An Experiment to Create Parallel Corpora for Odia

Rakesh Balabantaray, PhD.
Asst Prof., CLIA Lab, Dept. of CSE, IIIT,
Bhubaneswar, Odisha, India.

Deepak Sahoo
Project Fellow, CLIA Lab, IIIT,
Bhubaneswar, Odisha, India

ABSTRACT
The term parallel corpora are typically used in linguistic circles to refer to texts that are translations of each other. And the term comparable corpora refer to texts in two languages that are similar in content, but are not exact translations. In order to exploit a parallel text, some kind of text alignment, which identifies equivalent text segments (approximate sentences), is a prerequisite for analysis. Parallel corpora are very much essential in cross lingual or multilingual information retrieval. This paper presents an approach for automatic creation of English-Odia parallel corpus from comparable corpus. Generally Named entities, Proper nouns and common nouns play an important role in information retrieval. We tried to find the effectiveness of named entities, Proper nouns and common nouns in aligning English – Odia comparable document pair. We have taken the Odia parallel corpus (152 English-Odia documents) from TDIL, as well as we have crawled comparable Wikipedia pages for testing and the results are encouraging. We have used Stanford coreNLP tool and Google translator in our work.

KEYWORDS
Cross lingual information retrieval, Named entity, Comparable document, document similarity, key terms.

1. INTRODUCTION
As we know Odia is an official language of Odisha and it is the first language of more than thirty million people and second and third language of many people of India. Odisha and Odia language is bound by three distinct language areas that are Telugu in the south and south-west, Chattishgarhi (Hindi) in the west and Bengali in the north. It is also significance to say that the internal linguistic composition of Odisha state consists of many tribal groups, who linguistically belong to two distinct language families-Dravidian and Munda. So, the name entities in these languages have been nativised through borrowing in Odia. This language has history of more than one thousand and ten years. It is also rich in literature for which it deserves reconisation of classical language. But good quality and high volume of parallel corpora with Odia and English does not exist which motivates us to create an English-Odia comparable corpora.

Recently, there has been a surge of interest in the automatic creation of parallel corpora. A parallel corpus consists of document pairs that are more or less exact translations of each other but in a comparable corpus, the document pairs are not exact translations but both documents talk about same topic. Comparable corpora as a source of translation knowledge have attracted the attention of many researchers. Comparable corpora [1] are composed of document pairs describing the same topic in different languages. They are not parallel(mostly sentence -to- sentence translated) corpora composed of good bilingual sentence pairs, but still contain various levels of parallelism such as named entities, Proper nouns, common nouns and some key word.

2. RELATED WORK
P.Sheridan (et. al, 1996) [1] talked about an approach to multilingual information retrieval based on the use of thesaurus-based query expansion technique applied over a collection of comparable multilingual documents.

Now a days, NLP and IR researcher's interest to use Wikipedia as a prime resource is increasing. Wikipedia is a multilingual encyclopaedia that is online and freely available and it is developed for most of the world's language. So we can use Wikipedia as prime resource to build bilingual parallel corpora by extracting Bilingual terms from Wikipedia. [2]

Sunita Arora (et. al. 2010) [3] have discussed about two-pass approach at sentence level, for automatic creation of Hindi-Punjabi parallel corpus from comparable corpus.

Dragos Stefan Munteanu (et. al. 2006) [4] have discussed method for extracting parallel sub-sentential fragments from comparable, non-parallel bilingual corpora. By analyzing potentially similar sentence pairs using a signal processing inspired approach. They detect which segments of the source sentence are translated into segments in the target sentence, and which are not.

Braschler (et. al 1998) [5] discussed a method to find similarity at document level in different languages. They have used a set of indicators to find the similarity between multilingual documents. Such indicators include presence of proper nouns, numbers, dates etc.

3. RESOURCES
As we are experimenting to know the effect of proper nouns, common nouns and Named Entities to create parallel corpus for Odia. For our experiment we have taken two set of test data. First set (100 out of 152 English-Odia parallel documents) collected from TDIL on request. For second set, we collected 50 comparable English-Odia Wikipedia urls. Then we crawled the urls and dumped it. Then from the dumped data we manually created 50 pair of English Odia comparable documents. The statiststics of the test resource is given below.

Total No of Document pairs in Corpus	Total Number of sentences in Corpus		Total Number of Tokens in Corpus	
	English	Odia	English	Odia
100 English-Odia document form TDIL	10000	10000	244650	190207

50 paired wiki pages	46721	2325	201283	38087

6. SYSTEM ARCHITECTURE

To find a parallel document between an English document (E) and an Odia document (O), we have taken three different attribute set to extract important terms from comparable English document.

Set1 – Named Entities (only PERSON, LOCATION and ORGANIZATION);

Set2 - Named Entities (PERSON, LOCATION and ORGANIZATION), Proper nouns (But not tagged as NE by Stanford CoreNLP NER*)

Set3 - Named Entities (PERSON, LOCATION and ORGANIZATION), Proper nouns (But not tagged as NE by Stanford CoreNLP NER*) and common nouns.

We first extract important terms from English comparable corpora, based on a particular attribute set given above.

Then we used Google translate to translate the extracted important English terms to Hindi and then we have java code to translate Hindi to Odia (as English to Odia translate option is not there with Google translate and the accuracy of translation of English to Odia is very low).

Now we have list of important terms in Odia that is present in comparable English documents. Then we search in the comparable Odia document to get an approximate match of the English document.

An Odia document (O) is matched with an English document (E) if

The difference between the size of important term list of English document (translated to Odia) with that of the important term list of a comparable Odia document.

And

Frequency of important terms of the Odia document is higher than rest of the Odia documents.

For Example

Suppose we want to find a parallel Odia document for an English document E1.

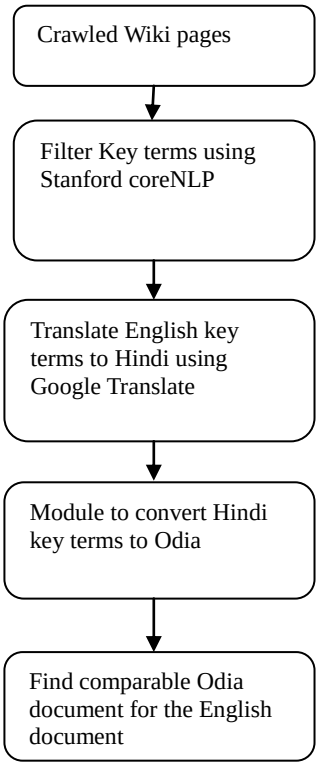
Let E1 have 10 important terms identified by Stanford CoreNLP NER (Then translated to Hindi and then to Odia).

Let there are 5 Odia documents as below

Document No	No. Of Important Terms	Total occurrence of important terms
OD1	5	8
OD2	4	10
OD3	6	12
OD4	3	8
OD5	6	14

In our case we choose document OD5 as parallel document to E1 as difference of No. Of important Terms are minimum and frequency of important terms are maximum.

FIGURE



7. RESULT AND DISCUSSION

To test the system we have two set of test data.

1St set of test data have 100 English-Odia parallel documents which we got it from TDIL on request.

2nd set of test data, we have crawled 50 comparable English-Odia Wikipedia pages.

We experiment both the test data set, with all the three attribute sets.

Experiment with 1st set of important term list (only PERSON, LOCATION and ORGANIZATION).

*http://nlp.stanford.edu/software/corenlp.shtml

No of Documents	No of documents properly matched	Accuracy
100(Parallel data from TDIL)	84	84%
50(Crawled wiki pages)	21	42%

Experiment with 2nd set of Important term list (PERSON, LOCATION and ORGANIZATION, Proper nouns (But not tagged as NE by Stanford CoreNLP NER*).

No of Documents	No of documents properly matched	Accuracy
100(Parallel from TDIL)	87	87%
50(Crawled wiki pages)	18	36%

Experiment with 3rd set of Important term list (PERSON, LOCATION and ORGANIZATION), Proper nouns (But not tagged as NE by Stanford CoreNLP NER*) and common nouns.

No of Documents	No of documents properly matched	Accuracy
100(Parallel from TDIL)	89	89%
50(Crawled wiki pages)	13	26%

One important thing we have observed from this experiment that, when the No. of features increases to extract the important terms from English document the accuracy of matching document decreases. The Reasons are

a. Content in Odia document is very less in comparison to the corresponding English document.

b. Some important terms are not properly translated to Odia.

For example

English Term	Hindi Term	Odia (Wrongly Translated)	Odia (Actual Translation)
Cancer	कैंसर	କ୍ୟାନସ୍ତର	କର୍କଟ
Advocate	वकील	ବକୀଲ	ଓକିଲ
Death	मौत	ମୌତ	ମୃତ
Daughter	बेटी	ବେଟୀ	ଝିଅ
Oriya	उड़िया	ଉଡିୟା	ଓଡ଼ିଆ
Orissa	उड़ीसा	ଉଡାସା	ଓଡ଼ିଶା

c. There are junk data in both English and Odia crawled data.

8. CONCLUSION AND FUTURE WORK

This is our initial work to create English-Odia parallel corpus. As we do not have a good NER for Odia and the accuracy to translate English named entities to Odia is very low, So in this work we have used Google translate to translate English named entities to Hindi, then we have a module to convert Hindi to Odia.We are trying to develop a English to Odia Translator which will translate English named entities to Odia with more accuracy so that the accuracy of document alignment will also increase. Another problem is that the content of Odia document is very less in comparison to the corresponding English document. We are also planning the sentence level similarity and hope that accuracy of document alignment will also increase.

9. ACKNOWLEDGEMENT

We are much indebted to the Department of information Technology (DIT), Ministry of Communication and Information Technology (MCIT), Govt. of India for this research work.

10. REFERENCE

[1] P.Sheridan & J.P.Ballerini, “Experiments in Multilingual information retrieval using the SPIDER system”, SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp.58-65.

[2] J.R. Harsita, R.Kotagiri and V. Padu (Eds.): DaSFAA 2008, LNCS 4947, pp. 380-392. 2008. Springer-Verlag Berlin Heidelberg 2008.

[3] Sunita Arora, Rajni Tyagi, Somi Ram Singla: “Creation of Parallel Corpus from comparable Corpus” Proceedings of ASCNT – 2010, CDAC, Noida, India, pp. 77 – 83.

[4] Dragos Stefan Munteanu, Daniel Marcu ”Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora ” Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 81–88, Sydney, July 2006. C 2006 Association for Computational Linguistics

[5] Braschler. M, Scauble. P.: “Multilingual information retrieval based on document alignment techniques”. Research and Advanced Technology for digital Libraries, 513-518 (1998).

*http://nlp.stanford.edu/software/corenlp.shtml