A Lemmatizer Tool for Assamese Language

Arindam Roy, Sunita Sarkar(✉), and Hsubhas Borkakoty

Assam University, Silchar, India
arindam_roy74@rediffmail.com,
sarkarsunita2601@gmail.com

Abstract. Word Sense Disambiguation (WSD) requires sense tagged corpora. Words in a corpus appear in inflected or morphed forms. Sense tagging can only be done with words in their root or lemmatized forms. Similarly for Part of Speech Tagging (POS), the words in a corpus are required to be available in their root forms. A WordNet, which is a lexical knowledge base, consists of words in their lemmatized forms. In this paper we present a scheme whereby a trie is constructed from words in the Assamese WordNet. An input word, which maybe in inflected form, is compared against an entry in the trie by searching the trie with the input word as key and following the principle of longest prefix match, the root form of an inflected word may be obtained. When there is a mismatch between an input word and an entry in the trie, a rule based morphological analyzer provides the lemmatized form of the inflected word.

Keywords: WSD · Corpus · Trie · WordNet

1 Introduction

In Natural Language Processing (NLP), Sense marked corpora or Sense annotated corpora is of utmost importance. It is used in Supervised WSD which generates the correct sense of a word in a given context. But a corpora is always available in inflected forms. The words in the corpora need to be converted to their root forms through Lemmatization. Similarly POS tagging of any corpora would require that the inflected words in the corpora are first rendered in their root forms and only then any POS tagging algorithm may be applied on the corpus containing words in their root forms. Lemmatization is also used in Information Retrieval to expedite the retrieval time and improve the relevance of retrieved documents [1].

Stemming is different from Lemmatization in the sense that stemming aims to convert a word into its base form which may or may not be a dictionary word. For example, a stemmer can produce "parti" from the word "parties" whereas a Lemmatizer has to produce the root word "party". A lemmatizer does not simply remove inflections but relies on WordNet to produce the correct root form of an inflected word. There are broadly three different approaches to Lemmatization. These are, namely, rule based, statistical and hybrid (both rule based and statistical).

In this paper we have also used a hybrid method which is different from the standard one. Here the longest prefix match has been used to generate the root word

J. K. Mandal et al. (Eds.): CICBA 2018, CCIS 1031, pp. 124–133, 2019.
https://doi.org/10.1007/978-981-13-8581-0_10

from the trie. In cases where longest prefix match has been found to be deficient, morphological rule based method has been used.

2 Assamese Language

Assamese is a diverse and morphologically rich language. The language has its own script and literary texts since the ancient times (from 14th century). It belongs to eastern sub group of Indo Aryan languages which falls under Indo European languages. Currently it has around 15 million native speakers (Census 2010) [2]. It is the lingua franca of the Indian state of Assam. It is also partly spoken in some areas of Indian state of Arunachal Pradesh. An Assamese based Creole Language called Nagamese is widely used in the Indian state of Nagaland. Magadhi Prakrit, a middle Indo Aryan language, is believed to be the source of Assamese language [14]. Eastern Magadhi Prakrit and Magadhi Apabhramsa can be divided into four dialect groups: (1) Rādha dialects which represent standard Bengali colloquial in Western Bengal and Oriya in the south west (2) Varendra dialects of North Central Bengal (3) Kāmarūpa dialects which represent Assamese and some dialects of North Bengal and (4) Vanga dialects which represent the dialects of East Bengal [16].

Assamese WordNet is part of IndoWordNet which is a linked lexical knowledge base of major Indian languages belonging to Indo-Aryan, Dravidian and Sino Tibetan families. English WordNet was the first WordNet [17] to be built. Hindi WordNet [18] was the first of its kind in India and it followed the principle of expansion from English WordNet. The nationwide project of building Indian language WordNets followed suit and it also follows the expansion approach from Hindi WordNet. The Assamese WordNet consists of 14,958 synsets. The Part of Speech (POS) subdivision is as follows - (i) Noun-9065, (ii) Verb-1676, (iii) Adjective-3805, (iv) Adverb-412.

The rest of the paper is as follows: Sect. 3 is on literature survey. Section 4 and its subsections provide a description of Word formation in Assamese, Sect. 5 and its subsections presents the framework and methodology for a lemmatizer in Assamese, Sect. 6 describes experimental results while Sect. 7 winds up the discussion by presenting the conclusions and future work.

3 Literature Review

Lovins was the first to develop a stemmer [3], which was meant for IR/NLP applications. His methodology consisted of the use of a manually developed list of 294 suffixes, each linked to 29 conditions, plus 35 transformation rules. Given an input word, the suffix with an appropriate condition is checked and removed. Porter developed the Porter stemming algorithm [4] which became the most widely used stemming algorithm for English language. It was described in a very high level language known as Snowball. Statistical approaches have been significantly used for stemming. Significant works are Goldsmith's unsupervised algorithm for learning morphology of a language based on the Minimum Description Length (MDL) framework [5, 6], Creutz's unsupervised morpheme segmentation [7, 8] which uses probabilistic

maximum a posteriori (MAP) formulation. Hidden Markov models have also been used in stemming [9]. In this approach each word is considered to be composed of two parts "prefix" and "suffix". HMM states are composed of two disjoint sets: Prefix state which generates the first part of the word and Suffix state which generates the last part of the word, if at all the word has a suffix. A complete and trained HMM can then perform stemming directly. A two level morphological analyser containing a large set of morphophonemic rules was developed by Karttunen et al., [10]. The work started in 1980 and the first implementation was available in 1983. An Arabic lemmatizer was proposed by El-Shishtawy [11]. Different Arabic language knowledge resources were used to generate accurate lemma form and its relevant features that support IR purposes and a maximum accuracy of 94.8% was achieved. A Turkish Morphological Analyzer called OMA gives all possible analyses for a given word with the help of finite state technology. As far as Indian languages are concerned, Ramanathan and Rao was the earliest work which performed longest match stripping on manually sorted suffix list to produce a Hindi stemmer [12]. Mazumder et al. [13] proposed a clustering based approach for discovering equivalence classes of root words and their morphological inflections. The equivalence classes are underpinned by a set of string distance measures to cluster the lexicon for a given text.

4 Word Formation in Assamese

The primary words or the Lemmas in Assamese have both Aryan and non Aryan origin. The secondary word formation of Assamese language is realized through two different approaches [14]. The approaches are affixation: addition of prefixes and suffixes and Compounding: addition of certain words to form a new word.

4.1 Affixation

Affixes are added before or after the word to create a new inflected form of the word with respect to number, person, tense, aspect and mood. There are 20 prefixes of Assamese words that can be added before the word to form a new word, such as

- প্র: প্রমাণ (proof)
- পৰা : পৰাজয় (defeat)
- বি : বিপদ (danger)
- নি: নিহিত (contain)
- সু: সুকৃতি (good deedInitial Population)

4.2 Compounding

The Assamese compound words are formed in three different forms, viz. closed Form, Open form and Hyphenated form. Apart from these three forms, there also exist a set of compounds in which one word is of native origin and the other is of foreign origin. For

example, the word ৰংঘৰ (palace) is compounded from the word ৰং which is a Thai word meaning palace and ঘৰ, which is an Assamese word meaning house.

Closed Form. In this formation, the words are formed by adding more than one word joined together to form a new word with new meaning. Some examples are given below:

- হিমালয় : হিম (ice) + আলয় (home)
- আগদিনা: আগ(before) + দিনা (day)
- চন্দ্রোদয়: চন্দ্র(moon) + উদয় (rise)
- যথোপযুক্ত: যথা(As) + উপযুক্ত(appropriate)
- শুক্রেশ্বৰ: শুক্র (name of a yogi) + ইশ্বৰ (god)

Open Form. In this form, two different words are added to form a new word. Here, combination of more than one word that work as a unit in order to convey different meaning. Examples include:

- কাঠৰ পুতলা (noun + noun): one who is entirely led by others
- মাটিৰ মানুহ (noun + noun): one who is humble,
- নপতা ফুকন (Adjective + Noun): A false leader
- হাত দীঘল (Noun+ adjective): Influential person
- গেলা গপ (Adjective + noun): vain boasting
- হাত টান (Noun + Adjective): one who is miser
- কপাল ফুটা (Noun + Adjective): one who is unlucky

Hyphenated Form: In this form, the words are joined together by a hyphen. Some of the examples are given below:

- হকা-বাধা (act of preventing something)
- মৰম-চেনেহ (act of love)
- মাত-বোল (act of calling someone)
- আলি-পদূলি (road)
- অহা-যোৱা (coming and going)
- উঠা-বহা (act of sitting and standing up)
- ঘৰ-বাৰী (Home)
- টকা-সিকা (money)
- মাৰ-পিট (to hit someone)
- হৰণ-ভগন (to lose something) etc

The noticeable thing about the words formed via this formation is that the parts of each word has similar meaning. For example, for the word মৰম-চেনেহ, both parts of the word (মৰম and চেনেহ) convey the same meaning, i.e. love.

4.3 Types of Suffixes in Assamese Language

Suffixes are added at the end of the words to form a new word. The suffixes are also termed as "প্রত্যয়" [12]. There exists four different types of suffixes in Assamese language. They are:

- বিভক্তি (suffix marker)
- স্ত্রীপ্রত্যয় (feminine suffixes)
- তদ্ধিত প্রত্যয় (derivational suffixes)
- কৃৎ প্রত্যয় (verbal suffixes)

বিভক্তি (Word Suffix markers). There exist seven different suffix markers for Assamese language. These are:

- Nominative: এ,ই :আখৰে (আখৰ+এ)) (of the letters, সামৰি (সামৰ+ই)) (finishing up)
- Accusative: অক: আখৰক (আখৰ+অক) (to the letters), মানুহক (মানুহ+অক) (to people)
- Dative: ৰে, দি, দৱাৰা: পানীৰে (পানী+ৰে)) (with water) , চকুৰেদি(চকু+দি) (with eyes), মানুহৰদৱাৰা (মানুহ+অৰ+দৱাৰা)) (by people)
- Genitive : লৈ: মানুহলৈ (মানুহ+লৈ)) (to people)
- Ablative:পৰা: মানুহৰপৰা(মানুহ+ৰ+পৰা))(from people)
- Instrumental: অৰ: পানীৰ (পানী+অৰ) (of water)
- Locative: অত: পানীত (পানী+ত) (in water)

স্ত্রী-প্রত্যয় (Feminine Suffixes). The feminine suffixes in Assamese consists of two different suffixes: **ঈ** and **অনী**. Some of the examples depicting its use are given as follows:

ঈ: কলা (Deaf)+ঈ=কলী,
বুঢ়া (Old) + ই=বুঢ়ী
পেটুলা (Fat) + ঈ=পেটুলী
অনী:কুকুৰ (Dog) + অনী=কুকুৰনী,
হস্তী (Elephant) + নী=হস্তিনী, নাতি (Grandchild) + নী=নাতিনী

তদ্ধিত প্রত্যয় (Derivational Suffix). The derivational suffixes are added after the word in order to form a complete new word. The formation of derivational suffixed words can be achieved using proper suffixes with the given word. Examples:

Sanskrit Derivational Suffixes: ষ্ণি: দশৰথ+ষ্ণি: দাশৰথি (son of King Dasarath), বন্ধু+ষ্ণ=বান্ধৱ(friend) etc.

Assamese Derivational Suffixes: ঈ: তাম+ঈ:তামী (of brass), লুণীয়া (লোণ+ঈয়া) (salty), খঙাল (খং+আল) (angry) etc.

কৃৎ প্রত্যয় (verbal suffixes). The suffixes that are added to the root of the verb (ধাতু) are called কৃৎ প্রত্যয়. The suffixes are added to the respective verbs to convey a new meaning for the verb. It is not necessary that the root word be verb, these suffixes can also be added to, for example, adjectives. But these work better in case of verbal roots.

Examples are: কৰা (কৰ+আ) (Do), শোৱন (শো+অন) (the act of sleeping), জিৰণি (জিৰ+অণি) (the act of taking rest), শিকাৱু (শিক+আৱু) (learner) etc.

5 Methodology

The approach that we have followed creates a trie data structure by inserting words into it from the Assamese WordNet. The words in a WordNet are in their root forms. It checks if a given word from the WordNet is in the trie or not. If not, then it is inserted in the trie. Whenever a word is given as input, the trie is searched using input word as key and longest prefix match is used to compare the input word with the corresponding entry in the trie. The corresponding branch of the trie which has the longest prefix match with the input word is considered as the root of the input word. If there is a mismatch between the input word and the corresponding entry in the trie, a morphological rule based analyzer is called which provides the lemmatized form of the inflected word.

For example let us take the word ৰাজ্যক. The root of the inflected word ৰাজ্যক is ৰাজ্য which is present in Assamese WordNet. So the word ৰাজ্য would be inserted in the trie. When the trie is searched with the inflected word ৰাজ্যক as key, by the principle of longest prefix match, the word ৰাজ্য would be output which is the root. Similar argument holds good for commonly occurring words in a corpus like প্রসঙ্গত, সম্প্রদায়ৰ, শব্দৰ, অসমৰ, সংস্কৃতৰ, ভাষাত, পদার্থৰ, মুখত, ভাষাতকৈ, মানচিত্রখন, অবচেতনত, , সপ্তাংশতকৈ, পুৰাণত, অলোচনাৰ etc.

5.1 Deviation from Trie Based Approach

Assamese language has a large number of verbs. There exist several irregularities in verbs in Assamese language also, like all other languages. In case of lemmatization, the

problem arises when the structure of the verb changes with change in tense [15]. Some examples are mentioned below.

যা: গৈছিলো,গলোহেতেন, গলিহেতেন, গলাহেতেন ,গৈছে, গল, গৈছ, গৈছা, গলো etc. (various forms of go)
অহা: আহা, আহক, আহিছা, আহিছিলা, আহিছিলি, আহিছিলে, আহিছিল, আহিছিলো etc. (various forms of come)
ৰোয়া : ৰুইছিল, ৰুইছা (to sow)

Apart from the verbs, there are derivational suffixes which don't easily lend themselves to straightforward derivation from trie structure mentioned above. A few examples, which are representative in nature, will be in order:

কাৰুণ্য = কৰুণা + ষ্ণ্য (কৰুণা is the root)
সামীপ্য = সমীপ + ষ্ণ্য (সমীপ is the root)
গাম্ভীর্য্য = গম্ভীৰ + ষ্ণ্য (গম্ভীৰ is the root)
পনীয়া = পানী + ঈয়া (পানী is the root)
এটীয়া = এটা + ঈয়া (এটা is the root)
ভদীয়া = ভাদ + ঈয়া (ভাদ is the root)
নুমলীয়া = নোমল + ঈয়া (নোমল is the root)
কপহৰা = কপাহ + উৰা (কপাহ is the root)

From these examples we can see that derivationally inflected words need to be split into its root form and suffix following the rules of the language. A morphological analyzer conforming to rules of suffix splitting for derivationally inflected words in Assamese language has been encoded in our system.

6 Experimental Results

The Assamese corpus was mainly taken from Assamese Corpora provided by TDIL (Technology Development for Indian Languages) under Ministry of Electronics and Information Technology, Government of India. The corpora consists of texts of Assamese history, Assamese society and community tourism, health etc.

A few snapshots of the output is shown below in Figs. 1, 2 and 3.

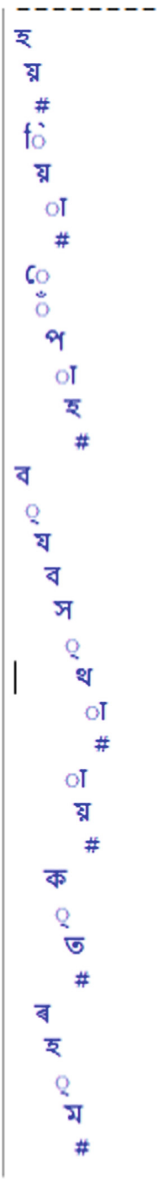

Fig. 1. Trie structure of a few words of Assamese Language

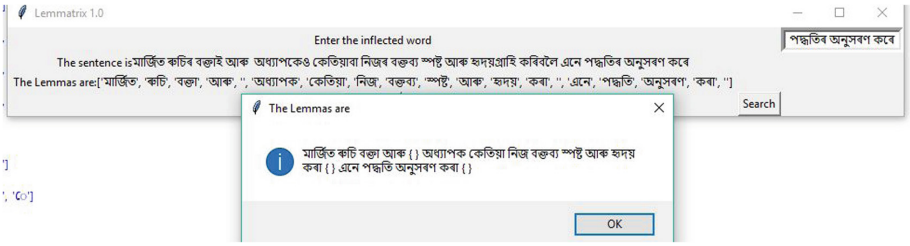

Fig. 2. Lemmatization of a sentence from Assamese corpus.

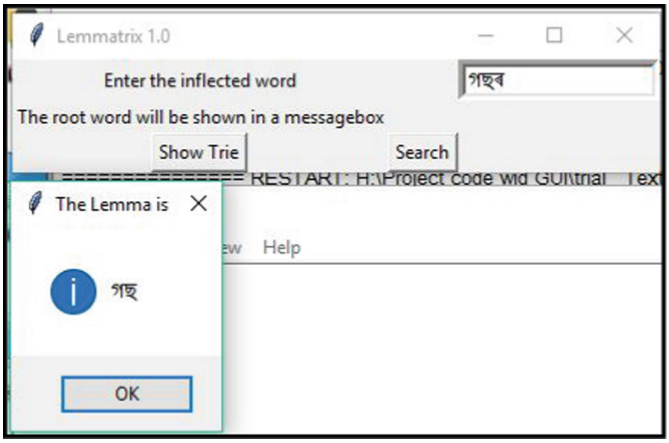

Fig. 3. A GUI depicting the lemma of an inflected Assamese word

Table 1. Result of Assamese Lemmatizer tool

	Noun	Verb	Adjective	Adverb
Total words in corpus	7312	6786	2134	517
Correctly analyzed words	5841	5264	1557	343

7 Conclusion and Future Work

We have tested our approach on significantly varied categories of text as mentioned above. Our method envisage obtaining the root of a word through prefix matching and suffix stripping with the aid of a trie data structure and rule based morphology. The results validate the efficiency of the proposed system. It has been noticed that, at times, Assamese WordNet does not contain the root form of inflected words in the corpus, specially, in the case of nouns and adjectives. Commonly used words in a corpus like

প্রতিবেদন, সামগ্রিক, ক্ষেত্র, প্রণালী, পোহ, সোমা, পিৰালি, প্রবর্তন, সাম্য

etc. are not present, till date, in Assamese WordNet. There are significant inflectional variations in case of verbs when there is a change of tense, all of which have not been addressed in the present system. As part of future work an exhaustive rule based morphological analyzer may be built to address the irregularities of verbs in Assamese language.

References

1. Christopher, M., Prabhakar, R.D., Hinrich, S.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
2. https://www.ne.se/språk/världens-100-största-språk-2010
3. Lovins, J.B.: Development of a stemming algorithm. Mech. Transl. Comput. Linguist. **11**(1-2), 22–31 (1968)
4. Porter, M.F.: An algorithm for suffix stripping. Program **14**, 130–137 (1980)
5. Goldsmith, J.A.: Unsupervised learning of the morphology of a natural language. Comput. Linguist. **27**(2), 153–198 (2001)
6. Goldsmith, J.A.: An algorithm for the unsupervised learning of morphology. Nat. Lang. Eng. **12**(4), 353–371 (2006)
7. Creutz, M., Lagus, K.: Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical report A81, Publications in Computer and Information Science, Helsinki University of Technology (2005)
8. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Trans. Speech Lang. Process. **4**(1), 1–34 (2007)
9. Massimo, M., Orio, N.: A novel method of stemmer generation based on hidden Markov models. In: CIKM 2003, New Orleans, Louisiana, USA (2003)
10. Karttunen, L.: KIMMO: a general morphological processor. In: Texas Linguistic Forum, vol. 22, pp. 163–186 (1983)
11. El-Shishtawy, T., El-Ghannam, F.: An accurate arabic root-based lemmatizer for information retrieval purposes. IJCSI Int. J. Comput. Sci. Issues **9**(1, 3) (2012). ISSN 1694-0814
12. Ramanathan, A., Rao, D.D.: A lightweight stemmer for Hindi. In: Workshop on Computational Linguistics for South-Asian Languages, EACL (2003)
13. Majumder, P., Mitra, M., Parui, S.K., Kole, G., Mitra, P., Datta, K.: YASS: yet another suffix stripper. ACM Trans. Inf. Syst. **25**(4), 18–38 (2007)
14. Kakati, B.: Assamese-its formation and development. Govt of Assam, Department of Historical and Antiquarian Studies, Narayani Handiqui Historical Institute, Guwahati, Assam (1941)
15. Bora, S.: Bahal Byakaran, Smt S Dey, College Hostel Road, Panbazar, Guwahati (2012)
16. Chatterjee, S.K.: Origin and Development of Bengali Language (ODBL), Rupa and Co. (2002)
17. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: An On-line Lexical Database. Princeton University, Cognitive Science Laboratory, Technical report (1993)
18. Narayan, D., Chakrabarty, D., Pande, P., Bhattacharyya, P.: An experience in building the indo WordNet-a WordNet for Hindi. In: 1st International Conference on Global WordNet (GWC 2002), Mysore, India (2002)