DECLARATION

We hereby declare that this thesis is based on the results found by ourselves. Materials of work found by other researcher are mentioned by reference. This thesis, neither in whole nor in part, has been previously submitted for any degree, but this paper is heavily influenced by one of our previous works targeted for an international conference which was never published.

Signature of
Supervisor

Signature of
Authors

ACKNOWLEDGMENTS

Special thanks to Dr. Mumit Khan & Matin Saad Abdullah who made us curious about natural language processing and WordNet in particular. We are also grateful to all the members of Center for Research on Bangla Language Processing (CRBLP), BRAC University, who endlessly supported us throughout the research. We must also mention the name of one of the members of CRBLP, Naushad UzZaman, who helped us extensively during our background study for the research.

ABSTRACT

This paper presents a method for creating a Bangla wordnet, semi-automatically from other source wordnets and Bilingual Dictionaries. The method requires an extensive involvement of Bangla Linguists to make the effort effective, and should be helpful when the computational resources at hand for a language like Bangla are very limited. It is a bootstrapping method where the cycle starts with some portion of the target WordNet be created automatically, then linguists edit the mistakes on that portion and then that corrected information is also used to generate next set of target entries. Our success will be evaluated against an (automatically building a Bangla WordNet using Princeton WordNet and a small test purpose bilingual dictionary).

TABLE OF CONTENTS

Page

1. INTRODUCTION

Bangla is among the top ten most widely spoken languages [1] with more than 200 million native speakers, but unfortunately very little research efforts has been made so far to enrich the language processing resources for Bangla. Furthermore, not having sufficient computational language resource for Bangla hinders the research efforts in this particular area of Bangla Language Processing. In contrast, English language has a rich set of language processing resources. We can use those to build or at least to start off building resources of Bangla language.

The importance of a wordnet for NLP applications can hardly be overestimated [2]. The Princeton WordNet (PWN) is now a mature lexical ontology which has demonstrated its efficiency in a variety of tasks (word sense disambiguation, machine translation, information retrieval, etc.). Inspired by the success of PWN many languages started to develop their own wordnets taking PWN as a model (cf. http://www.globalwordnet.org/gwa/wordnet_table.htm). Furthermore, in both EuroWordNet and BalkaNet projects the synsets from different versions of PWN (1.5 and 2.0) Ire used as ILI repositories. The created wordnets Ire linked by means of interlingual relations through this ILI repository. The rapid progress in building a new wordnet and linking it with an already tested wordnet (usually PWN) is hindered by the amount of time and effort needed for developing such a resource. To take a recent example, the development of core wordnets (of about 20000 synsets, as is the case with the Romanian wordnet) for Balkan languages took three years (2001-2004).

In what follows we present a methodology that can be used for automatically building wordnet for Balngla, strictly aligned (that is, using only EQ_SYNONYM relation) with an already available wordnet - PWN.

We call the wordnet already available Source wordnet (as mentioned before, this is usually a version of PWN) and the wordnet to be built and linked with the Source wordnet will be named Target wordnet. The methodology we present has three basic phases. In the first one some of the synsets for the target language are automatically generated and mapped onto the source language synsets using a bilingual dictionary (Bangla-English-Bangla). In the second phase the linguists must involve themselves to

correct the errors produced in the output from the target wordnet. In the third phase both the bilingual dictionary and the partial target wordnet is used to generate the next set of synsets. The paper has the following organization. Firstly we stated why we chose this approach rather than the conventional ones in the Literature Review section. Next, we state the implicit assumptions in building a wordnet strictly aligned with other wordnets. Then we shortly describe the resources that one needs in order to apply the method, and also the criteria we used in selecting the source language test synsets to be implemented. Finally, we state the problem to be solved in a more formal way and tried to implement as much as we can, so that the feasibility of such an effort can be measured.

2. LITERATURE REVIEW

Many different approaches have been made to build a wordnet based on an already built wordnet. Most of those are top down approaches [2, 3, 4]. They used a source wordnet as the starting point, and use that to generate a target wordnet by mapping the synsets of the two languages. This approach has a serious drawback. Concepts that are available in the source wordnet must have corresponding concepts in the target wordnet. Moreover, it requires a significant amount of resource (the set of sysnsets, strictly aligned with the source wordnet sysnsets) that we currently do not have in hand. A simpler approach is a bottom-up approach where we start with the words in the target language, rather than starting from the source wordnet. The process will be described in detail later in this paper.

3. ASSUMPTIONS

Firstly, we are assuming that English and Bangla have a significant amount of linguistic similarity that will cause this process to be a success. Secondly we assume there are Bangla word senses that can be clearly identified in the Bangla-English-Bangla Dictionary. This assumption is implicit when one builds a wordnet aligned or not with other wordnets. This premise was extensively questioned among others by Kilgarriff who thinks [5] that word senses have not a real ontological status, but they exist only relative to a task. We will not discuss this issue here.

4. RESOURCES NEEDED

To make this effort a success a well defined Bangla-English-Bangla dictionary, that describes the sense of a word as precisely as possible. This is important because these are the senses that are to be used to translate the source wordnet structure to our target wordnet. Another requirement is an interface to query the PWN to retrieve graphs of synsets and enter those graphs to our wordnet for Bangla.

5. PROCEDURE

The process we will describe here is fairly simple. For each Bangla word in the dictionary, we need to look up all possible English words. Then we out find out the synsets for those English words from PWN, extract the whole network of those synsets and copy that to our target wordnet for Bangla. Then, we try to translate the structure where ever possible, like name of the features attached with each word/synset, the features of these words and of course the actual words into Bangla. We agree that this will not be a perfect translation as many of the English synsets may not have direct translations in Bangla. In that case, we keep the structure unchanged where we cannot translate (words/sentences in English that cannot be translated are left unchanged). It's the job of the linguists to fix those errors and translate those into Bangla. After the linguists have edited the first set of sysnsets and approved that for later use, we can use this target wordnet as a resource for translation for the next iterations over the rest of the Bangla words. This might eventually increase the translation capability for the rest of the words in Bangla. This process keeps on going until all the words in the selected source dictionary has been explored and evaluated.

6. IMPLEMENTATION

As proper digitized Bangla-English-Bangla dictionary was not available at the time of this implementation, we used a small test purpose Bangla-English Dictionary (one-to-one) to bootstrap our program. We then availed an SQL script file that unifies WordNet 3.0, WordNet 2.0-2.1, 2.1-3.0, 2.0-3.0 sensemaps, VerbNet 2.1, XWordNet 1.1 compiled by Bernard Bou. A piece of software was written to translate words directly on that SQL script file, and then the partially translated script file was used to generate a MySQL database. In this first phase, about 650 words were translated automatically.

For the second phase, the manual translation by the linguists, we wrote a database editor for this SQL formatted WordNet3.0 database using PHP. The application is AJAX enabled, through which one can easily translate words, sense definitions and associated sample sentences into Bangla provided that a Bangla Unicode input method is present. Our research would be successful, if the linguists can use this translator effectively and easily.

7. FUTURE WORK

Based on this proposal, in the future one can try out this approach and may end up with a rich wordnet for Bangla. Or, one can do more extensive linguistic research on the both languages to find out, if this effort will be feasible or not.

REFERENCES

[1] The Summer Institute for Linguistics (SIL) Ethnologue Survey (1999). http://www.sil.org/sil/

[2] Barbu, Eduard and Verginica Barbu Mititelu, ``Automatic Building of Wordnets" In: Proceedings of the International Conference Recent Advances in Natural Language Processing , pp. pp. 329-332, Borovets, Bulgaria, 21-23 September 2005 https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2005/RANLP/papers/89_barbu .pdf

[3] Chakrabarti, Debasri and Pushpak Bhattacharyya, ``Creation of English and Hindi Verb Hierarchies and their Application to Hindi WordNet Building and English-Hindi MT" In: Proceedings of the Second Global WordNet Conference , pp. 83-90, Brno, Czech Republic, January 20-23,2004. http://www.fi.muni.cz/gwc2004/proc/125.pdf

[4] Farreres, Xavier, German Rigau and Horacio Rodriguez, ``Using WordNet for building WordNets." In: Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, 1998. http://xxx.lanl.gov/abs/cmp-lg/9806016

[5] (Kilgarriff 1997) A. Kilgarriff, We don’t believe in word senses. In Computers and the Humanities, 31(2), 91-113, 1997. *http://citeseer.ist.psu.edu/73081.html*