An Overview of the Basic NLP Resources Towards Building the Assamese-English Machine Translation

Nibedita Roy*, Apurbalal Senapati

Department of Computer Science and Engineering, Central Institute of Technology Kokrajhar, Kokrajhar-783370, Assam

*Corresponding author
doi: https://doi.org/10.21467/proceedings.115.7

ABSTRACT

Machine Translation (MT) is the process of automatically converting one natural language into another, preserving the exact meaning of the input text to the output text. It is one of the classical problems in the Natural Language Processing (NLP) domain and there is a wide application in our daily life. Though the research in MT in English and some other language is relatively in an advanced stage, but for most of the languages, it is far from the human-level performance in the translation task. From the computational point of view, for MT a lot of preprocessing and basic NLP tools and resources are needed. This study gives an overview of the available basic NLP resources in the context of Assamese-English machine translation.

Keywords: Machine Translation, Natural Language Processing, Resources, Language

1 Introduction

The Assamese language is one of the major regional languages of India spoken by the people of Assam and other northeastern states of India. The Assamese is a major official language in the state of Assam and is used for day-to-day conversation. Assamese is a morphologically rich language. It is one of the less computationally aware Indian languages belonging to the Indo-Aryan family. Assamese language has less amount of computational linguistic resources. The linguistic researches are in traditional mode. In the recent year most of researches have made a attempt to study Assamese language from the technological perspective. Machine Translation is the process of using software, with the help of which we can convert a source language to a target language. The Machine Translation task for Assamese language is extremely difficult because the amount of parallel corpus is extremely less. In the low-level view of the system, there are lots of sequential steps to perform the translation system. In the other words a lot of pre-processing tools like Pats-of-speech tagger (POS), Named Entity Recognizer (NER), Morphological Analyzer, Chunker, Parser, etc. are needed. The aims of this paper to investigate the availability of such resources for Assamese languages.

2 Resources for the Machine Translation

What are the resources needed to develop an end-to-end MT system is a fundamental question for the researchers. The answer to these questions is also depending on the method are following in the translation system. Machine Translation (MT) system that produce translations between any two particular languages. Machine translation transfers the most suitable target language words and phrases from the source language words and phrases. The advantages of MT system are quick translation, low price and confidential, online translation and overcome technological barriers. Currently Machine Translation is one of the heavily research area in NLP. Based on the technological approach MT can be Statistical MT, rule-based MT, neural MT,

Proceedings DOI: 10.21467/proceedings.115; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-947843-5-7

adaptive MT, hybrid MT systems, etc. A statistical approach derives with much better result once the scale of the corpus is enlarged. In Statistical approach, the most effective translation are performed supported on some decision. In Rule based approach, varied the present development on source language text are investigated, so extract them as some rules and analyze them to suit to implant for generating target language text. Neural machine translation (NMT) is an approach that uses an artificial neural network to predict the likelihood of a sequence of words, typically modelling entire sentences in a single integrated model [1]. Adaptive machine translation approach, this new technology claims to allow a MT system to learn from corrections on the fly [2]. Hybrid machine translation approach is the combination of rule based approach and statistical approach. There is a variant of resources is needed based on the chosen approaches for the translation. But, there are some resources is in common among the systems and we are focusing such resources as basic resources. Next section gives an outline of the basic resources and availability in literature.

3 Basic resources for Assamese language

3.1 POS in Assamese language

POS tagging is a basic preprocess tool for any NLP application. There are several POS taggers available in the literature [3,4]. There are two categories of POS tagging, which are supervised and unsupervised tagging. Supervised and unsupervised tagging can be of three sub types, they are rule based, stochastic based and neural network base are available. Navanath Saharia et al. in 2009 [5] have presents a work on POS tagging using stochastic POS tagger based on Hidden Markov Model (HMM) in the Assamese text corpus (Corpus Asm) of 3,00,000 words from the web version of the Assamese daily Newspaper "Asomiya Pratidin" where 10,000 words of this corpus were manually tagged by them for training. The tagset used by them have 172 tags which was larger in size with compared to the other Indian languages' tagsets. They have obtained an average tagging accuracy of 87%. According to their report, the HMM based experiments on various Indian languages, they have obtained the best accuracy level so far. Moreover, for the development of the system's accuracy, they need to proposed some additional works like, the dimension of the manually tagged a part of the corpus will need to be increased, a suitable procedure for handling unknown proper nouns will have to be developed. If this technique is often expanded to trigrams or may be n-grams employing a larger training corpus. Anup Kr. Barman et. al. [6] presents work on POS tagging for Assamese sentences, using Conditional Random Field (CRF) and Transformation Based Learning (TBL). they have reported the results 87.17 and 67.73 percent tagging accuracy for TBL and CRF respectively. Bipul Roy and Prof. Purkayastha [7] have discussed the POS tagging of Assamese language and its related issues like collection of annotated corpus, grammatical difficulties faced during POS tagging in Machine. They find that getting annotated Assamese Corpus is the toughest challenge faced by the language research. Some of them are resolve this issue by few schemes. Surjya Kanta Daimary, Vishal Goyal, Madhumita Barbora and Umrinderpal Singh [8], has developed POS tagger with good accuracy on Assamese language based on Hidden Markov Model (HMM). They work on an annotated corpus of 271,890 words with a BIS tagset consisting of 38 tag labels is used. The method is trained on 256,690 words and the remaining words are used in testing. The method obtained an accuracy of 89.21%. Dr. Karabi Kherkatary boro and Dr. Uzzal Sharma [9] give the brief idea about POS tagging in Assamese language and discus the challenges arise towards the tagging of POS and also discus various common techniques which are used in the POS tagging.

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

3.2 Parsing for Assamese Language

Parsing of a sentence is considered to be an important intermediate stage for semantic analysis in natural language processing (NLP) application such as information retrieval (IR), information Extraction (IE) and question answering (QA). There are three main categories, rule based parsing, statistical based parsing and generalized parsing are available. All the developed parsers belong to any of these categories and follow either 'top-down' or 'bottom direction'. Rahman, Mirzanur, Das, Sufal and Sharma, Utpal in 2009 [10] have developed a method to Parse Assamese text. In this work they have considered only limited number of sentences for developing rules and only seven main tags are used. They have analyzed the problems that arise in parsing Assamese sentences and produce an algorithm to unravel those issues. They produced a technique to check that grammatical structure of the sentences in Assamese text and made grammar rules by analyzing the structure of Assamese sentences. Their Parsing program can find the grammatical error, if any, within the Assamese sentences. If there is no error, their program can generate the parse tree for the input Assamese sentence. Their algorithm is a modification of Earley's Parsing Algorithm and they found the algorithm simple and efficient but the accuracy rate is not mentioned. Navanath Saharia et al. in 2011 [11] described a parsing criterion for Assamese text. They have discussed some salient features of Assamese syntax and the issues that simple syntactic frameworks cannot tackle. They have also described the practical analysis of Assamese sentences from a computational perspective. This approach can be used to parse the simple sentences with multiple noun, adjective, adverb clause. They have defined a context free grammar (CFG) to parse simple Assamese sentences. But the main drawback of this approach is that it can also generate a parse tree for a sentence which is semantically wrong. Again they have also found that if the noun is attached with any type of suffix, then the defined CFG can easily generate syntactically and semantically correct parse tree. Also to generate parse tree for the sentences which cannot be obtained using their CFG, they have applied Chu-LiuEdmond's maximum spanning tree algorithms. They have achieved an accuracy of 78.82% in this particular parsing approach. Bipul Roy and Bipul Syam Purkayastha [12] have developed a parser for Assamese language. Here they developed an Assamese tagset improving the BIS POS tagset with 31 tags. With the help of Context Free Grammar(CFG) in Python with Natural Language Toolkit(NLTK), they tag and parse the Assamese texts. This technique parse all the simple sentences of Assamese texts with an accuracy of 98.6% , in case of complex sentences it faced difficulty to parse.

3.3 Named Entity Recognizer in Assamese Language

Named Entity Recognizer is that the process of identifying and classifying proper nouns in text documents into pre-defined classes like person, location and organization. NER is an important comportment of NLP tasks like Information Extraction (IE), Question Answering (QE), and Automatic Summarization (AS). Different approaches to NER are Rule-based NER, Statistics-based NER/Machine Learning approaches and Hybrid approach. Named Entity Recognition for Assamese language has performed approaches mentioned here. The first NER was developed by Padmaja Sharma, Utpal Sharma and Jugal Kalita [13]. They developed a rule based Named Entity Recognition for Assamese. A corpus of about 50000 words was manually tagged from Assamese online Protidin articles. The approach found 500 person names and 250 location names. They were analyzed the tagged corpus to enumerate some rules for automatic Named Entity tagging. Another work was developed by them [14], where location named entities were found by suffix stripping approach to identify the root of the word. They collect the corpus from Asomiya Pratidin nearly 300,000 words. The approach is simple, interestingly, it performs reasonably well and gives an F-measure of nearly 90%.Gitimoni Talukdar, Pranjal

Protim Borah and Arup Baruah [15] developed an Assamese NER by using Naïve Bayes classifier. This is a machine learning approach. Their performance of the technique was reasonably good measure with F1-measure nearly 88.40%. Sharma, P., Sharma, U., Kalita, J [16], stateda hybrid NER approach which is a combination of both rule-based and ML approaches to improve the overall system performance. Where they stated that hybrid approach is more powerful then rule based and statistical approach. This hybrid approach is recognizing four sorts of Named Entity, Person, Location, Organization and Miscellaneous. Hybrid approach obtained an accuracy of 85%–90%. It is found that in NER community, the foremost studied types are three specializations of proper names like names of persons, locations and organizations.Gitimoni Talukdar, Pranjal Protim Borah and Arup Baruah [17], work on a supervised approach of Naïve Bayes classification model to recognize the names of person, location and organization from Assamese text. They have achieved a reasonable performance of F1-measure 89% for our system using the Naive Bayes approach when the size of the training corpus is 5000 words and numbers of named entities in the test corpus were 50.

3.4 Assamese Stemmer

The process of reducing inflection towards their root forms are called Stemming, this happen in such a way that depicting a branch of relatable words under the equivalent stem, albeit the root has no appropriate meaning. Stemmer is a rule based approach. Some of the work is done on Stemmer for Assamese, which are given below. Utpal Sharma, Jugal Kalita and Rajib Das [18] was trying to developed a stemmer for Assamese and have developed various method. They have suggested a method to find out the new root words in a corpus using a suffix list. They assume that a valid word is likely to occur at different places in the corpus with different suffices. In this approach the probability of occurrence of a word increases with the number of its occurrences in distinct cases. This approach of finding of root doesn't perform well with a little corpus. Navanath Saharia, Utpal Sharma, Jugal Kalita [19] have used suffix stripping approach in EMILLE corpus of size 123753 words to generate the stem words. They created a rule engine to generate all possible suffix sequences that can be attached after a root word and used the suffix stripping method to find root words by matching the stripped suffix in the list. In this approach they achieved an accuracy of 61%. To increase the accuracy, they need created a root list with most frequent and exceptional root words of size approximately 20,000 and matched words against this list. If the word is found within the list, then the word is marked as root else again suffix stripping is completed and matched within the root list. In this approached they have achieved an accuracy of 82%. The efficiency has increased due to the root-list. Swagata Seal and Nisheeth Joshi [20] have developed a stemmer for Assamese language. They applied rule-based approach for the Stemmer. The accuracy of the approach is (94.36%). They tested their result with 20,000 words.

3.5 Morphological Analyzer

Morphology studies the word structure and formation of word of a language. Morphological Analysis is a crucial branch of linguistics for any Natural Language Processing Technology. For Assamese language some of the reported work for morphological analysis have found. In this section we will try to summarize all related work to Assamese Morphological Analysis. Mona Parakh and Rajesha N [21], presented a Morphological Analyzers using the Suffix Stripping method for the four languages – Assamese, Bengali, Bodo and Oriya. In the proposed mechanism they need deals with only inflectional suffixes. The method involves identifying individual suffixes from a series of suffixes attached to a stem/root, using morpheme sequencing rules. The authors get 50 % coverage for 7000 to 8000 root entries. Navanath Saharia, Utpal Sharma and Jugal Kalita, [22], presented a

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

Suffix-based Noun and Verb Classifier for an Inflectional Language. In the proposed method they have consider only the morph syntactic properties of Assamese words. Assamese words are often categorized into inflected classes (noun, pronoun, adjective and verb) and un-inflected classes (adverb and particle. Sharma, Utpal and Kalita, Jugal K and Das, Rajib K [23], describe an approach to unsupervised learning of morphology from an un-annotated corpus for Assamese Language. In this paper they have present and discussed an unsupervised method for acquisition of Assamese morphology from a text corpus. This is the initial work towards unsupervised morphological analysis and it's very suitable for Assamese language. Describe approach, acquire the suffixation morphology of the language from a text corpus of about 300,000 words and build a morphological lexicon. The F-measure of the suffix acquisition is about 69%. Navanath Saharia, Kushori M Konwar, Utpal Sharma, Jugal K Kalita [24] have developed a stemmer employing a HMM based algorithm to resolve ambiguities in single letter suffixes and achieved an accuracy of 92%. They were defined two states: morphologically inflected and morphologically not inflected. Then identify the states for every words of the corpus and calculate the transition and emission probabilities from the training corpus. They have used 2000 words of EMILLE corpus because the training corpus and 1542 words as test corpus. This is a statistical approach of root word generation and therefore the choice of training set use to train the model effects the efficiency of the system. Mirzanur Rahman and Shikhar Kumar Sarma [25], they presented an Apertium based Finite-State-Transducers for developing morphological analyzer for Assamese Language with some limited domain and we get 72.7% accuracy.

3.6 Parallel Corpus

A parallel corpus is contains a collection of original texts in language L_1 and their translations into a set of languages L_2 ... L_n. In most cases, parallel corpora contain data from only two languages. There are very limited work is done for Assamese Parallel Corpus.The Research Centre for Indian Language Technology Solution(RCILTS) group, IIT Guwahati [26] has created an annotated parallel corpora for Assamese by manually tagging the sentences using the Sanchay tagging software developed by IIT Hyderabad. The Assamese Corpus developed during this project has been utilized in this work. Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, Sivaji Bandyopadhyay [27], have presented EnAsCorp1.0:English-Assamese Corpus for various tasks of NLP, specialy MT. They developed both parallel and Monolingual corpus collected from various online sources. The dataset will available in, https://github.com/cnlp-nits/EnAsCorp1.0. For implementation baseline systems with Statistical machine translation and neural machine translation approaches for the same corpus. It gives the result that NMT is better then SMT.

4 Conclusion

This paper shows some resources described in the literature. These are mandatory in almost all NLP applications. But though these are described in the literature but practically not available in executable mode or difficult to reproduce for practical use. Apart from these other resources like tag corpus or parallel corpus are needed for the MT system but these are not available in the public domain.

References

[1] https://en.wikipedia.org/wiki/Neural_machine_translation.
[2] https://www.linkedin.com/pulse/adaptive-machine-translation-nutshell-juan-mart%C3%ADn-fern%C3%A1ndez-rowda

[3] Antony P J and Dr. Soman K P "Parts Of Speech Tagging for Indian Languages: A Literature Survey*" International Journal of Computer Applications (0975 – 8887)* Volume 34– No.8, 2011.

[4] Neetu Aggwarl and Amandeep kaur Randhawa, "A Survey on Parts of Speech Tagging for Indian Languages*", International Conference on Advancements in Engineering and Technology (ICAET 2015).*

[5] Saharia, Navanath., Das, Dhrubajyoti ,Sharma, Utpal., Kalita, Jugal. "Part of Speech Tagger for Assamese Text*",In Proceedings of the ACL IJCNLP* 2009 Conference Short Papers, Suntec, Singapore, Pp. 33-36 (2009).

[6] A. K. Barman, J. Sarmah and S. K. Sarma, "POS Tagging of Assamese Language and Performance Analysis of CRF++ and fnTBL Approaches,*" 2013 UKSim 15th International Conference on Computer Modelling and Simulation*, Cambridge, UK, 2013, pp. 476-479, doi: 10.1109/UKSim.2013.91.

[7] Bipul Roy and Prof. Bipul Syam Purkayastha, *"* Annotating Assamese Corpus Using the standard POS Tagset*", International Journal of Advanced Research in Computer and Communication Engineering* ISO 3297:2007 Certified, Vol. 5, Issue 8, August 2016.

[8] Surjya Kanta Daimary, Vishal Goyal, Madhumita Barbora and Umrinderpal Singh, "Development of Part of Speech Tagger for Assamese Using HMM", *International Journal of Synthetic* Emotions Volume 9 • Issue 1 • January-June 2018

[9] Dr. Karabi Kherkatary Boro and Dr. Uzzal Sharma, "An In-depth Study on POS Tagging for Assamese Language", *ADBU-Journal of Engineering Technology*, Boro, AJET, ISSN:2348-7305, Volume 9,Issue 2, December. 2020,009021407(8PP).

[10] Rahman, Mirzanur, Das, Sufal and Sharma, Utpal (2009), "Parsing of part-of-speech tagged Assamese Texts*", IJCSI International Journal of Computer Science* Issues, Vol. 6, No. 1.

[11] Saharia Navanath ,Sharma Utpal, and Kalita Jugal (2011), "*A First Step Towards Parsing of Assamese Text",* Special Volume: Problems of Parsing in Indian Languages, May 2011.

[12] Bipul Roy and Bipul Syam PUrkayastha, *"*Parsing and Part-of-Speech tagging for Assamese texts*", Advance in Computer Science and Information*, volume 3, Issue 6, October-december, 2016, pp.517-512.

[13] Padmaja Sharma, Utpal Sharma and Jugal Kalita, "The first Steps towards Assamese Named Entity Recognition*", Brisbane Convention center, Brisbane* Autralia 2010.

[14] Padmaja Sharma and Utpal Sharma and Jugal Kalita, " Suffix Stripping based NER for Location Names", *Proceedings of 2nd National Conference on Computational Intelligence and Signal Processing (CISP)*. March 2-3. Pages:91-94. Year: 2012

[15] Gitimoni Talukdar, Pranjal Protim Borah and Arup Baruah, "Supervised Named Entity Recognition in Assamese language,*" 2014 International Conference on Contemporary Computing and Informatics (IC3I).*

[16] Sharma, P., Sharma, U., Kalita, J,"Named entity recognition in Assamese: a hybrid approach" *In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI),* pp. 2114–2120. IEEE, September 2016

[17] Talukdar, Gitimoni & Borah, Pranjal & Baruah, Arup. (2018), " Assamese Named Entity Recognition System Using Naive Bayes Classifier".DOI:10.1007/978-981-13-1810-8_4.

[18] Utpal Sharma, Jugal Kalita and Rajib Das, "Root Word Stemming by Multiple Evidence from Corpus*", in proceeding of 6th international conference on Computational Intelligence and Natural computing(CINC),* North Carolina 2003.

[19] Navanath Saharia, Utpal Sharma, Jugal Kalita, "Analysis and Evaluation of Stemming algorithms: A case study with Assamese", *In Proceeding of International Conference on Advances in Computing, Communications and Informatics(ICACCI)*, 2012

[20] Swagata Seal , Nisheeth Joshi, "Design of an Inflectional Rule-Based Assamese Stemmer", *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN: 2278-3075, Volume-8 Issue-6, April 2019

[21] Mona Parakh and Rajesha N, "Developing Morphological Analyzer for Four Indian Languages Using A Rule Based Affix Stripping Approach", *Linguistic Data Consortium for Indian Languages, CIIL*, Mysore, 2011.

[22] Navanath Saharia, Utpal Sharma and Jugal Kalita, "A Suffix based Noun and Verb Classifier for an Inflectional Language*" International Conference on Asian Language Proceesing(IALP-10),* China, 2010

[23] Sharma, Utpal and Kalita, Jugal K and Das, Rajib K. "Acquisition of Morphology of an Indic Language from Text Corpus*". ACM Transactions of Asian Language Information Processing (TALIP)*, vol 7, no. 3, article 9, p 1-33, August 2008.

[24] Navanath Saharia, Kushori M Konwar, Utpal Sharma, Jugal K Kalita, "An Improved Stemming Approach using HMM for a highly inflectional language"*, In Proceeding of 14 International on Computational Linguistics and Intelligent text processing (CICLing)*, Samos, Greece, March 24-23, 2013, pp.164-173.

[25] Mirzanur Rahman and Shikhar Kumar Sarma, *"*An implementation of apertium based assamese morphological analyzer", *International Journal on Natural Language Computing (IJNLC)* Vol. 4, No.1, February 2015.

[26] http://www.iitg.ernet.in/rcilts.

[27] Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, Sivaji Bandyopadhyay, " EnAsCorp1.0:English-Assamese Corpus*", Proceeding of the 3rd Workshop on Technologies for MT of low Resource Languages*, pages 62-68, December 04, 2020.