Research Report on Parallel Corpus Translation Challenges and Processes

Country Component	Bangladesh	**Report no.**	QPR – 01
Phase no.	1.1	**Report Ref no.**	PANL10n/Admn/QPR/001

Prepared By: Mumit Khan_____ **Designation: Project Leader___**
(Name)

Project Leader: Mumit Khan_____ **Signature: _____**
(Name)

Date: October 8, 2007_____

TABLE OF CONTENTS

Title

Parallel Corpus Translation Challenges and Processes

Matin Saad Abdullah, Naira Khan, and Mumit Khan
Center for Research on Bangla Language Processing
BRAC University

Abstract

We describe some of the challenges in developing English-Bangla parallel corpora, and look some of the established processes used by other language corpora for solutions to some of these challenges.

Introduction

Parallel Corpora is increasingly becoming an essential tool in various advanced language processing applications, starting from statistical machine translation to automatic dictionary creation. There have been significant work in creating Parallel Corpora for related languages, and the processes are now well established. However, creating Parallel Corpora for unrelated languages remain a challenge, and only now taking center stage in the research on linguistic resource (Garside, et. al., 1994; Hutchinson, Leech, McEnergy Collier and Takahashi, 1995; Koehn, 2002; Samy, Moreno-Sandoval, Guirao, 2005; Samy, et. al., 2006). There are a few challenges in building the Parallel Corpora for any two languages: (1) preprocessing the source texts, a series of steps that would convert the texts to a standard encoding, clean the texts, and define the word/sentence/paragraph boundaries, (2) aligning the corpora at paragraph/sentence/word level, and finally (3) tagging the corpora for part of speech. The challenges become more acute if the languages in question are not related, since the word alignment process may not be able to take advantage of cognateness between the two languages (Megyesi, Hein and Johanson, 2006); in addition, the differences in word ordering may also impact the performance and accuracy of the various alignment processes. In the following sections, we describe the challenges and processes for creating Parallel Corpora for English and Bangla, starting with English source text that is already available in standard encoding, and cleaned.

Pre-processing and Alignment

The pre-processing typically involves the following steps:

1. Converting source texts to a standard encoding such as Unicode;
2. Segmenting the source texts into paragraphs and sentences; and
3. Tokenizing the source texts into words.

For the English language source text, these processes are well established, and there are existing tools such as Uplug (Tiedemann, 1999) to automate the processes. For the Bangla language target text, since the translations are being done as part of the project, it is going to be available in Unicode. One of the challenges would be to develop the tools to for

segmentation and tokenization. However, these are not expected to be significant roadblocks, as the rules for Bangla segmentation and tokenization are well understood.

The alignment process presents a much larger challenge, one that will have to be investigated thoroughly as this is the first published effort at creating an English-Bangla Parallel Corpora. The process for automatically aligning the English Corpus is established by the literature, starting with aligning paragraphs and sentences using the length-based algorithm of Gale and Church (1993). This is then followed by aligning phrases and words using the Clue alignment approach of Tiedemann (Tiedemann 2003), and using GIZA++ (Och and Ney, 2003). However, these processes may not work as well for Bangla, given the difference in typical sentence lengths, differences in word order, and phrasal translation strategies. This problem is seen for the Turkish target text alignment in a Swedish-Turkish Parallel Corpus, where 31% of the words were misaligned using automatic alignment techniques presented earlier (Megyesi, Hein and Johanson, 2006). Fortunately, there is some existing work on alignment for Hindi, which may well be applicable to Bangla, a sibling language (Aswani, 2003).

POS Tagging

The last step involves tagging the corpora at the part-of-speech level. Again, there are extensively tested models and tools available for English, reaching accuracies of up to 96%. However, the existing automatic taggers for Bangla have been unable to achieve anything close to that, barely approaching 60% using a small training corpus (Hasan, UzZaman and Khan, 2006). One of our challenges would be to manually tag the translated Bangla corpus, using perhaps multiple human taggers to minimize errors.

Conclusion

We describe the major challenges in creating a English-Bangla parallel corpus, and look at some of the established processes used when creating parallel corpora for some unrelated languages such as Spanish-Arabic and Swedish-Turkish.

References

Aswani, N. (2003). Aligning Sentences And Words Using The English-Hindi Bilingual Parallel Corpora. Masters Thesis. The University of Sheffield.

Collier, N. and Takahashi, K. (1995). Sentence alignment in parallel corpora: The asahi corpus of newspaper editorials. Technical Report 95/11, Centre for Computational Linguistics, UMIST, Manchester.

Gale, W. and Church, K. (1993). A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19(1), 75 102.

Garside, R., Hutchinson, J., Leech G., McEnery A. M. and Oakes M. (1994): The exploitation of parallel corpora in projects ET10/63 and CRATER. In Jones D.B. (Ed.), New Methods in Language Processing, Manchester: UMIST, pp. 108-115.

Hasan, F., UzZaman, N. and Khan, M. (2006). Comparison of different POS Tagging Techniques (n-gram, HMM and Brill's tagger) for Bangla. In Proceedings of ICS2E.

Koehn, P. (2002). Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Information Sciences Institute, University of Southern California.

Megyesi, B., Sågvall Hein, A. and Csató Johanson, É. (2006). Building a Swedish-Turkish Parallel Corpus. In Proceedings of Language Resources and Evaluation Conference. May 22-28, 2006. Genoa, Italy.

Och, F. J., and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, volume 29, number 1, pp. 1951 March 2003.

Samy, D., Moreno-Sandoval, A. and Guirao J. M. (2005). A Proposal for an Arabic Named Entity Tagger Leveraging a Parallel Corpus (Spanish-Arabic). In Proceedings of International Conference on Recent Advances on Natural Language Processing RANLP 2005, Borovets, Bulgaria, pp. 459-465.

Samy, D., Moreno-Sandoval, A., Guirao, J. M. and Alfonseca, E. (2006). Building a Parallel Multilingual Corpus (Arabic-Spanish-English). In Proceedings of Language Resources and Evaluation Conference. 2006.

Tiedemann, J. (1999). Uplug A Modular Corpus Tool for Parallel Corpora. In Lars Borin (ed.), 2002, Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University. Amsterdam: Rodopi.

Tiedemann, J. (2003). Recycling Translations Extraction of Lexical Data from Parallel Corpora and their Applications in Natural Language Processing. PhD Thesis. Uppsala University.