Word Sense Disambiguation in Bengali: A Lemmatized System Increases the Accuracy of the Result

Alok Ranjan Pal
Detp. of Computer Sc. and Engg.
College of Engg. & Mgmt, Kolaghat
Kolaghat, India

Diganta Saha, Sudip Naskar
Dept. of Computer Sc. and Engg.
Jadavpur University
Kolkata, India

Niladri Sekhar Dash
Linguistic Research Unit
Indian Statistical Institute
Kolkata, India

***Abstract*-In the proposed approach, an attempt was made to disambiguate Bengali ambiguous words using Naïve Bayes Classification algorithm. The whole task was divided into two modules. Each module executes a specific task. In the first module, the algorithm was applied on a regular text, collected from the Bengali text corpus developed in the TDIL project of the Govt. of India and the accuracy of disambiguation process was obtained around 80%. In the second module, the whole training data and the test data were lemmatized and applying the same algorithm, around 85% accurate result was obtained. The output was verified with a previously tagged output file, generated with the help of a Bengali lexical dictionary. The implicational relevance of this study was attested in automatic text classification, machine learning, information extraction, and word sense disambiguation.**

***Keywords*-Natural Language Processing; Bengali Word Sense Disambiguation; Bengali WordNet; Naïve Bayes Classification**

I. INTRODUCTION

In all natural languages, there are a lot of words that denote different meanings based on the contexts of their use within texts. Word Sense Disambiguation (WSD) [1-6] is the technique for finding the actual meaning of the ambiguous word in a particular contextual environment. For example, in English, the word 'goal' may denote several senses based on its use in different types of construction, such as *He scored a goal, It was his goal in life,* etc. Such words with multiple meanings are ambiguous in nature and they posit serious challenges in understanding a natural language text by machine.

The act of identifying the most appropriate sense of an ambiguous word in a particular syntactic context is known as WSD. A normal human being, due to her innate linguistic competence, is able to capture the actual contextual sense of an ambiguous word within a specific syntactic frame with the knowledgebase triggered from various intra- and extra-linguistic environments. Since a machine does not possess such capacities and competence, it requires some predefined rules or statistical methods to do this job successfully.

Normally, two types of learning procedure are used for WSD. The first one is Supervised Learning, where a learning set is considered for the system to predict the actual meaning of an ambiguous word within a syntactic frame in which the specific meaning for that particular word is embedded. The system tries to capture contextual meaning of the ambiguous word based on that defined learning set. The other one is Unsupervised Learning where dictionary information (i.e., glosses) of the ambiguous word is used to do the same task. In most cases, since digital dictionaries with information of possible sense range of words are not available, the system depends on on-line dictionaries like WordNet [7-13] or SenseNet.

In the proposed approach, we have adopted the Naive Bayes [14] probabilistic technique for Sense Disambiguation task. The whole task was divided into two modules. First, the algorithm was applied on a regular text and the accuracy of the disambiguation process was obtained around 80%. In the second module, the whole training data and the test data were lemmatized and applying the same algorithm, around 85% accurate result was obtained.

The output was verified with a previously tagged output file, generated with the help of a Bengali lexical dictionary.

The organization of the paper is as follows: in Section 2, we present a short review of some earlier works; in Section 3, we refer to the key features of Bengali morphology with reference to English; in Section 4, we refer to the Bengali corpus we have used for our study; in Section 5, we explain the approach we have adopt for our work, in Section 6, we present the results and corresponding explanations; and in Section 7, we infer conclusion and redirect attention towards future direction of this research.

II. REVIEW OF EARLIER WORKS

WSD is perhaps one of the greatest open problems at lexical level of Natural Language Processing. Several approaches have been established in different languages for assigning correct sense to an ambiguous word in a particular context. Along with English, works have been done in many other languages like Dutch, Italian, Spanish, French, German, Japanese, Chinese, etc. And in most cases, they have achieved high level of accuracy in their works.

978-1-4799-8349-0/15/$31.00 ©2015 IEEE

For Indian languages like Hindi, Bengali, Marathi, Tamil, Telugu, Malayalam, etc., effort for developing WSD system has not been much successful due to several reasons. One of the reasons is the morphological complexities of words of these languages. Words have a complex morphology and because of that there is no benchmark in these languages (especially in Bengali). Keeping this reality in mind we have made an attempt to disambiguate word sense in Bengali. We believe this attempt will lead us to the destination through the tricky terrains of trial and error.

In essence, any WSD system typically involves two major tasks: (a) determining the different possible senses of an ambiguous word, and (b) assigning the word with its most appropriate sense in a particular context where it is used. The first task needs a Machine Readable Dictionary (MRD) to determine the different possible senses of an ambiguous word. At this moment, the most important sense repository used by the NLP community is the WordNet, which is being developed for all major languages of the world for language specific WSD task as well as for other linguistic works. The second task involves assigning each polysemic word with its appropriate sense in a particular context.

The WSD procedures so far used across languages may be classified into two broad types: (i) knowledge-based methods, and (ii) corpus-based methods. The knowledge-based methods obtain information from external knowledge sources, such as, Machine Readable Dictionaries (MRDs) and lexico-semantic ontologies. On the contrary, corpus-based methods gather information from the contexts of previously annotated instances (examples) of words. These methods extract knowledge from the examples applying some statistical or machine learning algorithms. When the examples are previously hand-tagged, the methods are called *supervised learning* and when the examples do not come with the sense labels they are called *unsupervised learning*.

A. Knowledge-based Methods

These methods do not depend on large amount of training materials as required in supervised methods. Knowledge-based methods can be classified further according to the type of resources they use: Machine-Readable Dictionaries; Thesauri; Computational Lexicon or Lexical Knowledgebase.

B. Corpus-based Methods

The corpus-based methods also resolute the sense through a classification model of example sentences. These methods involve two phases: *learning* and *classification*. The learning phase builds a sense classification model from the training examples and the classification phase applies this model to new instances (examples) for finding the sense.

C. Methods Based on Probabilistic Models

In recent times, we have come across cases where various statistics-based probabilistic models are being used to carry out the same task. The statistical methods evaluate a set of probabilistic parameters that express conditional probability of each lexical category given in a particular context. These parameters are then combined in order to assign the set of categories that maximizes its probability on new examples.

The Naive Bayes algorithm is the mostly used algorithm in this category, which uses the Bayes rule to find out the conditional probabilities of features in a given class. It has been used in diverse works applied to the task of WSD.

In addition to these, there are also some other methods that are used in different language for WSD task, such as, methods based on the similarity of examples, *k*-Nearest Neighbour algorithm, methods based on discursive properties, and methods based on discriminating rules, etc.

III. KEY FEATURES OF BENGALI MORPHOLOGY

In English, compared to Indic languages, most of the words have limited morphologically derived variants. Due to this factor it is comparatively easier to work on WSD in English as it does not pose serial problems to deal with varied forms. For instance, the verb *eat* in English has five conjugated (morphologically derived) forms only, namely, *eat, eats, ate, eaten,* and *eating*. On the other hand, most of the Indian languages (e.g., Hindi, Bengali, Odia, Konkani, Gujarati, Marathi, Punjabi, Tamil, Telugu, Kannada, Malayalam, etc.) are morphologically very rich, varied and productive. As a result of this, we can derive more than hundred conjugated verb forms from a single verb root. For instance, the Bengali verb খাওয়া (khāoyā) "to eat" has more than 150 conjugated forms including both *calit* (colloquial) and *sādhu* (chaste) forms, such as, খাই (khai), খাস (kkās), খাও (khāo), খায় (khāy), খান (khān), খাচ্ছি (khācchi), খাচ্ছিস (khācchis), খাচ্ছ (khāccha), খাচ্ছেন (khācchen), খাচ্ছে (khācche), খাচ্ছি (khāitechi), খেয়েছি (kheyechi), খেয়েছ (kheyecha), খেয়েছিস (kheyechis), খেয়েছে (kheyeche), খেয়েছেন (kheyechen), খেলাম (khelam), খেলি (kheli), খেলে (khele), খেল (khela), খেলেন (khelen), খাব (khāba), খাবি (khābi), খাবে (khābe), খাবেন (khāben), খাচ্ছিলাম (khācchilām), খাচ্ছিলে (khācchile), খাচ্ছিল (khācchila), খাচ্ছিলেন (khācchilen), খাচ্ছিলি (khācchili), etc. (to mention a few).

While nominal and adjectival morphology in Bengali is light (in the sense that the number of derived forms from an adjective or a noun, is although quite large, in not up to the range of forms derived from a verb), the verbs are highly inflected. In general, nouns are inflected according to seven grammatical cases (nominative, accusative, instrumental, ablative, genitive, locative, and vocative), two numbers (singular and plural), a few determiners like, -টা (-□ā), -টি (-□i), -খনা (-khānā), -খনি (-khāni), and a few emphatic markers, like -ই (-i) and -ও (-o), etc. The adjectives, on the other hand, are normally inflected with some primary and secondary adjectival suffixes denoting degree, quality, quantity, and similar other attributes. As a result, to build up a complete and robust system for WSD for all types of morphologically derived forms tagged with lexical information and semantic relations is a real challenge for a language like Bengali [15-25].

IV. THE BENGALI CORPUS

The Bengali corpus that was used in this work was developed under the TDIL (Technology Development for the Indian Languages) project, Govt. of India (Dash 2007). This corpus contains text samples from 85 text categories or subject domains like Physics, Chemistry, Mathematics, Agriculture, Botany, Child Literature, Mass Media, etc. (Table 1) covering 11,300 A4 pages, 271102 sentences and 3589220 non-tokenized words in their inflected and non-inflected forms. Among these total words there are 199245 tokens (i.e., distinct words) each of which appears in the corpus with different frequency of occurrence. For example while the word মাথা (māthā) "head" occurs 968 times, মাথায় (māthāy) "on head" occurs 729 times, মাথার (māthār) "of head" occurs 398 times followed by other inflected forms like মাথাতে (māthāte) "in head", মাথাটা (māthā□ā) "the head", মাথাটি (māthā□i) "the head", মাথাগুলো (māthāgulo) "heads", মাথারা (māthārā) "heads", মাথাদের (māthāder) "to the heads", মাথারই (māthāri) "of head itself" with moderate frequency. This corpus is exhaustively used to extract sentences of a particular word required for our system as well as for validating the senses evoked by the word used in the sentences.

V. Proposed Approach

In the proposed approach, the Naïve Byes probabilistic model has been used to disambiguate the sense of an ambiguous word, present in a particular sentence. As a data source, the Bengali text corpus developed in the TDIL project of the Govt. of India has been used. First, all the instances containing a particular ambiguous word, both inflected and non-inflected forms were retrieved from the corpus. Then few sentences carrying a particular sense of that ambiguous word were selected for preparing the learning sets and few sentences were selected for preparing the test set. Next, all these texts were passed through a series of manual normalization processes (refer Section V. *A*).

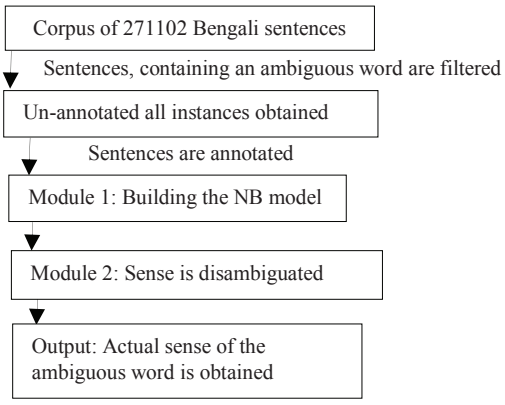

Fig. 1. The overall procedure is represented graphically.

A. Text annotation

At first, all the sentences containing a particular ambiguous word were extracted from the Bengali text corpus. As the sentences were not normalized adequately, these were passed through a series of manual normalization for (a) separation or detachment of punctuation marks like single quote, double quote, parenthesis, comma, etc. that were attached to words; (b) conversion of dissimilar fonts into similar ones; (c) removal of angular brackets, uneven spaces, broken lines, slashes, etc. from sentences; and (d) identification of sentence terminal markers (i.e., full stop, note of exclamation, and note of interrogation) that were used in written Bengali texts.

B. Modular representation of the proposed approach

The following modular representation depicts the overall procedure of sense disambiguation (refer Figure 1).

- **Explanation of Module 1: Building NB model**

In the NB model the following parameters were calculated based on the training documents:

i. |V| = the number of vocabularies, means the total number of distinct words belong to all the training sentences.
ii. P(ci) = the priori probability of each class, means the number of sentences in a class / number of all the sentences.
iii. ni = the total number of word frequency of each class.
iv. P(wi | ci) = the conditional probability of keyword occurrence in a given class.

To avoid "zero frequency" problem, Laplace estimation was used by assuming a uniform distribution over all words, as-

P(wi | ci) = (Number of occurrences of each word in a given class + 1)/(ni + |V|)

- **Explanation of Module 2: Sense Disambiguation**

To disambiguate an actual sense, the "posterior" probabilities, P(ci | W) for each class was calculated, as-

$$P(ci \mid W) = P(ci) \times \sum_{j=1}^{|V|} P(wj \mid ci)$$

The highest value of probability categorizes the test document into the related classifier.

For experimental purpose four mostly used ambiguous words in Bengali were used, as- "মাথা" (māthā), "তোলা" (tolā), "হাত" (hāt) and "দিন" (din) and the above approach was used to disambiguate each word in Lemmatized and Regular (without lemmatized) system .

VI. Results and Corresponding Evaluations

The algorithm was tested on 100 instances of 4 commonly used ambiguous words, as "মাথা" (māthā), "তোলা" (tolā), "হাত" (hāt) and "দিন" (din) and each word was disambiguated in Lemmatized and Regular (without lemmatized) system. The number of sentences selected for experiment purpose is given blow.

TABLE I. DETAILS OF INPUT DATA

Word	No. of	Senses considered

	sentences	
“মাথা” (māthā)	30	মস্তক, চিন্তা, প্রাপ্ত
“তোলা” (tolā)	30	উত্তোলনকরা, সৃষ্টিকরা, সংগ্রহকরা, উত্থাপনকরা, প্রত্যাহারকরা, অর্পণকরা
“হাত” (hāt)	20	অবদান, হাতপাতা হাতবদল, হস্ত,
“দিন” (din)	20	প্রতিদিন, দিবস, দিনকাটা, দেওয়া

The following steps were performed for output evaluation.

A. Preparing input data sets

Input data sets were provided in lemmatized and regular (without lemmatized) forms, as-

1) Regular input data: A partial view of the regular input data is presented in figure 2. In each sentence <Sentence x> tag represents the sentence number and <wsd_id=y, pos=z> tag represents the ambiguous word number in the sentence and corresponding Part-of-Speech of that ambiguous word in that particular sentence.

<Sentence 1> মার্কিন যুক্তরাষ্ট্রে ফিল্ম তৈরির সঙ্গে সঙ্গেই এদেশে হোটেল ব্যবসাতেও <wsd_id=1, pos=noun> হাত </wsd> লাগিয়েছেন বিজয় অমৃত রাজ।<Sentence 2> কোথাও সহানুভূতি নেই সহযোগিতার <wsd_id=1, pos=noun> হাত </wsd> এতটুকু প্রসারিত নেই একটি ভাল কথা কারও মুখ থেকে শোনা যায় না।<Sentence 3> এই সব কয়টি দেশেরই জাতীয় মুক্তি ও সাম্রাজ্যবাদ বিরোধিতার আন্দোলনে ভারত একনিষ্ঠভাবে পাশে দাঁড়াইয়াছে সাহায্য ও সহযোগিতার <wsd_id=1, pos=noun> হাত </wsd> প্রসারিত করিয়াছে।<Sentence 4> সঠিক ও বাস্তব মূল্যায়নের অভাব খাপছাড়াভাবে

Fig. 2. A partial view of regular input data.

2) Lemmatized input data: A partial view of the lemmatized input data is presented in figure 3. Here, the tags represent the same as regular text and all the words were lemmatized manually in the form- “word-instance/stem-word/POS”.

<Sentence 1> মার্কিন/মার্কিন/noun যুক্তরাষ্ট্রে/রাষ্ট্র/noun ফিল্ম/ফিল্ম/noun তৈরির/তৈরি/adj সঙ্গে/সঙ্গে/adj সঙ্গেই/সঙ্গে/adj এদেশে/দেশ/noun হোটেল/হোটেল/noun ব্যবসাতেও/ব্যবসা/noun <wsd_id=1, pos=noun> হাত/হাত/noun </wsd> লাগিয়েছেন/লাগানো/verb বিজয়/বিজয়/noun অমৃত/অমৃত/noun রাজ/রাজ/noun।<Sentence 2> কোথাও/কোথাও/prn সহানুভূতি/অনুভূতি/noun নেই/থাকা/verb সহযোগিতার/সহযোগিতা/noun <wsd_id=1, pos=noun> হাত/হাত/noun </wsd> এতটুকু/এত/prn প্রসারিত/প্রসারণ/noun নেই/থাকা/verb একটি/এক/adj ভাল/ভাল/adj কথা/কথা/noun কারও/কার/prn মুখ/মুখ/noun থেকে/থেকে/adj শোনা/শোনা/verb যায়/যাওয়া/verb না/না/abay।<Sentence 3> এই/এ/prn সব/সব/adj কয়টি/কয়েক/adj দেশেরই/দেশ/noun

Fig. 3. A partial view of lemmatized input data.

B. Preparing output data:

The actual output (refer Figure 4) was prepared initially, with the help of a standard lexical dictionary. When the output was generated by the proposed system, this was compared with that predefined output data using another program.

<Sentence 1> মার্কিন যুক্তরাষ্ট্রে ফিল্ম তৈরির সঙ্গে সঙ্গেই এদেশে হোটেল ব্যবসাতেও <wsd_id=1, pos=noun, sense=abadaan> হাত </wsd> লাগিয়েছেন বিজয় অমৃত রাজ।<Sentence 2> কোথাও সহানুভূতি নেই সহযোগিতার <wsd_id=1, pos=noun, sense=abadaan> হাত </wsd> এতটুকু প্রসারিত নেই একটি ভাল কথা কারও মুখ থেকে শোনা যায় না।<Sentence 3> এই সব কয়টি দেশেরই জাতীয় মুক্তি ও সাম্রাজ্যবাদ বিরোধিতার আন্দোলনে ভারত একনিষ্ঠভাবে পাশে দাঁড়াইয়াছে সাহায্য ও সহযোগিতার <wsd_id=1, pos=noun, sense=abadaan> হাত </wsd> প্রসারিত করিয়াছে।<Sentence 4> সঠিক ও বাস্তব মূল্যায়নের অভাব খাপছাড়াভাবে কলোনি উন্নয়নের কাজে <wsd_id=1, pos=noun, sense=abadaan> হাত </wsd> দেওয়া যথাযথ ব্যবস্থা না করে শিবির বন্ধ

Fig. 4. A partial view of output data.

The same actual output was used to compare the performance of the system in both the experiments.

C. Comparison of the outputs

The comparison of performances of Sense Disambiguation task in two types of systems is given below.

TABLE 2. RESULT WITH COMPARISON

Word	No. of sentences	Accuracy of result in without Lemm. system (%)	Accuracy of result in Lemm. system (%)
“মাথা” (māthā)	30	83	86
“তোলা” (tolā)	30	82	86
“হাত” (hāt)	20	80	85
“দিন” (din)	20	80	86

It was observed that with respect to a particular vocabulary, a regular (without Lemmatized) system gives a lesser accuracy, as all the inflected forms of a particular word in a regular text can’t match with a single vocabulary entry. But, in a lemmatized system, as the stems of the word-instances were considered, all the inflections of a particular word were mapped to its stem word, results the more number of matches with the vocabulary entry. Thus, the accuracy of the result was increased.

VII. Conclusion and Future Works

In this paper, with the four lexical examples, we have tried to show a lemmatized system can perform better due to greater lexical coverage. The results, obtained from the system was quite satisfactory according to our expectation. We argue that a stronger and properly populated learning set

would invariable yield better result. In future, we plan to disambiguate the most frequently used 100 ambiguous words of different parts-of-speech in Bengali.

Reference

[1] N. Ide and J. Véronis, "Word Sense Disambiguation: The State of the Art," *Computational Linguistics*, Vol. 24, No. 1, Pp. 1-40, 1998.

[2] R.S. Cucerzan, C. Schafer and D. Yarowsky, "Combining classifiers for word sense disambiguation," *Natural Language Engineering*, Vol. 8, No. 4, Cambridge University Press, Pp. 327-341, 2002.

[3] M. S. Nameh, M. Fakhrahmad, M.Z. Jahromi, "A New Approach to Word Sense Disambiguation Based on Context Similarity," *Proceedings of the World Congress on Engineering*, Vol. I. 2011.

[4] W. Xiaojie, Y. Matsumot, "Chinese word sense disambiguation by combining pseudo training data," *Proceedings of The International Conference on Natural Language Processing and Knowledge Engineering*, Pp. 138-143, 2003.

[5] R. Navigli, "Word Sense Disambiguation: a Survey," *ACM Computing Surveys*, Vol. 41, No.2, ACM Press, Pp. 1-69, 2009.

[6] R. Gaizauskas, "Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs," *Computer Speech and Language*, Vol. 12, No. 3, Special Issue on Evaluation of Speech and Language Technology, Pp. 453-472, 1997.

[7] H. Seo, H. Chung, H. Rim, S. H. Myaeng, S. Kim, "Unsupervised word sense disambiguation using WordNet relatives," *Computer Speech and Language*, Vol. 18, No. 3, Pp. 253-273, 2004.

[8] G. Miller, "WordNet: An on-line lexical database," *International Journal of Lexicography*, Vol.3, No. 4, 1991.

[9] S.G. Kolte, S.G. Bhirud, "Word Sense Disambiguation Using WordNet Domains," *First International Conference on Digital Object Identifier*, Pp. 1187-1191, 2008.

[10] Y. Liu, P. Scheuermann, X. Li, X. Zhu, "Using WordNet to Disambiguate Word Senses for Text Classification," *Proceedings of the 7th International Conference on Computational Science, Springer-Verlag*, Pp. 781 – 789, 2007.

[11] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller, "WordNet An on-line Lexical Database," *International Journal of Lexicography*, 3(4): 235-244, 1990.

[12] G.A. Miller, "WordNet: A Lexical Database," *Comm. ACM*, Vol. 38, No. 11, Pp. 39-41, 1993.

[13] A.J. Cañas, A. Valerio, J. Lalinde-Pulido, M. Carvalho, and M. Arguedas, "Using WordNet for Word Sense Disambiguation to Support Concept Map Construction," *String Processing and Information Retrieval*, Pp. 350-359, 2003.

[14] http://en.wikipedia.org/wiki/Naive_bayes

[15] N.S. Dash, Indian scenario in language corpus generation. In, Dash, Ni.S., P. Dasgupta and P. Sarkar (Eds.) Rainbow of Linguistics: Vol. I. Kolkata: T. Media Publication. Pp. 129-162 2007.

[16] N.S. Dash, "Corpus oriented Bangla language processing," Jadavpur Journal of Philosophy. 11(1): 1-28, 1999.

[17] N.S. Dash, "Bangla pronouns-a corpus based study," *Literary and Linguistic Computing*. 15(4): 433-444, 2000.

[18] N.S. Dash, Language Corpora: Present Indian Need, Indian Statistical Institute, Kolkata, 2004. http://www. elda. org/en/proj/scalla/SCALLA2004/dash. pdf.

[19] N.S. Dash, Methods in Madness of Bengali Spelling: A Corpus-based Investigation, South Asian Language Rewiew, Vol. XV, No. 2 2005.

[20] N.S. Dash, From KCIE to LDC-IL: Some Milestones in NLP Journey in Indian Multilingual Panorama. Indian Linguistics. 73(1-4): 129-146, 2012.

[21] N.S. Dash and B.B. Chaudhuri, "A corpus based study of the Bangla language," *Indian Journal of Linguistics*. 20: 19-40, 2001.

[22] N.S. Dash and B.B. Chaudhuri, "Corpus-based empirical analysis of form, function and frequency of characters used in Bangla," Published in Rayson, P. , Wilson, A. , McEnery, T. , Hardie, A. , and Khoja, S. , (eds.) Special issue of the Proceedings of the Corpus Linguistics 2001 Conference, Lancaster: Lancaster University Press. UK. 13: 144-157. 2001.

[23] N.S. Dash and B.B. Chaudhuri, Corpus generation and text processing, International Journal of Dravidian Linguistics. 31(1): 25-44, 2002.

[24] N.S. Dash and B.B. Chaudhuri,"Using Text Corpora for Understanding Polysemy in Bangla," *Proceedings of the Language Engineering Conference* (LEC'02) IEEE, 2002.

[25] L. Dolamic and J. Savoy, "Comparative Study of Indexing and Search Strategies for the Hindi, Marathi and Bengali Languages, " *ACM Transactions on Asian Language Information Processing*, 9(3): 1-24, 2010.