Dataset for Bengali Word Sense Disambiguation

Biplab Kumar Sarkar, Afrar Jahin, and Md Mahadi Hasan Nahid
(Dated: March 02,2021)

I. INTRODUCTION

We live in the era of information that can be significant or insignificant in a various manner is available on the internet in a digital format and that so in every language possible. Being an open-source platform the internet requires information in a processed way through numerous natural language processing tasks. To extract the exact meaning of a word in a different context the machine needs to be applied new technologies and that's a big challenge. It also takes some pre-processing tools to examine the lexical, semantic, syntactic, morphological meaning of a given context to extract information, to deal with the exact senses through natural language processing. One of the important tools is Word Sense Disambiguation(WSD) which plays a very important role in various language processing applications like machine translation, semantic mapping, ontology learning, information retrieval, information extraction, speech recognition, etc.

II. WORD SENSE

Word sense is primarily referred to as one of a word's meanings. There are two sets of words: a large set having more than one meaning (senses of words) and a word with only one meaning (word sense). The senses of the word are often connected within a semantic field to each other. A typical trend is that it is wider in one sense and narrower in another. This can activate the technical jargon, where narrower-sense target audience activities and the general audience appear to take the broader meaning.

III. WORD SENSE DISAMBIGUATION(ENGLISH)

In principle, a word can be polysemous, but there is very little actual uncertainty for a person in real text. In a context that appears to be a largely unconscious mechanism in humans, lexical uncertainty is nothing less than deciding the meaning of each expression. It is sometimes defined as "AI-complete" as a computational problem, i.e. a problem whose solution presupposes a solution to complete comprehension of natural language or reasoning of common sense (Ide and Véronis 1998).

In the field of linguistics, Word Sense Disambiguation (WSD) is a tedious problem and the problem is formulated as the determination of the "sense" (meaning) triggered by the use of the word in a particular context. The meaning many times had to be inferred from the term or the context of the document that surrounded it.
As a popular example ,consider the sentences:

- Bokul went to see the play Romeo and Juliet at the theater. (Drama)
- Luck comes into play. (Action)
- Each day's play starts at 10 am. (Game)

In each sentence, there are different meanings of play associated with the context. Depending on the context, terms that appear in a specific context may be interpreted in more than one way. Natural Languages such as Bangla also have several words of uncertainty to be discussed. WSD plays a crucial role in enhancing the efficiency of the system.

IV. WORD SENSE DISAMBIGUATION(BANGLA)

Bangla is an Asian language spoken by 228 million people around the world, and the WSD system of Bangla can distinguish the senses from the Bangla expression. There are essentially two kinds of WSD methods as follows:

i. Knowledge-based approach
ii. Machine Learning techniques.

Second, approaches to machine learning can be divided into methods of supervised and unsupervised learning. Knowledge-based deals with lexical corpus, viz. dictionaries, thesaurus, WordNet, and so on. In the supervised learning tech. the system is trained with labeled and feature encoded input whereby the unsupervised learning practices with unknown training sample.

Nowadays there are several techs. and algorithm that can solve the ambiguity very effectively. Deep learning and artificial neural networks have shown to be very satisfactory results practicing with complex problems and tasks. WSD in Bangla can be described as following:

Table I. Definitions of two of the senses of Bengali Words

Bangla Word	Meaning	Gloss
মাথা	গুরুত্বপূর্ণ ব্যাক্তি	কামাল সাহেব এই গ্রামের মাথা।
মাথা	অঙ্গ বিশেষ	মাথা উদ্দিপনায় সারা দেয়।
পাকা	দক্ষ	মেয়েটি গণিতে খুব পাকা।
পাকা	পক্ক	পাকা ফল কিভাবে তুলবেন?

V. THE DATASETS

The first and foremost responsibility is to interact with a well-prepared dataset for training a machine learning model in trying to obtain the best evaluation result. The collection of data is the collection of valuable information relating to the subject of research.

A. Data Collection

We have collected 10k sentences based on 20 very common ambiguous words from various Bengali daily newspapers, online portals, the giant information source, and so on. We searched for each ambiguous word and copied each sentence carrying the word. We have done it manually and thus it removes automatically hyperlinks, noises, and unnecessary texts. We cleaned the dataset by removing punctuation, numbers, brackets using a python script.

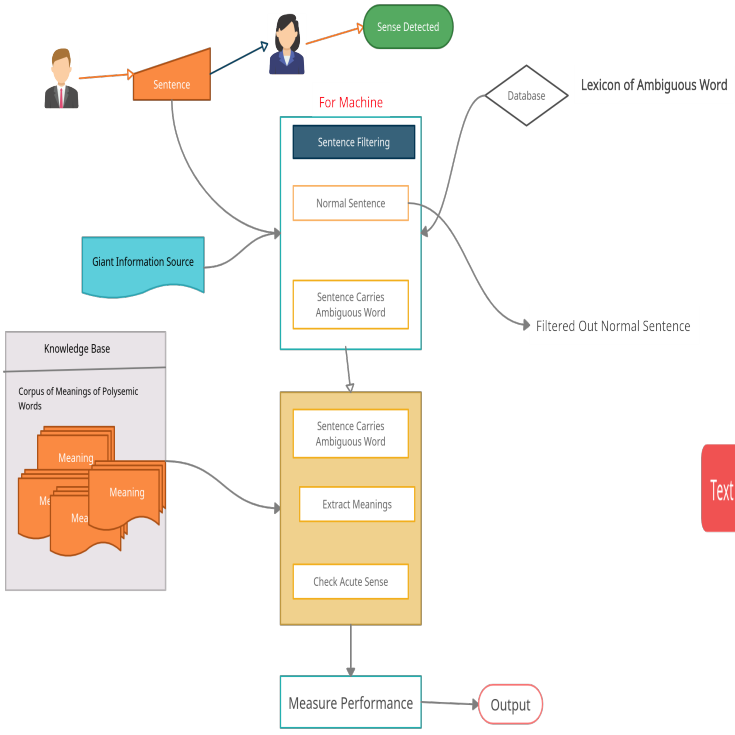

Figure 1. System Design of Word Sense Disambiguation

B. Data Preparation

Simply put, the task of taking raw data and making it ready for ingestion intoan analytics application is data preparation. The data must be processed, formatted, and converted into something digestible by the analytic software to reach the final stage of testing. These are large strokes, and a large variety of measures may be included in the actual process, such as leveraging fields and columns, modifying formats, removing redundant or junk material, and making data corrections. After collecting the data, we have preprocessedthe dataset. We have removed stop words and done tokenization.

- For any human, (Figure 1) it's completely easy to understand what is the intended meaning of a word in a context without even thinking. But in the case of a machine, we need to make sure the format and procedure to make it understood. And hence words can be polysemous so machines need to quantify the exact meaning.

- We will see the work cycle in section C where the working system is demonstrated in a figure through which the dataset has been made. It maintains a cyclic order to process the data and to enrich the size of the corpus.

C. WSD Work Cycle

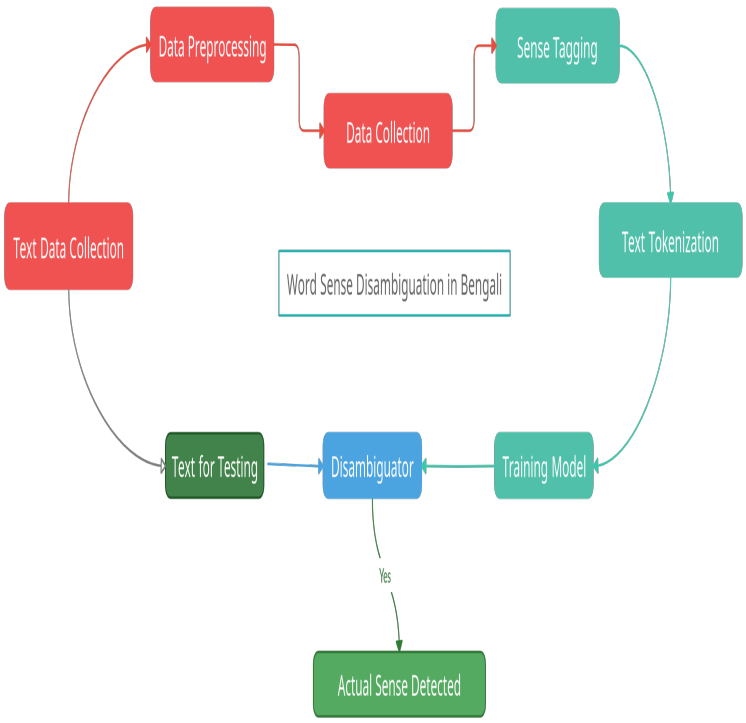

Figure 2. Work Cycle of Word Sense Disambiguation

Table II. Sample Dataset before Pre-processing

Ambiguous word	Meaning	Text
মাথা	অঙ্গ বিশেষ	মাথা উদ্দিপনায় সারা দেয়।
মাথা	বুদ্ধি	ছেলেটির মাথা খুব ভালো।
কাটা	কর্তন	তার হাত কাটা গেছে।
কাটা	বাধা	পথের গতি পথের কাঁটা।
পাকা	দক্ষ	সে আসলেই কাজে পাকা।
পাকা	পক্ক	পাকা ফল কিভাবে তুলবেন?
কাঁচা	কচি	গরু কাঁচা ঘাস খাচ্ছে।
কাঁচা	অদক্ষ	তার হাতের লেখা কাঁচা।
মন	মনোযোগ	সে মন দিয়ে বই পড়ছে।
মন	স্মরণ	মনে পড়ে অপারগতায়।

Contains 10k rows

Table III. Dataset Format

Ambiguous word	Count of senses per word	Count of texts per word
কড়া	13	166
গরম	7	118
গায়ে	5	12
ছাড়া	12	186
পড়া	41	1002
ফুল	7	121
মাথা	31	373
মুখ	24	543

VI. SAMPLE DATASET

The (.csv) file format dataset looks like the following TABLE II. Initially we have collected 20 ambiguous words that have more than one sense in Bengali Language. We have studied thoroughly and found out 1165 ambiguous words in Bengali. We are continuing to enlarge the dataset.

VII. DATA FORMAT

TABLE III describes the number of senses per ambiguous words and Text for ambiguous words in the corpus. Here only eight words are chosen to draw an overview of the data format.

VIII. DISCUSSION AND CONCLUSION

In this analysis, we have attempted to discuss how the basic system of wsd functions. We tried to incorporate the problems of creating a corpus and how to perceive the senses from scratch. We will enlarge our dataset with a satisfactory amount of ambiguous words and sentences. Word Sense Disambiguation is continuously improving due to its significant contribution to natural language

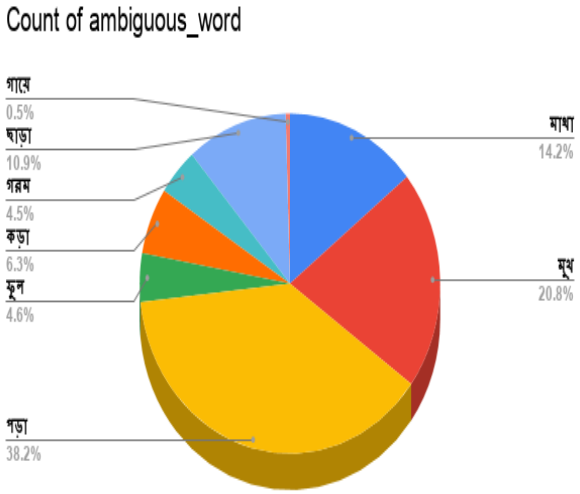

Figure 3. Count of ambiguous Words in Dataset

processing. Although there are so many problems with Word Sense Disambiguation, research work has begun. In Bengali Word Sense Disambiguation, minimal work is completed, which is not adequate. As Bengali is a relatively difficult and complex language this work will increase the coverage of correct sense detection. It will thereby pave the path to identifying the acute sense in Bengali Word Sense Disambiguation.

IX. TEAM UPBEAT

- Biplab Kumar Sarkar(Member 1)
  Email: bip.sec22@gmail.com
  Department of CSE
  Sylhet Engineering College

- Afrar Jahin (Member 2)
  Email: afrarjahin@gmail.com
  Department of CSE
  Sylhet Engineering College

- **Md Mahadi Hasan Nahid**(Mentor)
  Assistant Professor
  Email: nahid-cse@sust.edu
  Department of CSE
  Shahjalal University of Science and Technology

Thank You!