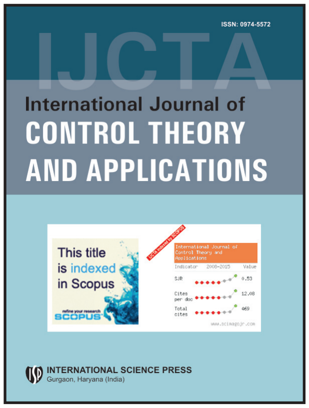

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 24 • 2017

Detail Survey on Parallel Corpora Generation Techniques

Chandramma[a], Swathi.K[b], Arun Biradar[c] and Piyush Kumar Pareek[d]

[a,b]Research Scholar, Dept of CSE R&D, East West institute of Technology, Bengaluru, VTU, Belagavi, India. Email: [a]rchandramma.vkit@gmail.com; [b]k.swathi980@gmail.com
[c]Research Supervisor, Visvesvaraya Technological University, Belgavi, Professor & HoD, Department of CSE, EWIT, Bengaluru. Email: hodcsea@gmail.com
[d]Research Guide, Visvesvaraya Technological University, Belgavi & Associate Professor, Department of CSE, EWIT, Bengaluru. Email: piyushpareek88@gmail.com

Abstract: This paper defines the study of the parallel corpus among the Natural Language processing applications. The parallel corpus means a text which can be obtained in a couple of languages, it might be a text that is initially its interpretation or it could be a text which includes being published by a consortium of writers in many different languages then posted in several language variations. Synchronous corpora are really a valuable way to obtain some sort of linguistic Meta knowledge as well as which will be commonly beneficial in numerous language that is normal. In this paper also explore the methods which are various constructions of synchronous corpora like Dictionary Based technique, POS Annotation Method, Cross Language Suggestions Retrieval Framework technique and Bootstrapping method.

Keyword: Parallel corpus, Natural Language Processing, Dictionary, Bootstrapping.

1. INTRODUCTION

Naturall Language Processing (NLP, also called Computational Linguistics) is really an industry of Computer Science which aims to come up with and understand the language that humans used to communicate. NLP carries a complete quantity that is big of tasks and applications; considered one of them is Machine Translation, which aims to straight away transform a text in a single language to a new one. To make Machine Translation possible, there are many practices based on dictionaries, information or examples. Even though these methods have actually real their advantages which are specific downsides, they share some techniques like text procedure that is positioning. Text procedure that is aligned in organizing parallel corpus in order to start an interaction between paragraphs, sentences and/or words from the supply texts and there, synchronous corpora are really crucial resources for tasks in the interpretation industry like linguistic studies, information retrieval systems development or language processing that is normal. These resources have to be accessible in reasonable quantities, because an application technique which is often the majority is dedicated to information to be useful. The typical associated with result depends plenty that is entire the dimensions concerning the corpora, meaning that the tools being robust had a need to build and process them.

2. PARALLEL CORPORA KINDS

To talk about text that is synchronous and perceive positioning dilemmas, we shall start by pointing away some interpretation traits. As presented in (Abaitua, 2000), we could classify translations based on the dependency between your text that is initially its interpretation:

- **Type A:** Once the translated text will totally replace the writin initially to the prospective language. This is actually the complete instance of literary translations (where visitors will elect to read only 1 form of them);
- **Type B:** Whenever translations will coexist in an area and time. Here is the instance of bilingual editions being literary. (where your reader will likely compare the texts in both languages);
- **Type C:** As soon as the translations are useful for exactly the same function once the initial, and work with a means that is symmetrical.

They are the situation for institutional papers associated with along with other organizations being multilingual or classify all of them with respect towards the interpretive goal:

- **Pragmatic:** The translated text is going to be employed for the interaction that is just like the first;
- **Stylistic:** The translated text attempts to keep up with the text that is initial and kind of language;
- **Semantic:** The translated text tries to transmit essentially the same message.

Parallel text alignment problems are highly dependent on these classifications:

- **Type A:** Translations can't be regarded as synchronous corpora. The translator usually changes your order of sentences plus some content8 when they keep up with the proven fact that is fundamentally the writing;
- **Type B:** Translations give reasonable outcomes on term positioning, because so many terms which can be particularly the corpora is supposed to be coherently translated between sentences;
- **Type C:** Translations are the very best variety of synchronous corpora for positioning. As this sort of synchronous corpora is usually consists of institutional papers with regulations as well as other information that is essential interpretation is completed accurately, to ensure that no ambiguities are placed into the text, in addition they keep symmetrical coherence.

3. TECHNIQUES UTILIZED IN THE GROWTH OF SYNCHRONOUS CORPUS ARE LENGTH BASED TECHNIQUE, LOCATION BASED TECHNIQUE AND LEXICAL BASED TECHNIQUE [3]

- **Length Based Method:** The strategy will align the sentences that are parallel fast sentences will be translated as brief sentences and long sentences such a long time sentences on the basis of the level of the sentences.
- **Location Based Method:** In this system texts which can be synchronized extracted based regarding the place of this sentence into the two texts. This method doesn't make an effort to align beads of sentences, but alternatively it aligns place offset in the 2 texts which are parallel.
- **Lexical Based Strategy:** In this system, it accounts the info that is lexical texts. Bilingual corpus can be used to fit this content terms in a single text using their correspondences into the other text and make use of these matches as fix points into the sentences procedure that is positioning.

4. METHODS FOR CONSTRUCTION OF SYNCHRONOUS CORPORA

Synchronous generation that is corpora a challenging task compare to aligning the sentences in synchronous corpora. The option of comparable corpora is more compared to the availability of synchronous corpora.

The strategy which are various creations of Parallel corpora depending on POS Annotation, Dictionary, built, Cross Language Suggestions Retrieval and bootstrapping practices

- **Dictionary Based technique:** The dictionary is definitely an electronic resource file is comprised of the language of the Language and their comparable terms in a language that is significantly different. Yulia Tsvetkov and Shuly Wintner [7] have actually proposed the purchase that is automated of Corpora from sites with Dynamic information. This method presents a totally automated construction of constantly growing corpora that is parallel. Right here it offers a straightforward and dictionary that is a beneficial algorithm to draw out parallel document pairs from the big number of articles that are retrieved on the internet with possibly containing manually translated texts.
- **Parallel Corpora Builder:** Parallel Corpora Builder (PCB), was created to get a corpus that is synchronous internet sites with powerful content which possibly have translated texts.
- **Online Crawling:** A Cron work can be used to perform a crawler many times every day also to harvest all articles which can be fresh. Right here it works with a script that is easy down load website pages from HTML tags and draw out just text and metadata like date, domain, supply Address etc.
- **Recognition of synchronous articles:** Yulia Tsvetkov, Shuly Wintnerrun a content-based contrast of most documents that is Hebrew-English which were collected through the past thirty days to extract translated papers. Here two documents E, H are thought as shared translations if E contains enough translated terms from H and vice versa. Morphological analysis tools for Hebrew (Itai and Wintner, 2008) [8] as well as for English (Minnen et. al., 2001) [9] are accustomed to reducing inflected types of terms up to a base type that is typical. Then after tokenization, lemmatization and prevent term reduction are done for each article that will be represented by its bag of words (BOW).

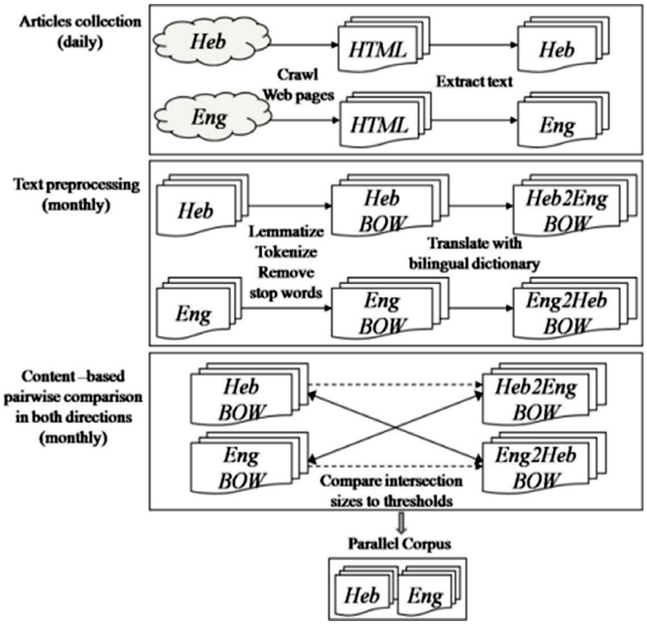
Articles collection (daily)
Heb
Eng
Crawl Web pages
HTML
HTML
Extract text
Heb
Eng
Text preprocessing (monthly)
Heb
Eng
Lemmatize Tokenize Remove stop words
Heb BOW
Eng BOW
Translate with bilingual dictionary
Heb2Eng BOW
Eng2Heb BOW
Content-based pairwise comparison in both directions (monthly)
Heb BOW
Eng BOW
Heb2Eng BOW
Eng2Heb BOW
Compare intersection sizes to thresholds
Parallel Corpus
Heb
Eng

Figure 1: Parallel Corpora Builder (PCB) architecture

- **Evaluation:** The benefit that is primarily of based algorithm is its simpleness without probabilistic models. Yulia Tsvetkov &Shuly Wintneruse the naive BOW contrast and reached positive results.

Table 1
PCB evaluation

Month	English articles	Hebrew articles	Parallel articles	Detected parallel articles	Precision	Recall
07	624	1530	168	145	100%	86.3%
08	548	1486	172	149	100%	86.6%
09	600	1341	165	143	100%	86.7%
average	573	1452	168	145	100%	86.5%

Munteanu and Marcu (2006) [10] delivered a way for extracting fragments which can be parallel sub-sentential a really non-parallel corpus. Each supply language document is translated into target language utilizing a lexicon/dictionary that is bilingual.

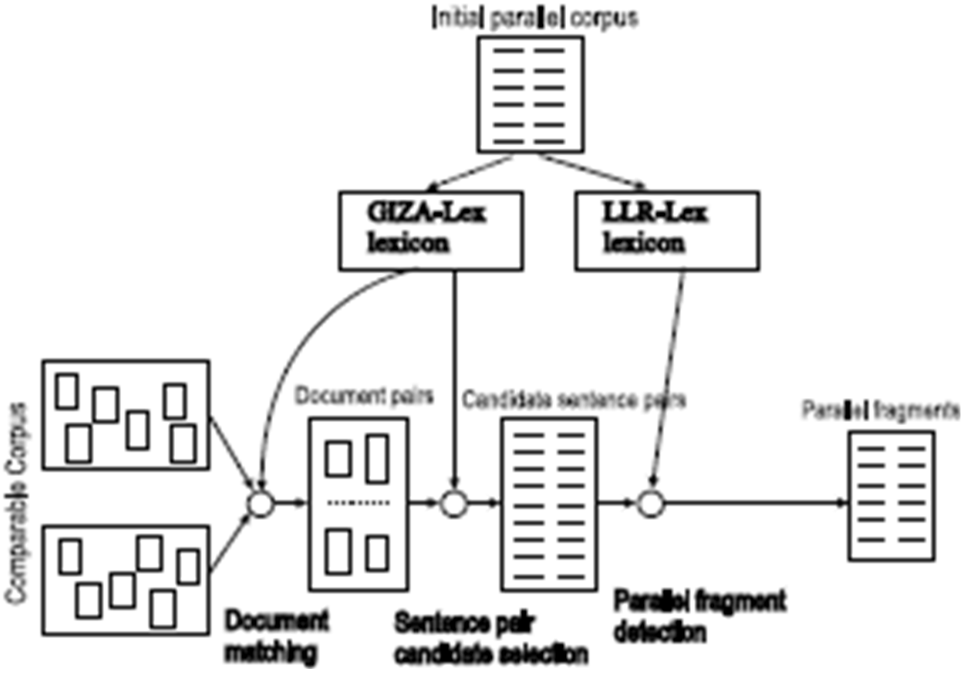

Figure 2: A Parallel Fragment Extraction System

Takehito Utsuro et. al., [11] as proposed Bilingual Text Matching is utilizing Bilingual Dictionary and Statistics he defines a framework that is unified bilingual text matching by combining current hand-written bilingual dictionaries and analytical strategies. The entire process of bilingual text matching contains two major actions: phrase positioning and matching that is structural of sentences. Statistical techniques are used to calculate term correspondences maybe not a part of bilingual dictionaries. Estimated word correspondences are helpful for enhancing both phrase positioning and matching that is structural.

5. POS ANNOTATION METHOD

There is not any corpus available with POS Tagset so Posh annotations are done manually for every single regarding the languages. Narayan Choudhary and Girish Nath Jha [12] have proposed the Creating Multilingual Parallel Corpora in Indian Languages based on areas of the message Annotation framework. He's got struggled

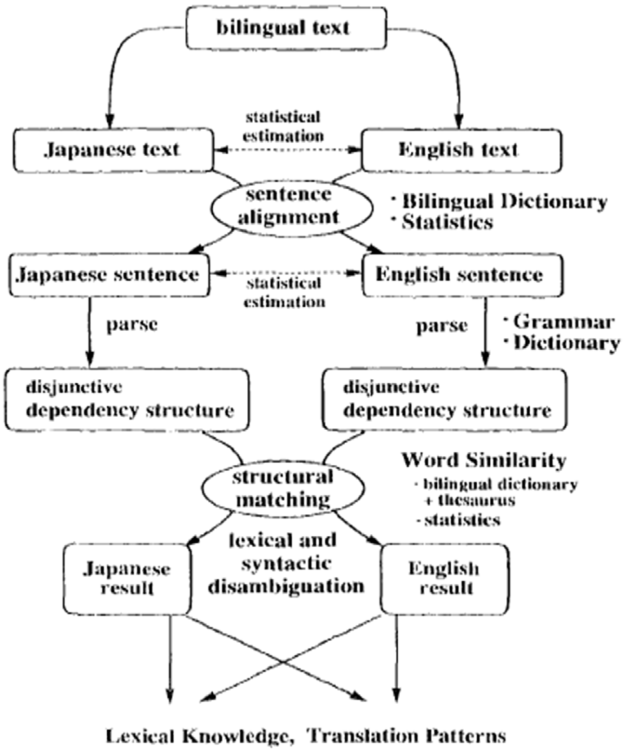

Figure 3: The framework of Bilingual text matching

to obtain 12 major Indian Languages with Hindi once the supply Language and two specified domains as health insurance and Tourism. Right here the two domain names are split into sub-domains after which seemed for the foundational text in those sub-domains which are specific be contained in the supplied text. The POs annotation framework employed for synchronous creation that is corpora.

Wellness domain: it had been split into an overall total of 16 sub-domains. These sub-domains had been made primarily to recapture the various procedures in the area that are medical.

Tourism Domain: it absolutely was divided into a complete of 17 sub-domains which can be major. They certainly were further divided into two categories according to requirement.

6. POS ANNOTATION

The annotations are done manually for every single of this language. There are a few POS taggers developed with various accuracies for the languages like Hindi (Shrivastava, M. & Bhattacharya, P., 2008)[13], Telugu (Avinesh, PVS & G. Karthick, 2007)[14], Bengali (Dandpat, S. et. al., 2007) [15] etc.. POS Tagset: two major kinds of targets employed for POS annotation of texts in Indian Languages. Those two add a tagset produced by IIIT Hyderabad (Bharti, A. et. al., 2006) [16] and Penn tagset proposed by Santorini, B. 1990 [17] and leadership of Microsoft analysis Asia (MSRI) called IL-POST (Baskaran, S. et. al., 2008)[18].

Creating a Parallel Treebank for the language that is dissimilar particularly Swedish-Turkish has proposed by Beta megyesi et. al., [19]. The Treebank has been just a syntax that is balanced corpus containing both fiction and technical papers. It consist of applications 1, 60,000 tokens in S and 14500 in T. The texts are annotated, making use of levels, which can be various POS tags and Morphological features to depend annotation.

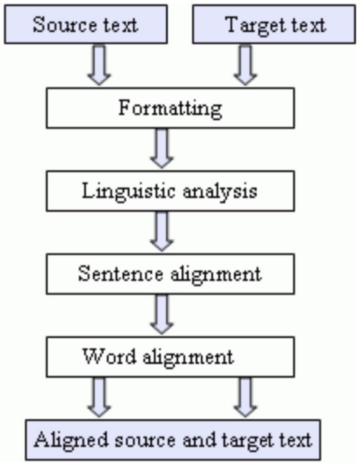

Figure 4: Corpus annotation procedure

To enable a synchronous corpus to be always a parallel Treebank, each language in the corpus has become annotated in the degree that is syntactic. In Treebank, a few annotation levels are utilized for the morpho-syntactic analysis. It is carried out by First annotate the information morphologically using externaltaggers. The Swedish texts are annotated utilizing the Trigrams'n' Tags tagger, trained in Swedish with a precision that is typical of a per cent. The Turkish product is morphologically analyzed and disambiguated using an analyzer that is turkish adisambiguator which immediately learns morphological disambiguation guidelines from the choice list induction algorithmachieving an accuracy of around 96% (Yuretand Ture, 2006) [20]. Both the Swedish and also the Turkish information had been annotated syntacticallyusing MaltParser (Nivre et. al., 2006a) [21], trained on the Swedish Treebank Talbanken05 (Nivre et. al., 2006b) [22] and in the Metu-SabancıTurkish Treebank (Oflazer et. al.,2003) [23], correspondingly. MaltParser had been the doing parser that is most beneficial for both Swedish and Turkish.

7. CROSS LANGUAGE INFORMATION RETRIEVAL FRAMEWORK METHOD

Cross Language Suggestions Retrieval is really a system that has retrieval managed to document in one single language to recover papers in other languages. This method gift suggestions a quick and accurate parallel phrase mining algorithm for comparable corpora called LEXACC (LUCENE-based Parallel Sentence Extraction from Comparable Corpora) considering their cross-language information Retrieval Framework, along with a trainable interpretation similarity measure that detects pairs of Parallel and sentences [24] that is quasi-parallel.

The writer has utilized the Indexing, looking and Filtering concepts for learning the corpus that is parallel additionally utilized the translation similarity function.

Indexing target Sentences: It includes splitting the mark corpus into sentences and transform them such that it keep just stemmed terms being non-functional. It also computes the distance that is normal terms (μ) and also the standard deviation (σ) for target sentences. A phrase be viewed because of it *s* become quicker; if length (s) $<= \mu + \sigma$ and

Very long: If length (s) $>= \mu - \sigma$. Look at the medium sized sentences as $\mu - \sigma <= s$ which is the length ($<= \mu + \sigma$ to be both quick and long. Browsing: It includes finding, interpretation applicants for supply sentences offered an input supply phrase, the part of the major search engines would be to get back a listing of interpretation applicants and includes *h* while the quantity of hits hands because the size associated with the search room that is brand new.

Then it provides the bigger synchronous sentences if h*s could be the bigger it's the greater.

A Lucene question is generate the following:it includes a GIZA++ dictionary by after Och and Ney,2000[25]. For each word that is content keep carefully the most useful 50 interpretation equivalents that are additionally content terms having interpretation probabilities above 0.1.each of those is stemmed and included being an disjunctive question term as MUST APPEAR, add two disjunctive question terms as MUST happen standing for the size of the foundation phrase like quick and long and put in a compulsory question term like SHOULD happen for indicating the prospective document where in actuality the supply phrase interpretation should really be searched. Following the question is built utilize it to interrogate the Lucene internet search engine to get the greatest hits.

Filtering: It is made to further lessen the search that is brand new finding just the most readily useful prospects for the last phase where the interpretation similarity measure is used. Right here we calculate a viability rating for every single prospect phrase set, then maintaining just those typical for the prospect set created by way of a supply phrase and a target phrase t.

$$\text{Viability score} = \alpha \times \beta \times \text{see} \times \text{SIM}$$

Where she represents the rating came back by the search SIM and the motor is just a similar rating

$$\alpha = 1 - \text{abs}(|s| - |t|)/\max(|s|, |t|)\ \beta = \min(|s|, |t|)/\lambda$$

Where abuse could be the absolute value, |s| may be the size in terms of phrases and λ is definitely an integer constant representing the distance limit from which it think about a phrase become lengthy ($\lambda = 100$). The similarity rating is $\text{SIM} = 2 \times \text{tefound} \times \text{te}/|s| + |t| \times 1/\sqrt{}\text{coh}$ tefound could be the quantity that is total of in us which is why it discovered interpretation equivalents in it, a coach could be the cohesion rating computed since the average distance between your sorted roles of the interpretation equivalents present in it.

The Translation Similarity Measure: It modeled as an amount that is weighted of functions that suggest in the event that supply bit of text is translated by the goal.

Offered two sentences *s* within the supply *t* and language within the target language then

The interpretation similarity measure $P\ (s, t)$ is $P\ (s, t) = t$ that is $\Sigma\ (s)$

Each function function *fi*(*s*, *t*) will get back an actual value between 0 for *s* and *t* aren't associated at all and 1 for it is really an interpretation of sand plays a role in the entire parallelism rating by having a certain small fraction if that is language-pair. If that is reliant (*s*, *t*) are calculated on such basis as content terms interpretation power, practical terms interpretation, strength, positioning obliqueness, strong interpretation sentinels and end with all the punctuation that is exact same.

8. BOOTSTRAPPING TECHNIQUE

This technique relates to the beginning of the procedure that is self-sustaining is meant to continue without outside input. Do Thi Ngoc Diep et. al., [26] presents an unsupervised way for extracting parallel phrase pairs from the corpus that can be compared. An interpretation system can be used to mine the comparable corpus and to detect parallel phrase pairs as well as a procedure that is iterative implemented to increase the amount of extracting parallel sentence pairs and to increase the general quality associated with interpretation system.

Sarikaya et. al., (2009) [27] Introduced an iterative bootstrapping approach where the extracted phrase pairs are included with the first synchronous corpus to rebuild the equipment interpretation system that is analytical. The technique contains 3 steps: (i) document combining. (ii) Sentence sets the positioning of this paired papers and (iii) context extrapolation to enhance the phrase set protection. People's assessment of the extracted data demonstrates that 95% associated with the sentences which can be removed helpful information for interpretation.

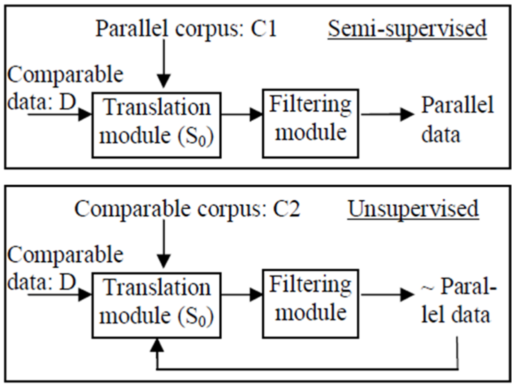

Figure 5: Semi-supervised V/S Unsupervised methods

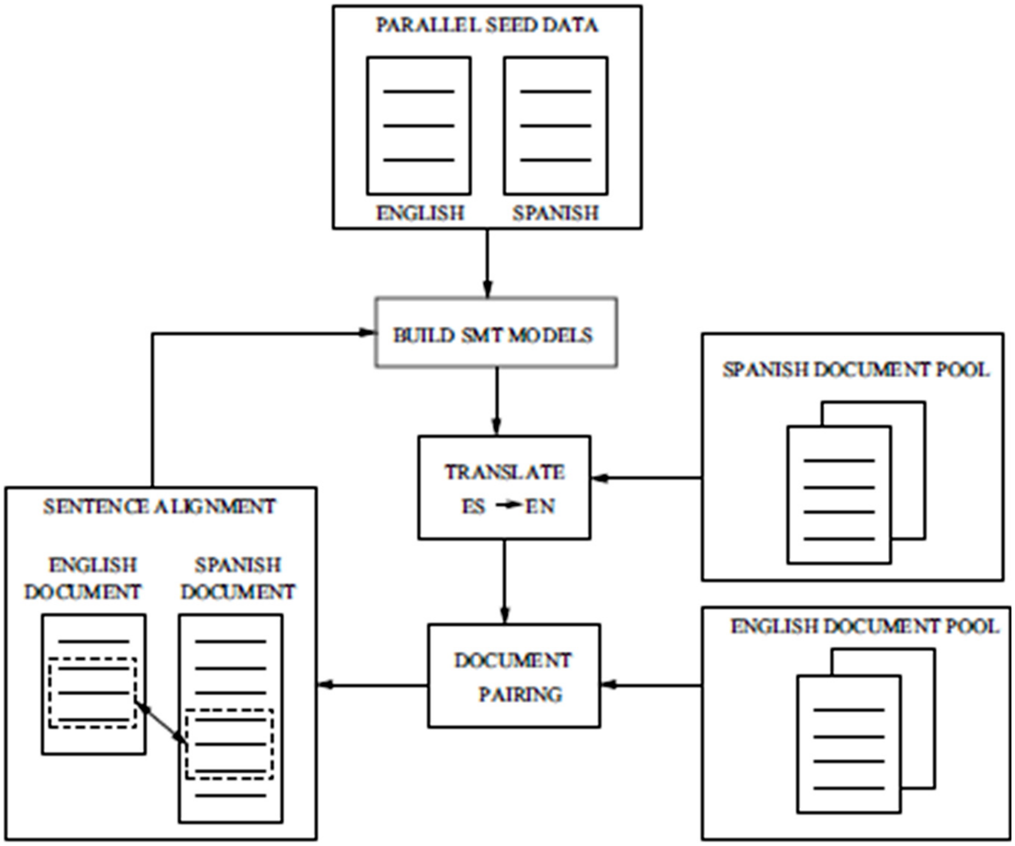

Figure 6: Flowchart for Iterative sentence extraction method

Tools: PEXACC is a phrase that is"Parallel from Comparable Corpora" which is a language approach that is a separate parallel expression mining from comparable corpora and here is a trainable and extensible unit along with other languages. [28] XCES generator that is parallel corpora the languages like English and Romanian from Romania.

Table 2
Comparison of Parallel corpora under Parallel corpora generation techniques

Reference	*Data used*		*Techniques*				*Accuracy*	
	Indian Language	*Foreign Language*	*Dictionary*	*POs Tagged*	*CLIR*	*Bootstrapping*	*Precision*	*Recall*
[7]		✓					100%	86.5%
[10]		✓	✓					
[11]		✓	✓					
[12]	✓			✓				
[19]		✓		✓				
[24]		✓			✓			
[26]		✓				✓		
[27]		✓				✓		

Precision = Number of correctly aligned Sentences/Number of aligned Sentences
Recall = Total number of correct aligned Sentences/Total number of Sentences in source
Accuracy = Number of aligned Sentences/Total number of Sentence

9. CONCLUSION

Synchronous corpora are essential resources for the language that is normal applications. In this paper, we've described the various practices employed for building corpus that is parallel. Therefore the substantial research gaps between these practices that are written by various writers which is seen that it's hard to get 100% accuracy, Recall and Frequency in virtually any Languages.

REFERENCES

[1] P. Resnik and N. A Smith. 2003. The web as a parallel corpus. Computational Linguistics, 29(3):349–380.

[2] D. S Munteanu and D. Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics, 31(4):477–504.

[3] Development of Hindi-Punjabi Parallel Corpus Using Existing Hindi-Punjabi Machine Translation System and Using Sentence Alignments Pardeep Kumar etl.2010.

[4] B. Zhao and S. Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In Proceedings of the 2002 IEEE International Conference on Data Mining, page 745. IEEE Computer Society

[5] Fung, P., P Cheung. 2004. Mining very-non-parallel corpora: parallel sentence and lexicon extraction via bootstrapping and EM. Conference on Empirical Methods on Natural Language Processing.

[6] Kumano,T.,H.Tanaka, T. Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability SMT model. Conference on Theoretical and MethodologicalIssues in Machine Translation.

[7] Yulia Tsvetkov & Shuly Wintner. Automatic Acquisition of Parallel Corpora from Websites with Dynamic Content.

[8] Alon Itai and ShulyWintner. 2008. Language resources for Hebrew. Language Resources and Evaluation, 42:75– 98, March. Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. Natural Language Engineering, 7(3):207–223.

[9] Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. Natural Language Engineering, 7(3):207–223.ling

[10] Dragos Stefan Munteanu and Daniel Marcu(2006). Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora

[11] Takehito Utsuro etl.Bilingual Text Matching using Bilingual Dictionary and Statistics

[12] Narayan Choudhary & Girish Nath Jha (2011).Creating Multilingual Parallel Corpora in Indian Languages.

[13] Shrivastava, Manish and Pushpak Bhattacharyya. (2008) Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information without Extensive Linguistic Knowledge. In: Proceedings of theInternational Conference on NLP (ICON08), Pune, India,

[14] Avinesh, P. V. S. & G. Karthik. (2007). Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation-Based Learning. In: Proceedings of the IJCAI and the Workshop OnShallow Parsing for South Asian Languages (SPSAL) (2007), pp. 21-24.

[15] Dandapat, Sandipan, Sudeshna Sarkar, Anupam Basu. (2007) Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario. In: Proceedings of theAssociation for Computational Linguistic, pp 221-224

[16] Bharati, Akshar, Dipti Misra Sharma, Lakshmi Bai, Rajeev Sangal. (2006). Anncorra: Annotating Corpora. LTRC, IIIT, Hyderabad

[17] Santorini, Beatrice. (1990). Part-of-speech Tagging Guidelines for the Penn Treebank Project. Technicalreport MS-CIS-90-47, Department of Computer andInformation Science, University of Pennsylvania.

[18] Baskaran, Sankaran, Kalika Bali, Monojit Choudhury, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha and K.V. Subbarao. (2008). A Common Parts-of-Speech Tag Set Framework for Indian Languages. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Daniel Tapias (Eds.) Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

[19] Be´ata Megyesi et. al., Swedish-Turkish Parallel Treebank.

[20] Deniz Yuret and Ferhan T¨ure. 2006. Learning morphological disambiguation rules for turkish. In Proceedings ofHLT NAACL

[21] Joakim Nivre, Johan Hall, and Jens Nilsson. 2006a. Maltparser:A data-driven parser-generator for dependency parsing. In Proceedings of the 5th International Conferenceon Language Resources and Evaluation (LREC2006).

[22] Joakim Nivre, Johan Hall, and Jens Nilsson. 2006b. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In Proceedings of the 5th InternationalConference on Language Resources and Evaluation(LREC 2006).

[23] Kemal Oflazer. 1994. Two-level description of Turkish morphology. Literary and Linguistic Computing, 9:2.

[24] Dan □tefãnescu et. al., (2012).Hybrid Parallel Sentence mining from Comparable Corpora.

[25] Och, Franz Josef and Hermann Ney. 2000. Improved Statistical Alignment Models.

[26] Do Thi Ngoc Diep, Laurent Besacier and Eric Castelli (2010).A Fully Unsupervised Approach for Mining Parallel Data from Comparable Corpora

[27] Sarikaya R. et. al., 2009. Iterative sentence–pair extraction from quasi–parallel corpora for machine translation.

[28] Radu Ion. PEXACC: A Parallel Sentence Mining Algorithm from Comparable Corpora