important task and plays a crucial role in many Natural Language Processing (NLP) applications. A Statistical Machine Translation (SMT) system translates a source into a target language based on phrase-based statistical translation. MT plays a crucial role in a WSD system, as a source language word may be associated with multiple translations in the target language. This study aims to apply WSD to the input of the MT system to enhance the disambiguation output. Hindi WordNet was used by selecting the most frequent synonym to obtain the most accurate translation. This study also compared Naïve Bayes (NB) and Decision Tree (DT) to test and build a WSD model. NB was more appropriate for the WSD task than DT when evaluated in the Weka machine learning toolkit. To the best of our knowledge, no such work has been carried out yet for the Assamese Indo-Aryan language. The applied WSD achieved better results than the baseline MT system without embedding the WSD module. The results were analyzed by linguist scholars. Furthermore, the Assamese-Hindi transliteration system was merged with the baseline MT system for the translation of proper nouns. This study marks a remarkable contribution to Assamese NLP, which is a low computationally aware Indian language.

Keywords-word sense disambiguation; machine translation; machine learning; assamese; natural language processing

I. INTRODUCTION

Any natural language contains words that have different meanings when used in various contexts. These words are termed ambiguous words, and Word Sense Disambiguation (SWD) is used to resolve lexical ambiguities based on the context. Ambiguity occurs in various NLP phases, such as lexical, syntactic, semantic, pragmatic, and discourse levels [1]. This study aims to disambiguate lexical semantic ambiguities, such as:

- সীতা এজনী কলা/artপেমী ছোৱালী (*Sita is an artistic girl*)
- ছোৱালীজনী জন্মৰে পৰা কলা/deaf sense (*The girl is deaf from her birth*).

Both of the above sentences are syntactically correct but the semantics or meaning of the lexicon "কলা"/kola are different in both sentences. WSD is initialized with the following main requirements: sense repository, document representation, and approach identification [2]. Many supervised, unsupervised, and semi-supervised learning approaches are adopted,

considering a machine-readable dictionary as a sense repository and representing the document with important features for the WSD task. Machine Learning (ML) approaches are widely used for sense classification tasks. Supervised WSD approaches use a large number of algorithms and achieve high accuracies [3]. The algorithms are trained on sense-annotated datasets and are tested on some unseen events. This study adopted supervised ML approaches, Naïve Bayes (NB) and Decision Tree (DT), for a word sense disambiguation task. The NB and DT algorithms were also used in [4-5] for word sense disambiguation tasks. A comparative analysis was performed on DT and NB, and since NB had a lower error rate than DT, the NB classifier was used for the proposed WSD method.

A Machine Translation (MT) system automatically translates a source into a target language. The rapidly increasing rate of digitalized documents requires an automatic MT system, as every human wants to retrieve information in an understandable language. Rule-based, statistical, and hybrid approaches are used to develop automated MT systems. Statistical translation systems, such as Moses, allow any language pair to develop automatic translation models. This study chose a statistical approach for the Assamese-Hindi MT system, which required training of several bilingual parameters in both the source and target languages.

Assamese is the official language of the northeastern state of India,. Assam. It is spoken by nearly 13-14 million people. Assam shares an international border with Bhutan and Bangladesh. It additionally shares a culture and climate similar to Southeast Asia. It is a computationally less aware language that belongs to the Indo-Aryan language family. Recently, some developments have been made for this Assamese language from a technological perspective. Information Retrieval (IR), Assamese WordNet, and Corpus [5], which are important resources for Assamese NLP, were developed. Assamese WordNet was framed in [6]. Based on Hindi WordNet [7], the Assamese WordNet structure was developed by lexicographers as part of the Indo-WordNet project. WordNet is a combination of four main components: ID, CAT, SYNSETS, and GLOSS. ID is the unique identification number, CAT indicates the category or parts of speech, SYNSET is the basic building block of WordNet, which has a vector of synonymous words, and GLOSS defines the meaning.

In [8], a combination of statistical approaches and a WordNet-based SMT system was investigated [8]. This study used Assamese WordNet to retrieve the most appropriate translated word in the English-Assamese SMT system. This lexical resource has not been used in any other study of the Assamese language. Furthermore, the Assamese-English transliteration system was combined with the baseline system for proper nouns and achieved acceptable accuracy when analyzed by linguistic scholars. In [9], another SMT system was presented for Assamese to Bengali. Both Assamese and Bengali share the same code letters, the only difference being the symbolic representation of the letter 'ৰ' (Assamese letter Ra) and the individual letter 'ক্ষ' (Assamese letter Khya). The system was tested using the Bi-Lingual Evaluation Understudy (BLEU) score and achieved satisfactory results. This study aims to integrate the WSD module into the baseline Assamese-Hindi MT system. To our knowledge, no other study has attempted it before in this Indo-Aryan language. Therefore, this study marks a great contribution to Assamese NLP.

II. THE PROPOSED APPROACH

For the WSD task, a table representing the average F-scores of both the NB and DT classifiers is shown when implemented on Weka. The C4.5 DT algorithm and the NB classifier were used to compare, choose, and build an appropriate WSD model. Figure 1 shows the system architecture.

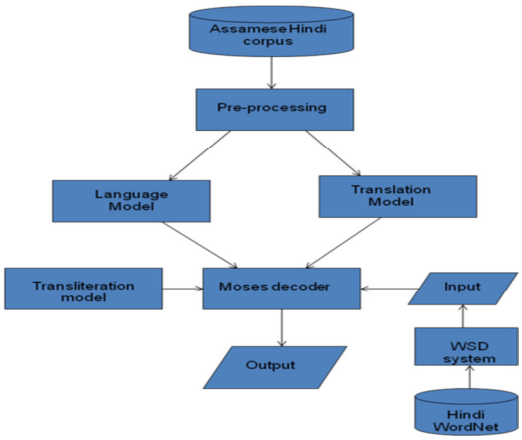

Fig. 1. Architecture of the proposed method.

A. Preparation of Parallel Corpus

Digitized Assamese-Hindi documents are few, so a parallel corpus of Assamese-Hindi was prepared by lexicographers from the GU NLP team. The Assamese-Hindi corpus contains 10K records. Only 3.6K records in the parallel corpus contain ambiguous terms. Some ambiguous terms have only one sense associated with the set of the parallel corpus. Among the 3.6K records, 300 instances were regarded as test data to evaluate the MT system. However, the whole 9.7K data was viewed as train data, as the MT system needs to translate all other words in the input other than the ambiguous term. A hold-out evaluation strategy, splitting the whole dataset into train and test data, was considered to evaluate the MT system. Various preprocessing phases, such as corpus cleaning and removal of junk values and blank spaces, were performed. As the tool was implemented in a Linux environment, the encoding of the parallel files was converted from Windows to Linux. It was also verified whether the source language in the parallel file was aligned with the same translated target lines or not.

B. Developing the Language Model

Given a string of words $W = w1, w2, \ldots, wn$, $P(w)$ according to Markov chain is given by:

$$P(w) = P(w1P(w2|w1)P(w3|w2)P(w4|w3)P(wn|wn-1)$$

This computation is for a 2-gram language model. The maximum likelihood estimation is given by:

$$P(w2|w1) = count(w1, w2)/count(w1)$$

Thus, a language model gives the probability of a sentence using the *n*-gram model. SRILM is a toolkit for building statistical language models, primarily for use in statistical tagging. This tool has been in development by the SRI Speech Technology and Research Laboratory since 1995. SRILM consists mainly of a set of C++ class libraries that implement language models, supporting data structures, and some utility functions [10]. This study used the SRILM toolkit to develop a language model for the Hindi corpus. The command executed in the Linux terminal was:

```
/media/sandhan/1214a349-753f-4189-83ed-
1024026066ae/home/salma/moses/srilm-1.7.1/bin/i686-
gcc4/ngram-count -order 3 -interpolate discount -unk-text
../corpus/train.clean.hi -lm surf1.lm;
```

C. Developing the Translation Model

This model helps to count the conditional probabilities of source and target language training pairs by the formula *P*(*S*/*T*). The computation was carried out by breaking the sentence into words or phrases and their probabilities were mastered. GIZA-PP [11] is a statistical MT toolkit that is used to train word classes using maximum-likelihood criteria. This generates a moses.ini file in the /model directory [12]:

```
/media/sandhan/1214a349-753f-4189-83ed-
1024026066ae/home/salma/moses/mosesdecoder/scripts/training/
train-model.perl/media/sandhan/1214a349-753f-4189-
83ed1024026066ae/home/salma/moses/mosesdecoder/corpus/--
corpus train.clean --e hi --f ass --
lm0:3:/media/sandhan/1214a349-753f-4189-83ed
1024026066ae/home/salma/moses/mosesdecoder/languageasshi/sur
f1.lm:0 -reordering distance,msd- bidirectional-fe -
external-bin-dir /media/sandhan/1214a349-753f-4189-83ed
1024026066ae/home/salma/moses/mosesdecoder/tools;
```

D. Baseline SMT System using the Moses Decoder

This phase of SMT maximizes the probability of translated text corresponding to the source text. The target word/phrase is selected if it maximizes the following formula:

$$P(S|T) = argmaxP(T)P(S|T)$$

Inputting the following command in the terminal, the output was:

```
echo "গৰম আৰু শীতকালত" ./moses-f
../working/ashi/model/moses.ini
```

TABLE I. OUTPUT OF BASELINE TRANSLATION

Sl no	Assamese sentences	English Translation	Hindi sentences
1	ইয়াত চৰাই আৰু জন্তু আছে (*Iyat sorai aru jantu ase*)	Here there are birds and animals	यहां पक्षी और जानवर है
2	আৰব এজন কলা ল'ৰা হয় (*Aarab ajon kola lora hoi*)	Aarab is a deaf boy	अरब एक बहरा लड़का है
3	তাজ মহোৎসৱ ভাৰতত হয় (*Taj mahotsav bharatot hoi*)	Taj festival is held in India	ताज महोत्सव भारत है
4	জামু পাহাৰৰ ওপৰত (Jammu paharor uporot)	Jammu is in the hills	जम्मू पहाडी के ऊपर
5	নৃত্য এটা কলা হয় (*Nitya eta kola hoi*)	Dance is an art	नृत्य एक कला है

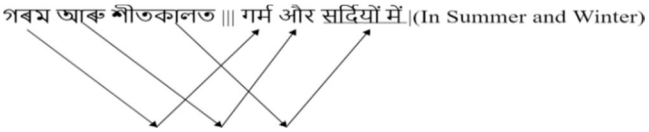

Fig. 2. Baseline translation output.

E. Assamese-Hindi Transliteration System

For the proper nouns, such as the one in the above table in Assamese sentence (2), আৰব/aarab was translated to আৰব by SMT. Thus, a list of parallel proper nouns (Assamese-Hindi) was prepared. Several named entities were prepared and translated by linguistic scholars, helping to some extent in the translation of Out-Of-Vocabulary (OOV) words. A look-up-based approach was used for the transliteration task.

TABLE II. OUTPUT OF TRANSLITERATION SYSTEM

S/N	Assamese sentences	Hindi sentences
1	ইয়াত চৰাই আৰু জন্তু আছে (*Iyat sorai aru jantu ase*)	यहां पक्षी और जानवर है
2	আৰব এজন কলা ল'ৰা হয় (*Aarab ajon kola lora hoi*)	अरब एक बहरा लड़का है
3	তাজ মহোৎসৱ ভাৰতত হয় (*Taj mahotsav bharatot hohoi*)	ताज महोत्सव भारत है
4	জামু পাহাৰৰ ওপৰত (*Jammu paharor uporot*)	जम्मू पहाडी के ऊपर
5	নৃত্য এটা কলা হয় (*Nitya eta kola hoi*)	नृत्य एक बहरा है

F. WSD and WordNet Embedded System

1) Comparative Analysis between NB and C4.5 DT to Build the WSD Model

The sense inventory was prepared by extracting ambiguous terms from both Assamese WordNet and Corpus. The number of unique ambiguous words in the sense inventory was 110, of which 50 (from 15K words from Assamese WordNet) and 60 (18K unique words from Assamese Corpus when mapped to 50K sentences) are derived to date. The senses/meanings were manually put in the database by the lexicographers for the ambiguous words derived from the Assamese corpus. The size of Assamese sense-tagged data was 4K, based on the sense inventory and considering unordered semantic features {-3, -2, -1, +1, +2, +3}. In supervised approaches, it is common to have high dimensional data, but sometimes it is difficult to collect large datasets or examples to represent each pattern or object class of an ambiguous word. The same is the case for the Assamese language. Data containing ambiguous terms of different senses were collected from digitalized content of Assamese Corpus and crawled data from a set of Assamese websites using Nutch1.4. The test involved ten different ambiguous words with 50 instances each of ambiguous words, and the dataset was unbalanced. The comparative experiments were carried out in the Weka ML toolkit. Table V shows the results of the average F-scores of seven ambiguous words. K-fold cross-validation (k = 10) was used, as the sense-tagged data was small. The NB classifier had a lower error rate than

DT. The low performance of DT for the WSD task can be attributed to overfitting. A model is efficient when it not only fits the training data well but also knows how to classify the test instances that it has not seen in the previous records. When a DT is fully grown, it loses generalization ability, which is known as the overfitting problem. There are various causes of the overfitting problem of DTs. One problem found is the lack of representative instances. The following example demonstrates the problem [13-14].

TABLE III. SAMPLE SHOWING THE OVERFITTING PROBLEM OF DT

Name	Body_ Temperature	Gives_ Birth	Four_ Legged	Hibernates	Class_ label
Salamander	Cold_blood	No	Yes	Yes	No
Guppy	Cold_blood	Yes	No	No	No
Eagle	Warm_blood	No	No	No	No
Poorwill	Warm_blood	No	No	Yes	No
Platypus	Warm_blood	No	Yes	Yes	Yes

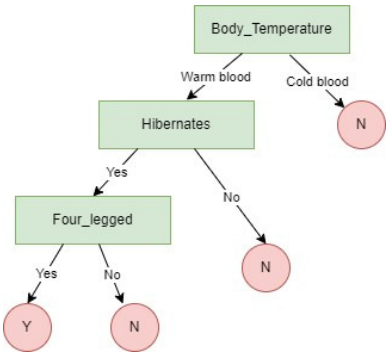

Fig. 3. Decision Tree (DT).

In Figure 3, the terminal nodes denote: Yes (Y): Mammals, and No (N): Non-mammals. Now, consider some test instances or tuples as given in Table IV.

TABLE IV. TEST SAMPLES ON THE DT FORMED

Name	Body_ Temperature	Gives_ Birth	Four_ legged	Hibernates	Class_ label
Human	Warm_blood	Yes	No	No	Yes
Anteatar	Warm_blood	No	Yes	Yes	Yes
Elephant	Warm_blood	Yes	Yes	No	Yes
Dolphin	Warm_blood	Yes	No	No	Yes
Platypus	Warm_blood	No	Yes	Yes	Yes

Among the testing instances, only Anteatar was classified correctly and Human, Elephant, and Dolphin were misclassified by the DT. This is because of a lack of training samples. There is only one tuple (Eagle) with such characteristics.

2) Naïve Bayes (NB) Chosen for the Disambiguation Task

NB was used to disambiguate the ambiguous terms in the input language. The Assamese sense tagged data (4K size) with the sense inventory of Assamese WordNet & Corpus was considered. A hold-out evaluation strategy was used for the disambiguation task. A search technique was applied to determine if the input language contains ambiguous terms by a look-up-based approach in the sense inventory. If any ambiguous term is detected, then the term is tagged with its corresponding sense by the NB classifier. Manual intervention was required to check if the appropriate sense was associated with the ambiguous term. Verified by the validators, the sense was later mapped to Hindi WordNet [14-18]. It should be noted that if the sense associated is from the sense inventory Assamese WordNet, then it can be mapped to Hindi WordNet. However, if the unique ambiguous terms with their senses are defined from Assamese Corpus, they cannot be mapped to Hindi WordNet as no such ambiguous words are available in Assamese WordNet. For example, consider an input sentence "নৃত্য এটা কলা হয়" (*Nitya eta kola hoi*). In the Assamese language, the word "কলা/kola" is ambiguous. The NB classifier recognizes the ambiguous word using the sense inventory and assigns the appropriate sense based on the learned sense-tagged module [15-16]. On verifying the sense tagged to the ambiguous term, it was found that the term "কলা/kola" has two distinct senses, as shown from the Assamese WordNet sense inventory in Table VI. A sample of Assamese sense-tagged data is shown below for the ambiguous word "আদি/aadi":

- @relation aadi
- @attribute text string
- @attribute sense {ইত্যাদি-etc-30, মূল-root-20}
- @data
- "মাউচ কিবর্ড আদি কিম্পউতাৰ আহিলা", ইত্যাদি, etc. ("*Mouse Keyboard aadi computer ahila*", *ityadi*)
- "হৰিণা গঁড় মহ আদি জন্তু আছে", ইত্যাদি, etc. ("*Harina Garh moh aadi jantu ase*", *ityadi*)
- "কামাখ্যা উমানন্দ বিশষ্ঠ আদি কামৰূপৰ তীর্থস্থান", ইত্যাদি, etc. (*Kamakhya Umanda Basistha aadi Kamrupor tirthosthan, ityadi*)
- "দোকানখন সাধুকথা ব্যঙ্গচিত্র আদি পোৱা □ায়", ইত্যাদি, etc. (*Dokankhon hadhukotha byngachitra aadi powa jai*, *ityadi*)
- "বুৰঞ্জী জ্যোতিষতত্ত্ব সাঁথৰ আদি প্রকাশিত হৈছিল ভাৰতীয়", ইত্যাদি, etc. (*Buranji jyotikhtatba hathor aadi prokakhit hoisil bharatiya*, *ityadi*)
- "মাইনা অৰুণ সৰিভ আদি কাকত আলোচনী প্রকাশেও", ইত্যাদি, etc. (*Moina Arun Surabhi aadi kakot alosoni prokasheo*, *ityadi*)
- "বিহীয়া মিচিং কার্বি আদি জনজাতীয় নৃত্য □থষ্ট", ইত্যাদি, etc. (*Bihiya Missing Karbi aadi janajatiya nritya jathesta*, *ityadi*)
- "নাদ খন্দা আদি বিভিন্ন ধৰণ কাম", ইত্যাদি, etc. (*Nad khanda aadi bibhinna dharan kam, ityadi*)
- "জেং বিহু হুচৰি আদি ৰঙালি বিহুৰ লগত", ইত্যাদি, etc. (*Jeng bihu husari aadi rangali bihur logat*)
- "খাৰু জোনবিবি বঢ়হমথুৰি আদি নাচনিয়ে বিহুত বা□ৰহাৰ", ইত্যাদি, etc. (*Kharu jonbiri borhamuthuri aadi nasoniye bihut byabahar*, *ityadi*)

- "মানস চৈখোৱা ওৰাং আদি অসম ৰাষ্ট্রীয় উদ্যান", ইত্যাদি, etc. (*Manas Saikhowa Orang aadi asom rastriya uddan, ityadi*)
- "বাল্মিকী আদিতে দস্যু আছিল", মল root (*Balmiki adite dashyu asil*, *mul*)
- "কালিদাস আদিতে অকৰা আছিল", মল root (*Kalidas adite okora asil*, *mul*)
- "ৰংঘ আদিতে কাঠ অট্টালিকা আছিল", মল root (*Rangha adite kath attalika asil*, *mul*)
- "শুনাঞায় তাজমহল আদিতে শিৱ মিন্দৰ আছিল", মল root (*Hunajai Tajmahal adite shiv mandir asil*, *mul*).

TABLE V. STATISTICS OF SENSE-TAGGED DATA

Dataset	Collected Total sense_ tagged instances	Number of senses of each ambiguous word	Type of evaluation	Average F-score of NB	Average F-score of DT
7 highly Assamese ambiguous words	350	Binary	10 fold cross validation	0.83	0.75

TABLE VI. SAMPLE OF ASSAMESE AND HINDI WORDNET

Assamese WordNet	Hindi WordNet
SynsetID : 1642 Synonyms :কলা, আৰ্ট, Gloss: কোনো কাম ভালকৈ কৌশল, বিশেষকৈ এনেকুৱা কাম ঢাকসম্পাদন কৰিবলৈ জ্ঞানৰ উপৰি কৌশল আৰু অভ্যাসৰ আৱশ্যক হয় SynsetID: 7196	SynsetID : 1642 Synonyms :कला ,फ़न, फन , हुनर , वद्या
Synonyms :কলা, বধিৰ,শ্রুতিহীন, শ্রুতিশক্তিহীন, শ্রৱণৰহিত, শ্রৱণশক্তিৰহিত, শ্রুতিৰহিত, কাণ_গধুৰ, Gloss :ঢিজনে শুনা নাপায়	SynsetID : 7196 Synonyms :बहरा ,बधरि , बहरि ,उच्चैश्रवा, श्रोतहीन

TABLE VII. OUTPUT OF EMBEDDED BASELINE SMT WITH WSD AND TRANSLITERATION SYSTEM

	Assamese sentences	Hindi Sentences
1	ইয়াত চৰাই আৰু জন্তু আছে (*Iyat sorai aru jantu ase*)	यहां पक्षी और जानवर है
2	আৰব এজন কলা ল'ৰা হয় (*Aarab ajon kola lora hoi*)	अरब एक बहरा लड़का है
3	তাজ মহোৎসৱ ভাৰতত হয় (*Taj mahotsav bharatot hohoi*)	ताज महोत्सव भारत है
4	জামু পাহাৰৰ ওপৰত (*Jammu paharor uporot*)	जम्मू पहाडी के ऊपर
5	নৃত্য এটা কলা হয় (*Nitya eta kola hoi*)	नृत्य एक कला है

The baseline SMT system Assamese-Hindi source word "কলা" is translated to the target word "बहरा", which is a wrong output of the MT system. So, a disambiguation task is performed on the input before passing it to the MT system to get correctly translated. The disambiguation performed by the NB classifier "কলা" was sense tagged with sense id 1642. After this, it was mapped with the same sense id of the Hindi WordNet, which is the most frequent synonym of id 1642 "कला" of Hindi WordNet. The WordNet's most frequent synonym was embedded to get the perfect translated Hindi word. Now, the input to the SMT was "নৃত্য এটা কলা হয়" (*Nitya eta kola hoi*) and the correct MT system output form "नृत्य एक कला है" was obtained. More such words have different Hindi translations, as shown in Table VIII, and some have the same translation even if used in different senses in different contexts, as shown in Table IX.

TABLE VIII. DIFFERENT ASSAMESE-HINDI TRANSLATIONS WITH DIFFERENT SENSES

Assamese ambiguous words	Different Hindi senses with different translated forms
কল (kal)	उद्योग, केला
আহাৰ (ahar)	आहार, असार
অন্তৰ (antar)	दलि, अंतर
আদি (aadi)	आदि, जड़
কবি (kobi)	गोभी ,कवि

TABLE IX. SAME ASSAMESE-HINDI TRANSLATION WITH DIFFERENT SENSES

Assamese ambiguous words	Different Hindi senses with different translated forms
কৰ *(kar)*	कर
উত্তৰ *(uttar)*	उत्तर

III. RESULTS AND DISCUSSION

The system was tested with 300 different Assamese sentences that contained ambiguous words. The whole output sample is shown in phases through the tables above. Table I is the sample output of the baseline SMT Assamese-Hindi system. Here, the proper nouns "আৰব" and "জাম" were not translated into Hindi and were the same on the output line. Moreover, the baseline system incorrectly translated the ambiguous word "কলা" in sentence 5 to "बहरा". In sentence 3 of Table I, "তাজ মহোৎসৱ ভাৰতত" was properly translated to "ताज महोत्सव भारत", as these OOV words may be trained by the MT system, but "মহোৎসৱ" was not. The baseline system in the next phase was combined with a transliteration system for out-of-vocabulary words. Here, the proper nouns "আৰব" and "জামু" were correctly transliterated to "अरब" and "जम्म", as shown in Table II. In sentence 5 of Table I, the output was not yet correct, so the WSD system was applied to help the baseline SMT system output correct the translated form. After associating the appropriate sense with the WSD module, the system was optimized with a large lexical resource, Hindi WordNet. WordNet provides the most frequent synonym of the same synset id in the sense denoted by the NB classifier. The translated output shown in Table VII was correct when a combination of the WSD module and Hindi WordNet was embedded into the baseline MT and transliteration system. Tables VIII and IX depict certain Assamese ambiguous words that have different translated forms and the same word form corresponding to their respective senses. The sentences tested achieved 76-77% accuracy. A bilingual Assamese-Hindi parallel corpus containing ambiguous terms can further increase the accuracy.

IV. CONCLUSION

This study embedded WSD to obtain the correct translation output from the SMT system. WSD is an important task in

many NLP applications. When embedding the WSD module into the baseline MT system, a significant improvement was observed in the output. At first, a statistical baseline MT system using Moses and OOV words was combined in the baseline with the transliteration system. Certain words have different meanings from different Hindi-translated words. At first, the ambiguous words were disambiguated in the appropriate sense using the NB classifier and the proper and most frequent synonym from Hindi WordNet before being passed to the SMT module. This improves the output result to a correct translated Assamese-Hindi form. The system was tested, and the results were analyzed by linguist scholars. More parallel and sense-tagged trained data comprising ambiguous terms with different senses, transliterated forms, and high-accuracy WSD models will enable the MT system to achieve even higher accuracy.

ACKNOWLEDGEMENT

The authors would like to thank the linguist scholars of the Gauhati University NLP Laboratory for their support in the output evaluation process.

REFERENCES

[1] R. Joshi, R. Karnavat, K. Jirapure, and R. Joshi, "Evaluation of Deep Learning Models for Hostility Detection in Hindi Text," in *2021 6th International Conference for Convergence in Technology (I2CT)*, Maharashtra, India, Apr. 2021, pp. 1–5, https://doi.org/10.1109/I2CT51068.2021.9418073.

[2] A. Kumari and D. K. Lobiyal, "Efficient estimation of Hindi WSD with distributed word representation in vector space," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, Part B, pp. 6092–6103, Sep. 2022, https://doi.org/10.1016/j.jksuci.2021.03.008.

[3] M. Sheth, S. Popat, and T. Vyas, "Word Sense Disambiguation for Indian Languages," in *Emerging Research in Computing, Information, Communication and Applications*, 2018, pp. 583–593, https://doi.org/10.1007/978-981-10-4741-1_50.

[4] R. L. Singh, K. Ghosh, K. Nongmeikapam, and S. Bandyopadhyay, "A Decision Tree Based Word Sense Disambiguation System in Manipuri Language," *Advanced Computing: An International Journal*, vol. 5, no. 4, pp. 17–22, Jul. 2014, https://doi.org/10.5121/acij.2014.5403.

[5] S. K. Sarma, H. Bharali, A. Gogoi, R. Deka, and A. K. Barman, "A Structured Approach for Building Assamese Corpus: Insights, Applications and Challenges," in *Proceedings of the 10th Workshop on Asian Language Resources*, Mumbai, India, Dec. 2012, pp. 21–28.

[6] D. S. K. Sarma and R. Medhi, "Foundation and Structure of Developing an Assamese Wordnet," presented at the 5th International Conference of the Global WordNet Association, Mumbai, India, Jan. 2021.

[7] P. Bhattacharyya, "IndoWordNet," in *The WordNet in Indian Languages*, N. S. Dash, P. Bhattacharyya, and J. D. Pawar, Eds. Singapore: Springer, 2017, pp. 1–18.

[8] A. K. Barman, J. Sarmah, and S. K. Sarma, "Assamese WordNet based Quality Enhancement of Bilingual Machine Translation System," in *Proceedings of the Seventh Global Wordnet Conference*, 2014, pp. 256–261.

[9] N. J. Kalita and B. Islam, "Bengali to Assamese Statistical Machine Translation using Moses (Corpus Based)." arXiv, Apr. 05, 2015, https://doi.org/10.48550/arXiv.1504.01182.

[10] A. Stolcke, "SRILM-an extensible language modeling toolkit," presented at the Seventh International Conference on Spoken Language Processing, Denver, CO, USA, Sep. 2002.

[11] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, Mar. 2003, https://doi.org/10.1162/089120103321337421.

[12] B. Nethravathi, G. Amitha, A. Saruka, T. P. Bharath, and S. Suyagya, "Structuring Natural Language to Query Language: A Review," *Engineering, Technology & Applied Science Research*, vol. 10, no. 6, pp. 6521–6525, Dec. 2020, https://doi.org/10.48084/etasr.3873.

[13] A. Alqahtani, H. Alhakami, T. Alsubait, and A. Baz, "A Survey of Text Matching Techniques," *Engineering, Technology & Applied Science Research*, vol. 11, no. 1, pp. 6656–6661, Feb. 2021, https://doi.org/10.48084/etasr.3968.

[14] P. Sharma and N. Joshi, "Knowledge-Based Method for Word Sense Disambiguation by Using Hindi WordNet," *Engineering, Technology & Applied Science Research*, vol. 9, no. 2, pp. 3985–3989, Apr. 2019, https://doi.org/10.48084/etasr.2596.

[15] A. Alblwi, M. Mahyoob, J. Algaraady, and K. S. Mustafa, "A Deterministic Finite-State Morphological Analyzer for Urdu Nominal System," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 11026–11031, Jun. 2023, https://doi.org/10.48084/etasr.5823.

[16] D. Chopra, N. Joshi, and I. Mathur, "Improving Translation Quality By Using Ensemble Approach," *Engineering, Technology & Applied Science Research*, vol. 8, no. 6, pp. 3512–3514, Dec. 2018, https://doi.org/10.48084/etasr.2269.