Rule Based Replacement of Pronoun by Corresponding Noun for Bangla News Documents

Md. Majharul Haque, University of Dhaka, Dhaka, Bangladesh
Suraiya Pervin, University of Dhaka, Dhaka, Bangladesh
Zerina Begum, University of Dhaka, Dhaka, Bangladesh

ABSTRACT

The object of this research work is to replace pronoun by corresponding noun for Bangla news documents. To the best of our knowledge, this is the first initiative to solve the problem of dangling pronoun where corresponding noun is not available. If the information retrieval procedures extract any sentence with dangling pronoun, it may raise confusion to the user. To mitigate this problem, a method has been proposed here by using general and special tagging, dependency parsing, full name identifying and finally pronoun replacing. For achieving the target of this method, 3000 Bangla news documents have been analyzed and some grammar books have been studied. Seven knowledgeable persons in the arena of Bangla language also helped us in this research work. Finally, the proposed method shows 71.80% accuracy in the evaluation for replacing pronoun.

KEYWORDS

Bangla News Documents, Corresponding Noun, Dangling Pronoun, Dependency Parsing, Information Retrieval

INTRODUCTION

Bangla is the 7th most spoken language in the world from 3500 languages and around 250 millions of people are using Bangla (Chowdhury, Khalil, & Mofazzal, 2000). It is the mother language of Bangladesh (Islam, 2003) and the second most spoken language in India (Olivet, 2015). Based on the economic survey - 2015, there are 62.30% literate people in Bangladesh where most of them are used to Bangla language only. Nowadays, many computerized contents are being developed in Bangla and online version of Bangla newspaper is also growing rapidly. So, electronic version of Bangla text is increasing without any bounds in the cyber world and people are overloaded with huge volume of texts. To alleviate this burden of large volume of text, very few research works have been conducted for Bangla (Islam & Masum, 2004; Sarkar, 2014; Haque, Pervin, & Begum, 2016). In this situation, the Bangla-speaking people have been deprived from the advantage of information technology. So, for this large community of Bangla-speaking people, more research work is very much necessary especially for Bangla information retrieval. But research works for Bangla language is difficult for the following issues:

DOI: 10.4018/IJTD.2017040102

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.
1. Based on our study, automatic procedures are hardly available for Bangla language to facilitate research work.

2. For Bangla language, there is no lexical database like WordNet (Miller, 1995). Though a similar tool is ongoing to be developed it has limited features (Indian Statistical Institute, 2015).

3. There is no database of ontological meanings for Bangla words that can be used programmatically.

4. Subject and object of all sentences need to be identified for proper recognition of structures of sentences which is complex in Bangla than that of English. Because, the placement of subject in English sentence is generally before the verb phrase, auxiliary verb or it may appear after the word ‘by’ in passive voice but subject may be existed in several places in Bangla sentence.

Some other problems have been discussed in (Karim, Kaykobad, & Murshed, 2013; Zaman, 2015) about the research work on Bangla. Even, the scope of knowledge sharing is also limited as there are a few researchers in this arena. Despite these difficulties, a method has been presented here which focuses upon a problem in the output of Bangla information retrieval procedures.

In the output of information retrieval, some sentences may be available with dangling pronoun(s) where the corresponding noun is missing. These pronouns will make the information incoherent. So, the systems, which have been developed for burden minimization from large volume of text, may deliver wrong message. Other than receiving a direction, the user will often be misguided with misinformation.

The objective of this research work is to make the output of Bangla information retrieval procedures free from dangling pronoun. Otherwise, there is a huge probability to misunderstand the text by user because only a single dangling pronoun is enough to deliver wrong message. In these circumstances, a method has been proposed here to resolve this problem with the following major contributions:

1. Detecting the nature of each word in the sentence of Bangla news document as noun, pronoun, verb, subject, object, numerical figure, acronym, name of people and places, etc. In this regard, words are tagged in two phases as general and special tagging.

2. Dependency parsing of words to verify the nature of each word because a word may have dynamic nature for the effect of surrounding words. Some untagged words are also tagged using dependency parsing.

3. Locating pronouns and distinguishing each of them as subject or object.

4. Identifying the corresponding noun of the pronoun and replacing the pronoun in suitable format.

Based on the authors’ findings, this is the first initiative for the replacement of pronoun in Bangla. For English language, one system has been developed by Stanford NLP (Stanford CoreNLP, 2015) for noun-pronoun matching but this can’t be used for Bangla as the structure of Bangla sentence is much different from English (Chowdhury et al., 2000).

To accomplish this research work, the authors have scrutinized 3000 Bangla news documents (news documents of around one month from the Daily Prothom-alo which is the most popular newspaper of Bangladesh). Seven knowledgeable persons of Bangla language, who have completed four years’ graduation on Bangla language (their mother tongue is Bangla and they read Bangla newspapers regularly), helped in this research work. After a detail discussion with those persons regarding the structure of sentences of Bangla language and analysis of news documents, some rules have been utilized here. Based on these rules, special tagging, dependency parsing, subject and object recognizing and after all pronoun replacing have been accomplished.
Related Content

A Qualitative Study on the Adoption of Bring Your Own Device (BYOD) Practice
[www.igi-global.com/article/a-qualitative-study-on-the-adoption-of-bring-your-own-device-byod-practice/146574?camid=4v1a](www.igi-global.com/article/a-qualitative-study-on-the-adoption-of-bring-your-own-device-byod-practice/146574?camid=4v1a)

“Evaluator”: A Grading Tool for Spanish Learners
Paz Ferrero, Rachel Whittaker and Javier Alda (2013). *Technologies for Inclusive Education: Beyond Traditional Integration Approaches* (pp. 244-269).
[www.igi-global.com/chapter/evaluator-grading-tool-spanish-learners/71877?camid=4v1a](www.igi-global.com/chapter/evaluator-grading-tool-spanish-learners/71877?camid=4v1a)
Economic Growth, Technical Progress and Labor Productivity: Knowledge Economics and New Forms of Technical Progress
www.igi-global.com/article/economic-growth-technical-progress-labor/51581?camid=4v1a

Understanding Crowdsourcing of Agricultural Market Information in a Pilot Study: Promises, Problems and Possibilities (3Ps)
www.igi-global.com/article/understanding-crowdsourcing-of-agricultural-market-information-in-a-pilot-study/188699?camid=4v1a