Journal of Computer Science

Original Research Paper

A Systematic Literature Review on Extraction of Parallel Corpora from Comparable Corpora

[1]Dilshad Kaur and [2]Satwinder Singh

[1]*Department of Computer Science and Technology, Central University of Punjab, Bathinda, Punjab, India*
[2]*Associate Professor, Computer Science and Technology, Central University of Punjab, Bathinda, Punjab, India*

Article history
Received: 22-05-2021
Revised: 31-08-2021
Accepted: 05-09-2021

Corresponding Author:
Dilshad Kaur
Department of Computer Science and Technology,
Central University of Punjab,
Bathinda, Punjab, India
Email: dilshadkaur@outlook.com

Abstract: In today's Globalized Scenario, the requirement for translation is high and increasing rapidly in the number of fields, but it is difficult to translate everything manually. Machine Translation, which is dependent on corpora availability, is a medium for meeting this high demand for translation. Parallel corpora are used to gain most translation knowledge. But, the number and quality of parallel corpora are critical. Because parallel corpora are not readily accessible for many different language pairs, comparable corpora that are widely accessible can be used to extract parallel corpora. A systematic literature survey is performed on 188 research articles that are published in premier journals, conferences, workshops and book chapters. The research process is carried out while considering the research questions. Different MT systems along with their features are identified. Several datasets and techniques for bilingual lexicon extraction, parallel sentence and fragment extraction are revealed. A proposed architecture and a mind map are also showcased in this review article to provide better clarity regarding parallel data extraction using comparable corpora. The study of the paper will increase readers' understanding of parallel data mining through bilingual lexicons, parallel sentences and fragments.

Keywords: Machine Translation, Statistical Machine Translation, Parallel Corpora, Comparable Corpora

Introduction

In today's era of globalization, a lot of data is accessible on the internet in diverse languages and domains. Due to diversity in languages all over the world, it is impossible to learn every language. People who are accessing the internet come from different language backgrounds. To overcome this problem the content available on the internet requires translations. In any case, it is difficult to perform the job of translation manually. There is the requirement of the machine to do the translation. With this requirement, the existence of Machine Translation (MT) emerged. MT (Machine Translation) is a medium to achieve high demand of translation. It is among the applications that fall under the umbrella of Natural Language Processing (NLP). It is an incredible tool to increase competence and decrease the cost of translation. For the process of translation, there is a need for some kind of dataset or corpus to train the machine. Thus, translation highly depends upon the availability of corpus. Corpus is an enormous assortment of text used to analyze how the words, phrases and language are used. It is used by linguists, social scientists, natural language processing experts, etc. It tends to be arranged into two prime classes, namely Parallel Corpora (PC) and Comparable Corpora (CC). The parallel corpus comprises two different language corpora where one is the translation of another. Parallel corpus is sentence-aligned bilingual texts. Whereas a comparable corpus is a set of two or more different language corpora which is not the exact translation of each other and hence are not aligned. There are two fundamental ways to make a corpus specifically, rule-based and statistical analysis (Babych *et al*., 2012). In the early days of MT research, rule-based played a keen role. In this, all the conversion rules are composed manually and afterward, the encoding is done into the MT framework. However, languages are very vast and complex, it is quite impossible to write the rules manually in a relatively short period. To address this issue, the emphasis was shifted to statistical analysis. In due course of recent decade or two, MT research has started working in the branch of Statistical Machine Translation (SMT) (Koehn, 2009; Och and Ney, 2003; Brown *et al*., 1993) and it has risen as a key method in the field of both research and business area. In SMT (Koehn, 2009; Och and Ney, 2003; Brown *et al*., 1993), translated information is consequently procured from PC (parallel

Science Publications

corpora) which is a kind of sentence-aligned bilingual text. Therefore, huge growth is seen in the MT framework for various language sets. Nowadays, most machine translation research is conducted with this approach. In SMT, due to high reliance on Parallel Corpora (PC), the quality and quantity of PC are serious (Ali *et al*., 2010; Srivastava and Bhat, 2013; Post *et al*., 2012). Nonetheless, aside from a couple of language sets and in some specialized fields, a top-notch PC of adequate size stays a scant asset. The insufficiency of Parallel Corpora (PC) has become SMT's primary challenge. There is no abundant Parallel Corpora (PC) available for performing the task of translations. Making the use of Comparable Corpora (CC) is a compelling method to solve the problem of insufficiency of PC for SMT (Statistical Machine Translation). The main reason behind using Comparable Corpora (CC) are, first these are undeniably more accessible for different fields than PC, such as Wikipedia, bilingual articles, bilingual websites, patent documents, e-newspapers, social media and research-related academic papers which are easily available. Second, single language corpus is easy to obtain and in using comparable corpora, work is performed on single language corpus only. Third, a lot of parallel data like bilingual lexicon, parallel sentences and fragments can be obtained from comparable data.

Motivation

The motivation for this review is derived from the fact that a detailed insight study is required for studying the mining of Parallel Corpora (PC) in the form of lexicons, sentences and fragments from Comparable Corpora (CC). When the search was conducted in the relevant literature, it did not reveal a clear review regarding PC and CC. A collaborated work that includes the extraction processes of lexicons, fragments and sentences for Parallel Corpora (PC) was not available in a systematic format. Also, no relevant review was available that could focus on statistical machine translations for different languages and domains. There was a need to study this area and give an accredited overview.

This study goal is to review and contribute towards:

- A feasible study that focuses on strengths and weaknesses of the research in the concerned domain,
- A systematic review is required in the branch of parallel data mining from comparable data. Therefore, this study explores the on-hand research on different data extraction techniques.
- A combined effort for extraction of lexicons, sentences and fragments from comparable data.
- Also, this survey will give us insight into the number of languages and domains which use machine translation.

Background of Related Work

Different authors have studied the PC and CC and their usage among Indian languages as well as European languages. A few reviews in this field are given by some researchers like Maskara and Bhattacharyya (2018), Khosla and Acharya (2018), Iyer (2015), Kulkarni (2013), Lehal *et al*. (2018), Saini and Sahula (2015) and Padhya and Sheth (2019).

Maskara and Bhattacharyya (2018) focused on the recent developments in the field of parallel sentence mining from CC using techniques like word embedding, deep learning and machine translation systems. Different classifiers were used by different authors but most of the work was done with maximum entropy-based classifier and SVM classifier (Chang and Lin, 2011; Zhu *et al*., 2011; Bouamor and Sajjad, 2018). Some research projects have made use of Solr (Zhang and Zweigenbaum, 2017) and Lucene (Azpeitia *et al*., 2017) search engines which are information retrieval-based frameworks.

Khosla and Acharya (2018) depicted in their literature survey, the existing methods by which parallel corpus can be built. The survey focused on corpus built aligned at the document level and sentence level. This study discussed three approaches to create parallel corpus i.e., Sentence Alignment approach, Web Mining approach and Manual approach.

Iyer (2015) performed the literature survey on comparable corpora. Different existing methods to extract PC from CC at sentence, phrase and word level were reviewed. The survey report was categorized in the form of chapters. The introductory part included approaches to machine translation. It was followed by different techniques used for mining the parallel sentence from comparable data. The report also gave a glimpse of different approaches to extract the phrases from comparable data. Lastly, the report focused on the extraction of bilingual lexicons, an application of CC.

Lehal *et al*. (2018) offered a review of different processes and approaches for bilingual lexicon extraction. It included Correlation-based extraction, Vector Depiction, Projection-Based approach, Classifiers-based mining, PC based approach, Linguistic Knowledge-based extraction for mining of bilingual lexicons for Comparable Corpora (CC). The review also mentioned the limitations of different extraction techniques like level of complexity, parameters, corpus size, accuracy, etc. A suggestion was also made to combine either two or three methods to improve efficiency and overcome the limitations.

Kulkarni (2013) revealed the investigation of literature by exploring different approaches for parallel sentence mining, parallel phrase extraction and bilingual lexicons abstraction from the Comparable Corpora (CC).

Padhya and Sheth (2019) conducted a review of the literature on numerous Machine Translation (MT) systems for Indian languages. The survey presented us with the findings that "Statistical Machine Translation" and Example-Based MT are the best approaches when working with a large corpus. Rule-Based Machine Translation is useful when there is no corpus. And for the same ordered language, Direct Machine Translation is best suited. The survey report concluded that all Indian languages have future enhancement scope. Similarly, Saini and Sahula (2015) also depicted the current situation of machine translation research in India. The survey also provides the difference amongst the methodologies used.

Literature surveys, according to the aforementioned researchers, include the latest updates on previously completed work in the concerned domain. It's also a simple way to look through the literature on a particular subject. The systematic review in this study has followed the footsteps of Singh and Kaur (2018). According to their study, a systematic literature survey traces the available and relevant literature by framing research questions. Literature is collected by following the criteria of inclusion and exclusion, keeping in mind the main topic. But, before we get into the specifics of the job, it is important to define "Machine Translation", "Statistical Machine Translation", Comparable Corpora (CC), Parallel Corpora (PC) and bilingual lexicons. The study will further review the steps to extract parallel data from comparable data.

Corpus

Corpus is a very large collection of text used to analyze how the words, phrases and languages are used. Its plural is corpora. Linguists, social scientists and specialists in natural language analysis, etc. use it. Corpus is used in different domains and has been of keen importance. It is used in Discourse analysis, literary studies, translation work, forensic linguistics, Pragmatics, political discourse and social discourse (O'Keeffe and McCarthy, 2010). There are different kinds of corpora that are used for various purposes. Tognini Bonelli and Sinclair (2006) presented in their study about the topology of corpora i.e., Sample corpora, CC, Special corpora, Corpora with the time dimension, Bilingual and Multilingual corpora, PC, Spoken corpora, Non-native Speaker corpora and Normative corpora. But in this particular literature survey, the focus will only be on Comparable and PC which are to be used for translation work.

Comparable Corpora

CC are a group of transcripts that are closely related to one another but are different in some of the other aspects (Kenning, 2010). The texts in Comparable Corpora (CC) are linked together based on criteria like a set of topics, the text of a certain size, time of the text, etc. CC are the set of two or more different language corpora that are not the exact translation of each other and are not aligned. Some sources of comparable corpora are Wikipedia (Adafre and Rijke, 2006), Bilingual articles from newspapers and web, etc. (Tillmann, 2009; Zhao and Vogel, 2002)

Parallel Corpora

Parallel Corpora (PC) are collections of transcripts in two or more languages that are precise translations of each other. In this, the relationship amongst the text in two languages or more language pairs lies in shared meaning (Kenning, 2010). Parallel corpora are not available easily because of the scarcity. The parallel texts such as lexicons, fragments and sentences need to be mined from the comparable corpora. Section "Parallel Sentences and Fragments Extraction" in this literature paper gives more clarity to the extraction process of fragments and sentences. Examples of already created PC are the English-Norwegian Parallel corpus, WHO bilingual articles which are in English and Spanish etc.

Machine Translation

"Machine Translation" is a device that is used to create translations from one normal language into other, with/without human intervention (Hutchins and Somers, 1992). Nirenburg and Wilks (2000) gave an overview regarding Machine Translation along with the crucial issues and highlights of the latest applications under machine translation. The survey report explained the different areas where there is the use of MT like linguistics, neuroscience, artificial intelligence, software designing, philosophy, etc. It provides an opportunity for software engineers to experiment and construct non-numerical complex systems. It is also used by field computational linguists for encoding the syntax and semantics of different languages into computer understandable form. Computational linguistics has a subfield called machine translation. Isabelle and Foster (2006) gave an overview of Machine Translation (MT). It defines MT as a process that translates two human languages: The source and target languages. This study stated machine translation as the study of different ways and methods that make the machine produce translations. It revealed that there is always a requirement of understanding the source language, grammar of target language and relevant knowledge to fulfill the informational gap between the target language and source language. This study gave insight into segmenting texts into words, word forms, word co-occurrence, dealing with unknown words, finding idioms in the sentences, solving the problem of ambiguity in the source language, word insertion and deletion, order of the words, etc. Machine translation has two levels: Metaphrase and Paraphrase (Tripathi and Sarkhel, 2010). Metaphrase refers to the word-to-word translation but the text converted may not convey the same meaning. There can be a difference in the semantics from the original text. Paraphrase is not the word-to-word translation but provides the user with exact meaning as of original text. Two models play a role in MT which are Rule-based MT and SMT (Statistical machine translation) (Babych *et al*., 2012). Rule-Based machine translation is based on creating a group of rules manually using linguistic information whereas Statistical Machine Translation (SMT) (Koehn, 2009; Och and Ney, 2003; Brown *et al*., 1993) is a machine translation method in which translations are produced on the base of statistical models, the parameters of which are extracted from a bilingual text corpora analysis. MT systems can be classified into several ways based on the specific approach by which the translation is carried out (Chand,2016). Figure 1 depicts the classification of MT approaches. There are many methods under MT like Rule-based, Corpus-based, Hybrid and Neural based.

Table 1: Indian Machine Translation systems based on approaches and features

MT approach	MT system	Target language	Features	Citation
Rule-Based MT	English- Sanskrit MT	English- Sanskrit	Uses Artificial Neural network (ANN)	Mishra and Mishra (2010)
	English-Urdu MT	English-Urdu	Uses ANN with RBMT	Khan and Mishra (2011)
	Etrans	English- Sanskrit	Uses Synchronous Context -Free grammer	Bahadur *et al*. (2012)
	TranSish	Sanskrit-English	Uses Artificial intelligence with the dictionary	Upadhyay *et al*. (2014)
	Transmuter	English-Marathi	Uses word sense disambiguation and Stanford parser	Gajre *et al*. (2014)
	English-Marathi MT	English-Marathi	Offers sentiment analysis, spell testing and idiom translation	Pisharoty *et al*. (2012)
	English- Kannada MT	English- Kannada	Uses Morphological Generator	Basavaraddi and Shashirekha (2014)
Direct MT	ETSTS	English-Sanskrit	Uses Morphological markings	Rathod and Sondur (2012)
	Punjabi-Hindi MT	Punjabi-Hindi	Uses word-by-word translation	Josan and Lehal (2008)
	Hindi-Punjabi MT	Hindi-Punjabi	Uses Unicode characters	Goyal and Lehal (2011)
	English- Devanagari	English, Marathi, Hindi, Gujarati	Uses the concept of transliterations and human aided concept	Dhore and Dixit (2011)
	Anusaaraka	Bengali, Kannada, Marathi, Punjabi, Telegu, Hindi	Uses paninian grammar	Bharati *et al*. (1997)
Transfer Based MT	Punjabi-English MT	Punjabi-English	Uses the concept of transfer approach and Morph analyzer	Batra and Lehal (2010)
	Telugu-Tamil MT	Telugu-Tamil	Used divergence index	Krishnamurthy (2015)
	Matra	English-Hindi	The rule bases and heuristics are used in the human-aided translation project.	Ananthakrishnan *et al* (2006)
	Shakti	English, Hindi, Marathi, Telugu	Uses rule basis	Sangal (2004)
	Mantra	English, Gujarati, Telugu, Bengali, Marathi, Hindi	Uses tree adjoining grammer	Darbari (1999)
	Sampark	Hindi, Telugu, Tamil, Urdu, Punjabi,	Uses paninian grammer	Ahmad *et al* (2011)
	Bengali-Hindi MT	Bengali, Hindi	Uses lattice-based data	Chatterji *et al*. (2011)
Interlingual MT	English- Sanskrit MT	English, Sanskrit	Uses semantic mapper and lexical parser	Barkade *et al*. (2010)
	Anglabharati	English, Tamil, Hindi	Uses intermediate structure pseudo lingual	Sinha *et al*. (1995)
	English-Bengali MT	English, Bengali	Uses Context-free grammar	Ashrafi *et al*. (2013)
Statistical MT	English-Kannada /Telugu MT	English, Kannada, Telugu	Uses transliteration model	Reddy and Hanumanthappa (2011)
	English-Urdu MT	English, Urdu	Uses Moses and language model toolkit, IRSTLM	Ali *et al*. (2013)
	English-Urdu MT	English. Urdu	Uses Moses and Giza++	Ali *et al*. (2010)
	English- Sanskrit MT	English, Sanskrit	Uses Statistical machine decoder	Warhade *et al*. (2012)
	English-Malayalam MT	English, Malayalam	Uses hand-made rules and statistical decoder	Sebastian *et al*. (2010)
	Kriya	Hindi-English	Uses hierarchical phrases	Sankaran *et al*. (2012)
Example-Based MT	Malayalam-English MT	Malayalam, English	Uses MATLAB functions	Anju and Manoj (2014)
	English-Hindi MT	English, Hindi	Uses similarity, tagging and training matrix	Sinhal and Gupta (2014)
	Vaasaanu baada	Bengali, Assamese	Uses pseudo-code and Backtracking	Vijayanand *et al*. (2002)
	Anubharti	Hindi, English	Uses pattern based and example -based approach	Jain *et al*. (2001)
Hybrid MT	English-Sanskrit MT	English, Sanskrit	Combined Rule-based and Example-based	Rathod (2014)
	Urdu- English MT	Urdu, English	Combines Rule-based, Example-based & SMT	Malik and Habib (2013)
	Angla Hindi	English, Hindi	Combines Rule-based, Example-based & SMT	Sinha and Jain (2003)
	Anubaad	English-Bangla	Combines Transfer Based and Example-based	Bandyopadhyay (2000)

Table 2: European and Asian MT systems along with approaches and features

MT approach	MT System	Target language	Features	Citation
Rule-Based MT	Japanese-English MT system	Japanese-English	Uses structural matching in parse trees	Winiwarter (2007)
Interlingual MT	ICENT	Chinese-English	Uses Syntactic parsing and semantic analyzing	Qi *et al*. (2002)
	English-Korean MT	English-Korean	Included word-sense disambiguation and had a plug and play architecture	Lee *et al*. (2002)
	English-Turkish MT	English-Turkish	Uses Knowledge-based system	Hakkani *et al*. (1998)
Statistical MT	English-to-Czech Factored MT	English-Czech	Made the use of morphology	Bojar (2007)
	Hebrew-German MT	Hebrew- German	Uses sense disambiguation	Dagan and Itai (1994)
	Farsi-German SMT	Farsi-German	Uses English as bridge language	Bakhshaei *et al*. (2010)
Example-Based MT	English-Turkish MT System	English-Turkish	Uses Synchronous Structured String Tree Correspondence	Alp and Turhan (2008)
	Chinese-Japanese MT system	Chinese- Japanese	Use of Super Functions	Sun *et al*. (2009)
Hybrid MT	Japanese-to-English MT	Japanese- English	Combined Rule-Based and Statistical method	Terumasa (2007)
	Korean- Chinese MT	Korean-Chinese	Combined Rule-Based and Statistical method	Knag *et al*. (2005)
	English-to-Persian MT	English=Persian	Combined Rule-Based and Statistical method	Motazedi and Shamsfard (2009)

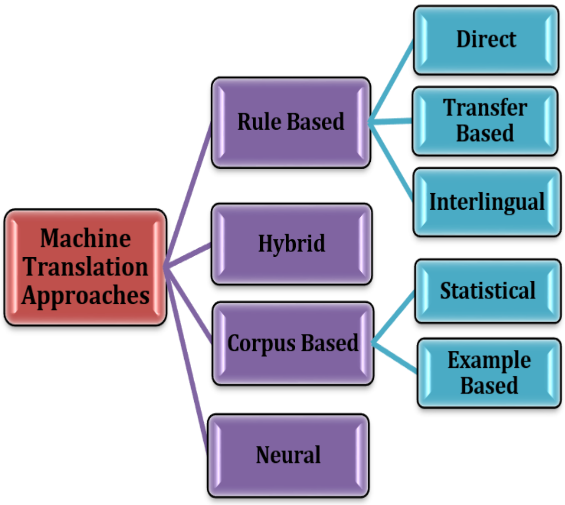

Fig. 1: Classification of MT approaches

Different languages and domains follow various approaches to perform translation.MT not only focuses on Indian languages but has worked tremendously for European and Asian languages as well. Table 1 presents MT approaches with diverse MT systems especially for Indian languages. It is observed from the study that most of the work is done under the Rule-based approach and other approaches that involves human assistance for Indian languages. The work performed under SMT involves transliterations and human-aided tools mostly. Whereas Table 2 refers to some of the MT systems created for European and Asian languages (Other than Indian languages). There must also be the number of other MT systems available such as Neural Machine Translation, but this review has only focused only on those MT systems which are mentioned in the collection of 188 papers of this literature survey.

Statistical Machine Translation

"Statistical Machine Translation" (Koehn, 2009; Och and Ney, 2003; Brown *et al*., 1993) is a method of creating a machine that automatically decides translation rules from a collection of the translated manuscript by integrating the contribution and production of the translation process and getting the results from the data figures (Koehn, 2009). Brown *et al.* (1993) presented a mathematical logic about the working of SMT. It is said that to convert a sentence in a foreign language into a sentence in English, there is a need for logic to make the SMT system work. SMT (Koehn, 2009) has emerged as a key method in both the academic civic and the marketable sector over the last decade or so, with machine translation research taking a turn towards it.

In SMT (Koehn, 2009; Och and Ney, 2003; Brown *et al*., 1993), Parallel Corpora (PC) are used to automatically gain translation knowledge (Veronis, 2000) and there is the rapid creation of MT frameworks for various language pairs and domains. The scale and quality of PC have a tremendous effect on the quality of translation in the "Statistical Machine Translation" system. But there are very few resources available of PC for different language domains. "Statistical Machine translation" has emerged as the main tool for conversion work over the past two decades. It has emerged out to be fruitful for the research society and commercial community (Koehn, 2009). Babhulgaonkar and Bharad (2017) suggested that the problem related to translation can be reduced by restricting to certain domains and languages only.

Research Methods

The systematic approach for reviewing the literature is chosen. It is a process for identifying, evaluating and understanding all the available research in the particular domain. Singh and Kaur (2018) literature review technique was followed to direct the systematic approach in the paper. A systematic approach is chosen to give better insight into the concerned subject. Moreover, while collecting the literature, to the best of our knowledge there was no such systematic review in this particular field. There were undoubtedly a few surveys that are already mentioned in "Background of Related Work" but none of them was done systematically.

Procedure of the Review

Having study questions, collecting data, analysis of data, applying inclusion and exclusion criteria, reviewing and assessing the research results and concluding with the discussions are part of the analysis protocol. Both electronic and manual databases, including journals, conference proceedings and researcher thesis, are searched for the literature review. There is a need for doing a literature review as it deals with collecting all the related pieces of evidence as per the research questions regarding the specified topic. A proper research procedure is required to be followed as it provides more clarity of the topic and also makes our research work more organized.

Research Questions

The research question is the key component for designing a systematic survey. To keep the study focused on the specific goal, research questions are mentioned. Research questions motivate to work towards a particular direction and carry out the survey. The main aim of research questions in this survey is to reveal different MT approaches, datasets used in the extraction process, techniques followed for bilingual extraction and various methods for parallel sentence extraction. Table 3 lists a series of research questions that can be used to perform a systematic literature review in the current study.

Sources of Information

To collect the relevant studies, current work identifies and evaluate the pool of articles. In performing a literature survey, extensive searching is done. Before initializing the review, some proper databases are to be chosen. Then the searching of databases is done by using the keywords. The study also checked databases of academic resources and publishers on a general and iterative basis such as:

a) "ACM Digital Library" (http://dl.acm.org)
b) "ScienceDirect" (https://www.sciencedirect.com)
c) "IEEE eXplore" (https://ieeexplore.ieee.org/)
d) "ACL" (https://www.aclweb.org/)
e) "Springer" (https://www.springer.com/)

For the review of the work, the study included international journals, review articles, the thesis of researchers, book chapters and conference proceedings that we have mentioned under "Other" academic resources. It contains all the research works that are indexed in "Google Scholar" and "Citeseer". Some papers presented in MT "Summit" are also included. Papers published in Journals like IJET, IGI Global, IJCA, Tand Fonline, CFILT, etc. are mentioned under the category of "Other". Figure 2 elaborates the percentage of papers included in this survey from the above-mentioned data sources. A total of 188 papers have been added from various databases. The pie chart in Fig. 2. presents a percentage of papers added from different data sources in this survey.

Vital Keywords

Keywords play a vital role in the process of the systematic review. During the implementation of research process, a set of keywords were defined. These keywords were used in searching the databases for the relevant papers. Every database mentioned in "Sources of Information" was searched for the given keywords. After obtaining the papers based on keywords, the title of the paper was read. Inclusion and exclusion of papers were then done according to the title. If the title seemed satisfactory then the abstract reading was done. Keywords made the search easy and relevant to the field. Figure 3 depicts the percentage of papers included in the survey on a particular keyword. Following are the keywords that have been used for each data source:

- Machine Translation,
- Statistical Machine Translation,
- Text Alignment,
- Comparable Corpora,
- Parallel corpora,
- Bilingual Lexicon Extraction,
- Parallel Fragment Extraction,
- Parallel Sentence Extraction.

Inclusion and Exclusion Criteria

Lots of literature is available with the above keywords. While exploring different databases, similar papers were seen in multiple repositories. Therefore, to ensure that the search is easily manageable, review established certain conditions for inclusion and exclusion in the selection of articles as follow.

Table 3: Research questions for systematic review

RQ1	What is the current status of data extraction in respect of parallel and comparable data?
RQ2	What are the various datasets used in the process of extraction?
RQ3	What type of machine translation approaches are used for translation in different language domains?
RQ4	What kind of parallel data can be mined by using CC ("comparable corpora")?
RQ5	What are the different kinds of parallel sentence and fragment extraction techniques followed?
RQ6	What are the different ways to extract Bilingual Lexicons from comparable data?

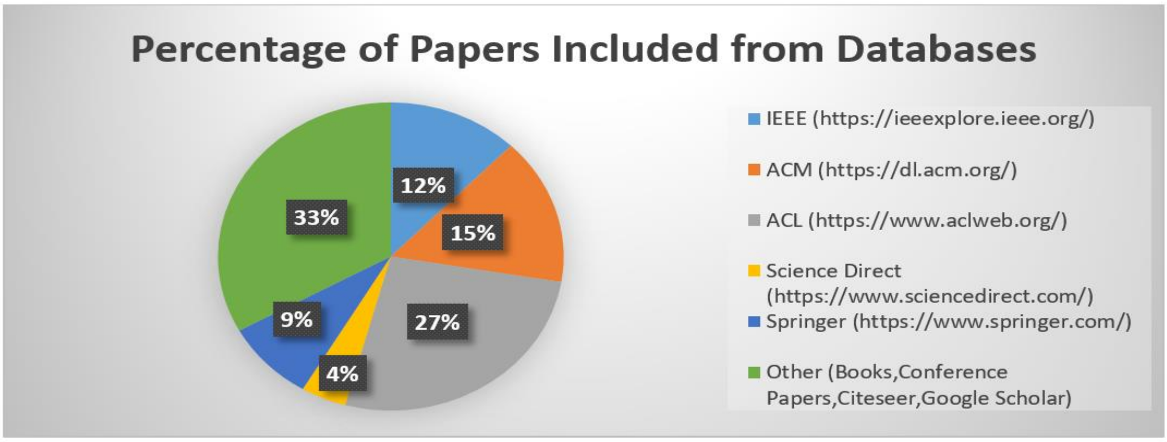

Fig. 2: Papers included from different databases

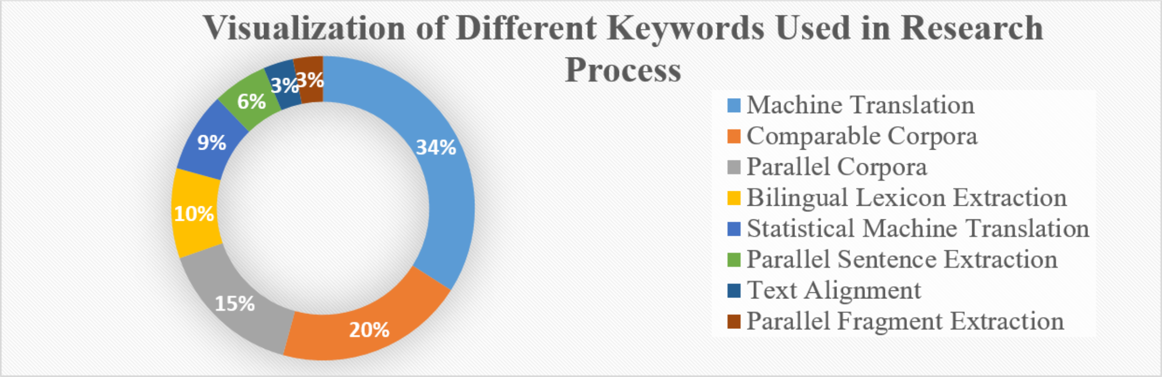

Fig. 3: Keywords used in conducting the research process

Inclusion Criteria:

- Criteria the study follows to include the articles in the survey are:
- Articles relating only to computer science and engineering have been included because the term "corpora" are multi-disciplinary and is found in different branches.
- The papers written in English were included.
- Conference papers were also included.
- Book chapters were included.
- Papers indexed in Google Scholar were included with relevancy with the keywords.

Exclusion Criteria:

- To exclude the unwanted articles for review, the following criteria are followed
- All other articles on different subjects like medical, animal sciences, biomechanics, etc. were excluded.
- Informal studies like unknown conferences or journals were discarded
- Papers irrelevant to the research questions were also excluded
- Wikipedia writings are excluded
- Predatory journals were left
- Information or articles available in Blogs were not included

The inclusion and exclusion process were divided into the following 4 levels. Figure 4. Showcases the levels of inclusion/exclusion:

I. At level 1, papers were searched keeping in mind the keywords and inclusion-exclusion criteria. A total of 1270 papers were collected. After performing exclusion rules, 480 papers were included in the literature
II. At level 2, papers were added to the literature survey based on title and abstract reading. A total of 230 papers was added out of 480
III. At level 3, papers were added after reading the full article. Irrelevant articles were removed
IV. At last, the survey was done with 188 papers

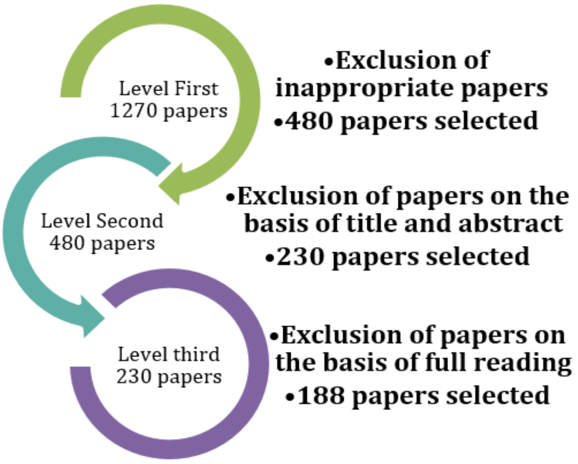
Level First
1270 papers
•Exclusion of inappropriate papers
•480 papers selected
Level Second
480 papers
•Exclusion of papers on the basis of title and abstract
•230 papers selected
Level third
230 papers
•Exclusion of papers on the basis of full reading
•188 papers selected

Fig. 4: Levels depicting inclusion/exclusion criteria

Parallel Data Extraction: The Proposed Architecture

The extraction of parallel data involves the number of tasks which are elaborated in Fig. 5. Firstly, there is the requirement of data resources for performing the task of extraction. But as mentioned earlier also in the survey, parallel data is not easily available in desired languages. To overcome this problem of scarcity, CC which are available in huge amount but in raw form can be used for the extraction of parallel data. There are three types of parallel data in CC i.e., “parallel sentences”, “parallel fragments” and “bilingual lexicons”.

A parallel data extraction consists of the following steps:

1. Potential resources, like comparable corpora in the desired language pair, Extraction of Bilingual Lexicons and a seed parallel dictionary.
2. Document Alignment model, to get similar document pairs.
3. Parallel Sentence and Fragment Extraction, to get parallel sentences and fragments from the aligned documents.
4. Improving SMT accuracy

Different techniques and methods used by enormous researchers for performing these steps are elaborated below.

Potential Resources

- Comparable Corpora

Comparable Corpora is composed of two languages textual data which are the raw translations of each other. The documents in CC are not properly aligned. Research scholars used various kinds of datasets as a source of comparable data like bilingual newspapers (Zhao and Vogel, 2002; Munteanu and Marcu, 2005; Tillmann,2009; Do *et al*., 2010), bilingual articles (Munteanu *et al*., 2004; Utiyama and Isahara, 2003; Abdul-Rauf and Schwenk, 2011; Abdul-Rauf *et al*., 2017), Web (Jiang *et al*., 2009; Hong *et al*., 2010), Wikipedia (Stefanescu and Ion, 2013; Chu *et al*., 2012; Chang *et al*., 2008; Adafre and Rijke, 2006; Smith *et al*., 2010; Mohammadi and Ghasem Aghaee; 2010; Archana *et al*., 2015, Chu *et al*., 2014b) and Social media (Ling *et al*., 2013). Non-parallel and non-aligned bilingual records make up a quasi-comparable corpus (Fung and Cheung, 2004b; Quirk *et al*., 2007). The TDT3 Corpus, which is a transcription of radio and TV reports in bilingual sentences and paraphrases, is an example of a quasi-comparable corpus. Data can also be collected from the “Internet Archive” (Resnik and Smith, 2003). It is a non-profit organization that archives the entire Web and the material is freely accessible via a Way back Machine Web Interface. Hindi and Punjabi data for the development of lexicons can also be taken from two conventional dictionaries available at Bhasha Vibhag and the National Book Trust. But this data has to be converted into digital format manually (Goyal and Lehal, 2010). Data can also be obtained from websites where bilingual transcripts are available such as Vikaspedia.in, e-books, film captions, online freely available encyclopedias, Quran, Bhagavat Gita and Bible (Premjith *et al*., 2019). Jindal *et al*., 2018 collected the raw data from different sources for creating the PC. English and Punjabi textual data were collected from online as well as offline resources. The raw form of data was collected from Gyan Nidhi, EMILLE, Bible, Guru Granth Sahib corpus available in electronic form, PSEB E-books, Bilingual Newspapers, tourism and health-related corpus from the web. Figure 6 depicts the percentage of different datasets used by different researchers mentioned in this survey. From the figure, it's prominent that news datasets and Wikipedia are common when creating comparable data.

- Parallel Seed Dictionary

The seed dictionary is the kind of glossary that contains the source word and its target translation. Seed dictionary is very important for training the machine. There is always the requirement of an external source like a seed dictionary along with the CC for sentence extraction and fragment extraction. Lakshmi *et al.* (2020) revealed that the dictionary is always a good alternative to CC and both can work hand to hand also. Table 4 provides some of the seed dictionaries used by researchers with various language pairs. The seed dictionary can be created manually (Utiyama and Isahara, 2003; Fung and Cheung, 2004; Adafre and Rijke, 2006; Lu *et al*., 2010; Jindal *et al*., 2018a; Deep *et al*., 2018) or a seed parallel corpus (Zhao and Vogel, 2002; Kumar and Goyal, 2010; Munteanu and Marcu, 2006; Ling *et al*., 2013; Smith *et al*., 2010; Tillmann, 2009; Lakshmi and Shambhavi, 2020; Gahbiche-Braham *et al*., 2011; Stefanescu and Ion, 2013; Stefanescu *et al*., 2012; Abdul and Schwenk, 2011) available can be utilized. Lu *et al*. (2010) provided a broad parallel corpus derived from an Internet-sourced corpus of comparable English-Chinese patents. First, parallel sentence pairs were formed using Champollion, a publicly available sentence aligner and then the candidates were filtered using MS Aligner, another publicly available sentence aligner. Around 7 million high-quality parallel sentences were chosen as the final parallel corpus from a pool of over 22 million bilingual sentence pair applicants. This is one of the patent domain's largest corpora of parallel sentences. Later, Zhu *et al*. (2011,2012) also designed a system that mined PC from web pages automatically. The system identified a decent number of parallel texts based on heuristic information extracted from web content for minority languages like Chinese-Mongolian. A similar kind of work was also done by Tan and Zhou (2010) for English and Chinese language pairs. It was also a web-based corpus that was parallel in nature. Kumar and Goyal (December 2010a) created a Hindi-Punjabi parallel corpus of 50,000 sentences based on a freely accessible Hindi-Punjabi machine translation system. The corpus is in .xml and .doc formats. The parallel corpus created was sentence-aligned. Few errors from categories such as out-of-vocabulary, grammar, inflection generation, transliteration, etc. were found when the parallel corpus was created. The current Hindi-Punjabi Machine Translation System was used to analyze the errors. The terms discovered during the study were applied to the machine translation dictionary that already existed. Jindal *et al*. (2018a) focused on creating an English-Punjabi corpus of big size. The use of a parallel corpus is important for statistical machine translation training. The creation of a corpus had huge challenges as raw data was not easily available in the required language pairs. English-Punjabi Corpus was generated because basic data was not available for regional language pairs. The raw text was obtained from different resources like Gyan Nidhi, EMILLE, Bible, Guru Granth Sahib electronic version, PSEB e-books, Bilingual newspapers, tourism and health websites. Also, Jindal *et al*. (2018b) worked upon English to Punjabi machine translation using free translation software called Moses (Koehn *et al*., 2007). In their research, they created a corpus of 20000 sentences that were of different domains. The sentences were aligned using the GIZA++ alignment tool. The accuracy was checked using BLEU scripts. Lakshmi and Shambhavi (2020) revealed that one of the promising resources to extract dictionaries is PC. Their study found that Comparable Corpora (CC) could be an alternative to extracting a dictionary. The proposed solution was to extract the dictionary for a low-resource language pair of English and Kannada using Comparable Corpora (CC) collected from Wikipedia dumps and corpus collected from the Indian Language Corpus Initiative (ILCI). Dictionary constructed comprises both translation and transliteration entities with term level associations from English to Kannada. The resulting dictionary is of size 77545 tokens with a precision score of 0.79.

Bilingual Lexicon Extraction

The oldest method of using CC is to extract bilingual lexicons. Artificial Intelligence and CLIR (Cross-Lingual Information Retrieval) (Widdows *et al*., 2002) both depend heavily on bilingual lexicons. A bilingual lexicon consists of words that are almost synonyms for one another (Haghighi *et al*., 2008). The bilingual lexicon is either hand-crafted or automatically produced from a Parallel Corpus (PC). Different systems for extraction have been elaborated in Table 5. From earlier work (Rapp, 1995) Bilingual Lexicon Extraction (BLE) has exploited Comparable Corpora (CC) for SMT. BLE's main aim is to create and enable bilingual dictionaries or seed lexicons, which are critical for both SMT and CLIR (Pirkola *et al*., 2001; Jagarlamudi and Kumaran, 2007; Chinnakotla *et al*., 2007). Their manual creation necessitates a high level of proficiency in both languages involved and can be a time-consuming operation. Vectors, Projections, Classifiers, Correlations, Linguistic information and other techniques may be used to derive bilingual lexicon from Comparable Corpora (CC). Goyal and Lehal (2010) worked on a direct translation approach. The data was collected for two closely related languages, Hindi and Punjabi in terms of grammar and vocabulary. Data were available in the form of hardcopy. It was then digitized and molded as required for machine translation. A lexicon of 1,00,000 words was manually created for word-to-word translation. The problem of ambiguity was resolved using a tri-gram approach. Using BiLDA topic models, Liu *et al*. (2013) developed a method for translating CC into a parallel aligned corpus, which is an advanced version of the LDA model (Blei, 2003) and with the aid of word alignment, defining word translations. Large-scale experiments in this study demonstrated that the proposed model introduced a range of benchmarks using both automated measures and manual assessments. The research also demonstrated that their subject-dependent translation systems are capable of capturing a few of the important poly-semi concepts in dictionary construction. Bouamor *et al*. (2013) later proposed a diverse approach for constructing a domain precise "bilingual lexicon" based on Wikipedia. These large multiple languages encyclopedia paved the way for the development of lexicons for a vast range of language pairs. Gaussier and Li (2010) proposed a comparability metric and then created a model for improving a CC by eliminating a subpart and completing the left subpart with external tools. They showed how to improve bilingual lexicon extraction using information gathered during the building process.

Fung and Yee (1998) described an associate algorithmic rule for mining bilingual lexicon from CC for the English-Chinese language domain. This algorithmic rule was language independent and took into account the burden of bilingual seed words. Additional language sets, such as English-French or English-German, were also benefited. This computational rule can also be implemented in a repetitive manner where better bilingual word pairs are added to the seed word list, yielding additional new bilingual similar words. Xu *et al*. (2011) explained the context-based approach for the creation of bilingual lexicons from CC. The experiments showed the mapping of context words, directions and types of dependency relationships. The proposed method surpassed the state-of-the-art scheme in bilingual lexicon creation for language sets of English and Chinese. Later Qian *et al*. (2012) discussed a comparable corpus, a bilingual dependency mapping model for bilingual lexicon building from English to Chinese. This model considers both dependent words and their relationships when measuring the similarity between bilingual words and thus offers a more precise and less noisy representation. It also illustrated that bilingual dependency mappings can be created and optimized automatically without human input, contributing to a medium-sized set of dependency mappings and that their impacts on Bilingual Lexicon Construction (BLC) can be fully exploited through weight learning using a simple but effective perceptron algorithm, making their approach quickly adaptable to several other language pairs.

For BLE from CC, Bouamor *et al*. (2013) presented the associated degree approach. This research focuses on the unresolved issue of polysemantic words discovered by dictionaries and suggests the need for an acceptance clarification approach to boost the appropriateness of context vectors. Empirical experimental findings on two advanced French English CC showed that the technique outperformed two state-of-the-art approaches. The most widely used methods for the BLE from CC were evaluated in comparison by Hazem and Morin (2013a). Their observations supported the hypothesis, that using a re-estimation methodology of word co-occurrence in a similar corpus can improve the accuracy of the standard method. A year apart, Chu *et al*. (2014a) developed a scheme for extracting bilingual lexicons that combined topic-based (Vulic *et al*., 2011) and context-based (Rapp, 1999, Harastani *et al*.2013) methods. Experimental studies on Chinese–English and Japanese–English Wikipedia data revealed that their proposed approach outdoes a state–of–the–art technology. Cao *et al*. (2007) also identified a system that extracts English-Chinese translation combinations mechanically from a substantial quantity of monolingual Chinese web material. Candidate translations are derived using pre-specified models in this method. On over 300GB of Chinese content online, the study compares a variety of approaches to aligning transliterations and mining translations.

To provide a new perspective on BLE, Gaussier *et al*. (2004) demonstrated the geometric form of BL extraction from CC. Evaluations of the strategies were proposed on a comparable corpus extracted from the CLEF collection and showed the strengths and weaknesses of each technique. The final results showed that the mixture of comparatively straightforward strategies helped in improving the average preciseness of BL extraction approaches from CC by ten points.

Document Alignment Model

CC can be huge in size so it is quite difficult to examine every sentence in the corpora. So, the concentration is made only on those documents and sentences which have similar kinds of content. For finding similar or comparable documents, techniques like topic alignment (Zhu *et al*., 2013), content alignment, text alignment and cosine similarity can be employed. In parallel data extraction, different authors employed various document alignment techniques which are mentioned in Table 6.

Table 4: Seed dictionary used by authors for various language pairs

Language Pair	Citation
Japanese-English	Utiyama and Isahara (2003)
English-Chinese	Fung and Cheung (2004)
English-Chinese	Tan and Zhou (2010)
Dutch-English	Adafre and Rijke,2006
Chinese-Mongolian	Zhu *et al*. (2012)
English-Chinese	Lu *et al*. (2010)
English-Punjabi	Jindal *et al*. (2018a)
Punjabi-English	Deep *et al*. (2018)
English-Chinese	Zhao and Vogel (2002)
Hindi-Punjabi	Kumar and Goyal (2010)
Spanish-English	Tillmann (2009)
English-Kannada	Lakshmi and Shambhavi (2020)
Arabic-French	Gahbiche-Braham *et al*. (2011)
French-English	Abdul and Schwenk (2011)
English-German, English-Romanian, English-Spanish	Stefanescu and Ion (2013)

Table 5: Bilingual lexicon extraction techniques

Technique of bilingual lexicon extraction	Citation	Language pair used
Correlation based extraction	Fung and McKeown (1994)	Asian/Indo-European
Correlation based extraction	Rapp (1999)	English-German
Domain Specific bilingual extraction	Chiao and Zweigenbaum (2002)	French-English
Associative algorithmic rule	Fung and Yee (1998)	English-Chinese
Context Heterogeneity	Fung (1995)	English-Chinese
Iterative Extraction	Fung and Cheung (2004b)	Chinese- English
Context and Lexical combined extraction	Déjean *et al*. (2002)	German-English
Topic model-based extraction	Vulic *et al*. (2011)	English-Italian
Linguistic knowledge with Topic Distribution	Vulic and Moens (2012)	Dutch, Italian, English
Geometric Interpretation	Gaussier *et al*. (2004)	English-French
Signal Processing and Parallel Corpora approach	Munteanu and Marcu (2006)	Romanian-English
Support Vector Machine	Brockett (2005)	English
Lexico Syntactic Method	Otero (2007)	English-Spanish
Clustering-based approach	Li *et al*. (2011a)	English-French
Victimization applied math learning	Cao *et al*. (2007)	English- Chinese
Topic and context-based combined	Chu *et al*. (2014a)	Chinese, English, Japanese
Smoothing strategy	Hazem and Morin (2013b)	English-French
Associated degree approach	Bouamor *et al*. (2013)	English-French
Bilingual dependency mapping	Qian *et al*. (2012), Xu *et al*. (2011)	English-Chinese
comparability metric	Gaussier and Li (2010)	English-French
BiLDA model	Liu *et al*. (2013)	English-French
Direct translation approach	Goyal and Lehal (2010)	Hindi-Punjabi

Table 6: Document alignment techniques with citations

Document alignment techniques	Citations
Cosine similarity	Fung and Yee (1998; Garera *et al*.,2009; Prochasson and Fung, 2011; Tamura *et al*., 2012; Lehal *et al*., 2019)
Topic alignment	Fung and Cheung (2004; Gahbiche-Braham, 2011; Li, 2011; Zhu *et al*., 2013; Goyal *et al*., (2020)
Context alignment	Gale and Church (1991, Resnik and Smith, 2003, Zesch and Gurevych, 2010)

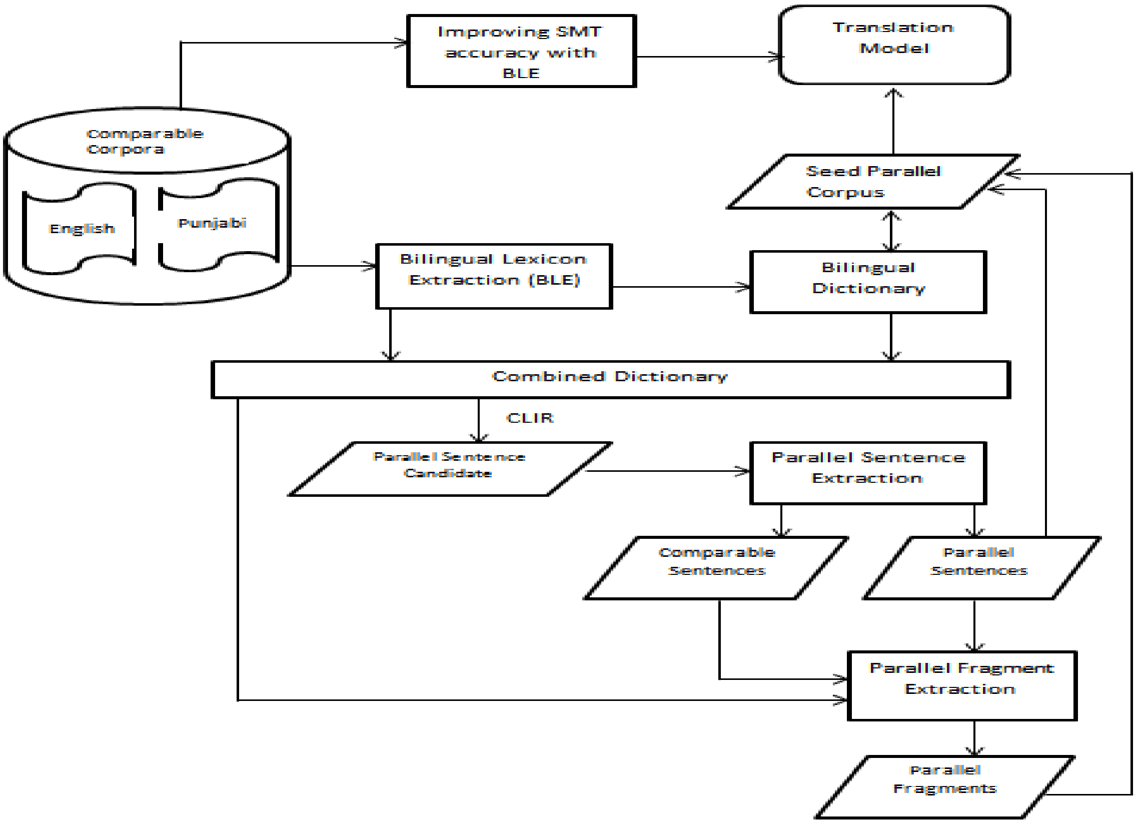

Fig. 5: Proposed model for parallel data extraction

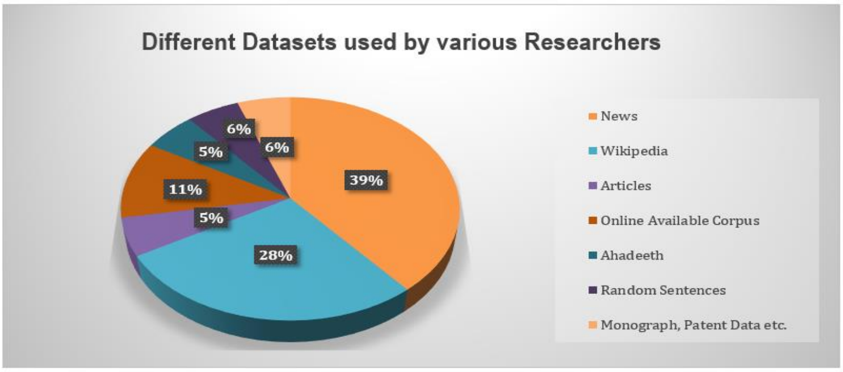

Fig. 6: Datasets used by different researchers mentioned in this survey

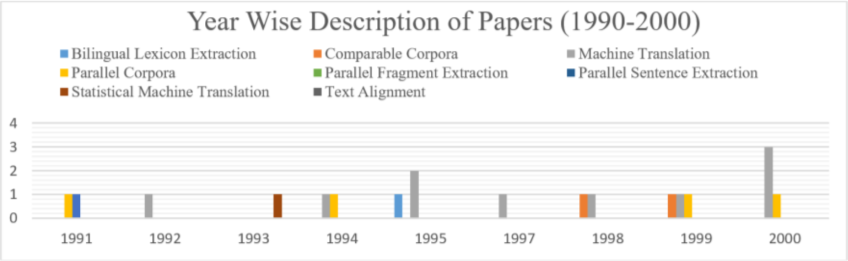

Fig. 7: Year-wise description of papers (1990-2000)

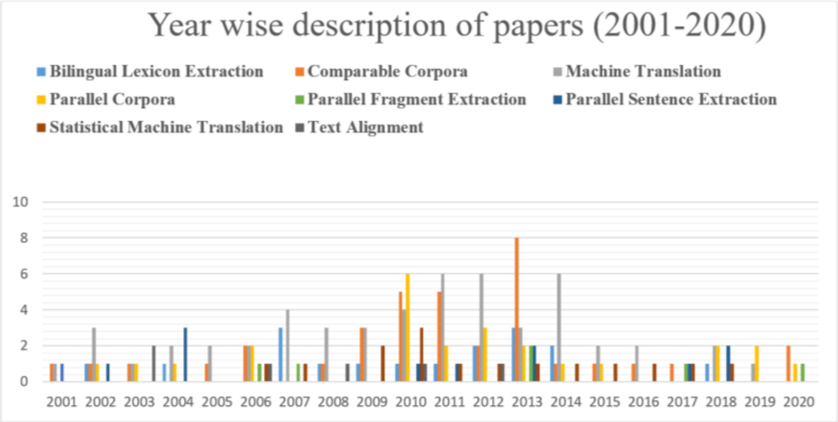

Fig. 8: Year-wise description of papers (2001-2020)

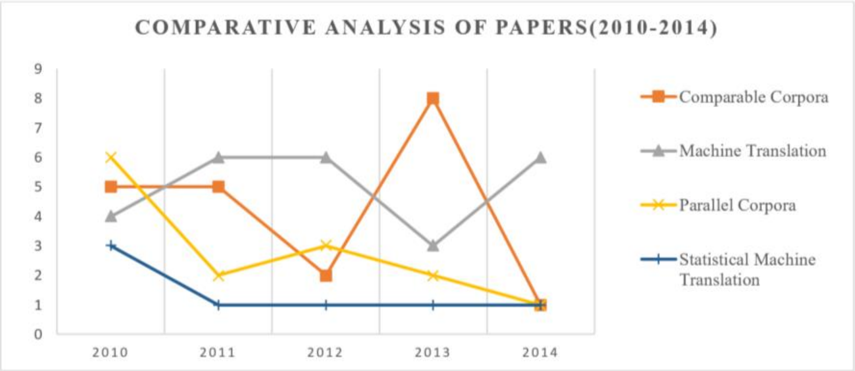

Fig. 9: Comparative analysis with respect to keywords for the year 2010-2014

Cosine Similarity

Cosine similarity is used to compute the similarity amongst the two documents described as vectors of the terms they contain. Cosine similarity is defined as the Dot product of the vectors (Fung and Cheung, 2004a). Lehal *et al.* (2019) compared the similarity and distance measures. Their research has analyzed and compared cosine similarity, Jaccard coefficient, Hamming distance and Euclidean distance (Fung, 1995; Yu and Tsujii, 2009). The accuracy levels were found using these metrices. It was concluded that if the data is imbalanced, accuracy will lack in providing the true efficiency. In that scenario, precision and recall will give better results. In terms of Euclidean Distance, Cosine Similarity and Jaccard Similarity, precision gave high result above 95% as compared to Hamming distance. However, the value of recall was high than precision in Hamming Distance. In analyzing cosine similarity, the f1 score and accuracy was much better, than in any other similarity measures. The data was taken from Wikipedia in Punjabi and English language. Numerous works are done in the field of web alignment. Online available alignment software and websites help in achieving the target of alignment for all kinds of textual data (Nie *et al*., 1999; Zhang *et al*., 2006; Fung *et al*., 2010; Uszkoreit *et al*., 2010).

Topic Alignment

Goyal *et al.* (2020) described the process of aligning the documents based on topics. A comparable corpus of English-Punjabi originated from the dump taken from Wikipedia. PHP scripts were developed for fetching and aligning articles. Articles have been aligned in two separate directories. Its corresponding English record was detected for each Punjabi record. A corpus was created which could be used for parallel data extraction. Utiyama and Isahara (2003) suggested two measures to ensure the correct alignment of article and sentence. Similarities in sentences associated with Dynamic Programming (DP) matching and similarities in papers matched with Cross-Language Information Retrieval (CLIR) (Utiyama and Isahara, 2003; Fung and Cheung, 2004b; Munteanu and Marcu, 2005; Gahbiche-Braham, 2011) for sentence alignment are used in the article alignment test. The experiments involved the enhancement of each other and permitted the accurate mining of the related article and phrase alignments from the excessively noisy parallel Japanese-English corpus. An effective large-scale article and sentence alignment corpus was built and made available to the public using these steps. Li, (2011b) also introduced a technique that can select candidate sentences for sentence alignment. The technique has mainly experimented on bilingual Comparable Corpora (CC) obtained from Wikipedia for English and Chinese language pairs.

Context Alignment

Context alignment was anticipated by Gale and Church (1991) and Brown *et al.* (1993). Brown *et al.* (1993) defined a set of five applied mathematics variants of the interpretation method and provided algorithms for estimating their parameters, resulting in a set of pairs of sentences that are translations of each other. The study tended to illustrate a thought of word-by-word alignment between certain pairs of sentences and also offered an associated formula for finding the highest likelihood of such alignment. Though its formula is sub-optimal, the alignment thus provided good accounts for word-by-word associations within the combined sentences. Resnik and Smith (2003) also revealed word-to-word translation. This technique employs translation similarities based on the word-by-word translation lexicon. It is also known as content-based alignment. BLE from CC is built on the distributional hypothesis (Harris, 1954) that terms with identical meanings feature in identical dispersals across languages. Srivastava and Sanyal (2012) presented an approach that increased the performance of word alignment for small PC of the English-Hindi language pair. Their model used POS tagging with word alignment and expressed the significant decrease in Alignment Error Rate. Post *et al.* (2012) compiled and fine-tuned PC at the document level between English and six verb-final languages: Bengali, Hindi, Malayalam, Tamil, Telugu and Urdu. The set of six Parallel Corpora (PC) containing four-way redundant translations of the source-language text was identified in their research. They revealed that the Indian languages of these corpora are low-resource and understudied and exhibit markedly different linguistic properties compared to English. Their study included performing baseline experiments quantifying the translation performance of several systems, investigated the effect of data quality on model quality and suggested many approaches that could improve the quality of models constructed from the datasets. They also concluded that the PC provides a suite of SOV languages for translation research and experiments.

Parallel Sentences and Fragments Extraction

The data collected through different web sources in the form of comparable, quasi-comparable, or noisy parallel is used to mine the parallel data in the form of sentences and fragments. PC are phrase-aligned bilingual documents. They are vital tools for natural language production in bilingual or multilingual contexts (Zhu *et al*., 2012). PC provides the majority of translation expertise, but the quality and quantity of PC are limited. A significant portion of the phrases encountered at run-time in such language pairs is unknown. Integration paraphrases into applied mathematics computational linguistics, according to Burch *et al*. (2006), would hold crucial improvements in coverage and translation accuracy. Paraphrases, in essence, introduced a degree of generalization into applied mathematics and computational linguistics. Their study was able to take advantage of information outside of the interpretation paradigm, such as terms with similar meanings and apply it to the translation process. Parallel sentences can be recognized dependent on classification (Munteanu and Marcu, 2005; Tillmann, 2009; Smith *et al*., 2010; Bharadwaj and Varma, 2011; Stefanescu *et al*., 2012) or by utilizing similarity procedures (Utiyama and Isahara, 2003; Fung and Cheung, 2004; Fung *et al*., 2010; Abdul-Rauf and Schwenk, 2011; Abdul-Rauf *et al*., 2017).

Various techniques for sentence extraction along with language pairs are mentioned in Table 7. Also, the techniques are elaborated below which were used by authors for classifying and mining the sentences from the aligned documents:

Classification

Munteanu and Marcu (2005) used a classifier for mining parallel sentences. The model uses linear functions. It classifies the sentences into parallel and non-parallel classes. But there was an error in this classification process as the maximum of the sentences were termed as non-parallel. This created an imbalance in extraction. Chu *et al*. (2013a) also suggested a procedure for extracting sentences from a quasi-comparable corpus. The system trained and tested a unique classifier that stimulates parallel sentence extraction. The study used linguistic information of Chinese characters for extraction.

Maximum Entropy Ranking Model

Smith *et al*. (2010) used the Maximum entropy model to rectify the problem faced in the above technique. The same model which was used in the classification technique was used here also. In this, the sentences are chosen based on probability scores. The higher the score, the more is the chance of the sentence being parallel.

Sentence Similarity

Fung and Cheung (2004a) suggested a multi-level bootstrapping method for parallel sentence extraction from quasi-comparable corpora. The research examined the suitability of various bilingual corpora for a trilingual natural language system. Beginning with parallel, comparable and non-parallel corpora, a variety of bilingual corpora were contrasted and differentiated. A lexical alignment score measured for the bi-lexicon tried within the matched bilingual sentence pairs is then used to test the usability of each corpus type. Fung *et al*. (2010) introduced a new multilingual web crawler and sentence extracting method for mining and extracting parallel sentences from trillions of websites with no regard for domain or address architectures or publication dates. Their primary goal is to improve applied computational machine translation frameworks.

Conditional Random Field

For aligning the parallel sentences, Smith *et al*. (2010) made use of the conditional random field. In this, only the sentences which are present in the aligned documents can be extracted. This same technique was also followed by Blunsom and Cohn (2006). The study worked on a small set of data and made use of GIZA++ for training purposes. In this technique, every word gets aligned to its target word and in reciprocation, the target word can get aligned to the number of source words. Wolk and Marasek (2014) presented a method that constructed PC from noisy parallel and CC. Wikipedia data as a source was selected for Polish and English languages. A web crawler was used for obtaining the bilingual articles from Wikipedia. The Hunalign tool was used for sentence alignment. Freely available translators were used for Polish language translation to English. MGIZA++ tool was used for word and sentence alignment. At last, the training was done using Moses which is an open-source SMT-related toolkit (Koehn *et al*., 2007). For evaluation BLEU was utilized. At last, for evaluating the quality and quantity of evaluation, human translators were used for manually aligning the articles on the sentence level. This study lacked due to human intervention and also due to fewer data available.

LEXACC

It stands for “Lucene based Parallel Sentence Extraction from Comparable Corpora”. Stefanescu and Ion (2013) identified a series of parallel sentences for three sets of languages: English-German, English-Romanian and English-Spanish, which were extracted from Wikipedia. To do so, they used a method called LEXACC, which was developed during their project and was used to extract parallel sentences from CC. Stefanescu *et al*., 2012 made use of CLIR to find parallel sentences. With the help of a seed dictionary, the source words are translated to target words. Rahimi *et al*. (2016) explained CLIR (“Cross-Language Information Retrieval”) and extraction of translations from CC for CLIR. CLIR is directly linked with the translation quality, so there is the requirement for a proper translation model from the available CC. The experimental work involved the gathering of English-Persian CC which was obtained from news articles in both languages. A successful translation model was built from CC available without any additional linguistic tools. To extract correlations between each pair of bilingual terms, a language modeling method was proposed. Integration of monolingual relations of word co-occurrences was done with translational relations for the translation of low-frequency terms. Various estimates of translation probabilities from word correlations have been compared. It was, therefore, claimed that the calculation affected the efficiency of cross-language information retrieval.

Some other authors also contributed in "Parallel Sentence Extraction". Kumar and Goyal (2018) used a mathematical approach to investigate the design of a Hindi to Punjabi machine translation method. The set of 3 lakh parallel sentences was the starting point for the creation of a machine translation method. The parallel sentences have been developed using different tools like Akhar, Microsoft's bilingual sentence aligner, spell checker, Tokenizer and translation software mentioned in the study of Kumar and Goyal (2012). The parallel corpus used by Kumar and Goyal (2018) was supplemented with approximately one lakh Hindi-Punjabi lexicons. For statistical analysis of the Hindi and Punjabi languages, pre-processing and post-processing modules were developed. For pre-processing, "Word Tokenizer" and "Text Normalization" modules were developed. For precision, the Transliteration and Grammatical Error Correction modules were used. The GIZA++ tool was used to create the translation model and Moses (Koehn *et al*., 2007) software was used as the decoder. The BLEU and NIST scores are used to assess quality. Deep *et al*. (2018) provided in the research different sources to collect the English data and Punjabi data. Their work presents the Punjabi -English parallel corpus and named it Pun Eng. They used the human translation approach and online translators for converting the data into the required language. Entire data was cleaned and unnecessary tags were removed manually. After removing all the tags, translations by google translate and human verification, they were left with parallel sentences. Premith *et al*. (2019) presented a neural MT technique for building four Parallel Corpora (PC) in the language combinations English-Malayalam, English-Hindi, English-Tamil and English-Punjabi. The information was gathered in the form of text from both online and offline sources. The models obtained were tested both automatically and physically. The BLEU score was used for automatic evaluation and three criteria, fluency, rating and adequacy, were used for manual evaluation. Long sentences were found in the English-Malayalam and English-Hindi corpora, which influenced the translation. In addition, the attention mechanism was applied to the issue of translating long sentences. Their findings revealed that, in addition to the corpus' size and coverage, the length of sentences plays an important role in translation efficiency. Later, Agic and Vulic (2020) created a parallel corpus for 300 languages with nearly a lakh of sentences in a single language. Their study work on extracting parallel sentences and creating a corpus of them. The corpus thus created was named JW 300 and is freely available online. The corpus created could be used for part of speech tagging projects as well as for cross-lingual procedures.

When there are fewer parallel data, then the focus is turned towards non-parallel data. The extraction of sentences from non-parallel is not feasible. So, there arises the requirement of fragment extraction. Fragments are the phrases present in the sentences.

Various techniques to mine the fragments are used by the number of authors. Below mentioned are the approaches for fragment mining.

Log-Likelihood Ratio

Munteanu and Marcu (2006) used this technique to extract the segments. For using this approach, the system is provided with some sentence pairs from the corpus. These sentence pairs are obtained by making the use of the GIZA++ tool on the given data. After getting sentence pairs, fragments are extracted from only those sentences which have an exact translation. The only drawback concerned with this technique is that the system has to be provided with correct translated words in the seed dictionary.

Sentence Splitting for Phrase Alignment

Hewavitharana and Vogel (2013) made use of this technique for the extraction of fragments from the available nonparallel corpus. In this technique, a source fragment and sentence pair are taken. Then the alignment of words is done. The words are combined with the help of heuristics. On the translated side, some split points are looked at. Splits points are searched based on the probability of word alignment. In this, words inside the source phrase align with the words inside the target phrase. Words outside the source phrase get aligned to the words outside the target phrase. Alignment goes hand in hand on both the source and target sides.

Chunking Approach

Chunk is the small part/phrase of the sentence. These small phrases are extracted from the main sentence and translation is done on that phrase. The chunk can be placed anywhere in the target sentence. The words in the chunk remain the same even after translation. Gupta *et al*. (2013) used the chunking method to translate a source fragment and measured the similarity between the translated source and target fragments to classify the target fragment. The study revealed the use of an automated method for extracting parallel English-Bengali text fragments from CC generated using Wikipedia materials. The method takes advantage of Wikipedia's multilingualism. The study also found that using an out-of-domain corpus was beneficial in training a site-specific MT system.

Table 7: Parallel sentence extraction techniques

Extraction technique	Citation	Language domains
Classification	Munteanu and Marcu (2005)	Chinese, Arabic, English, German
	Tillmann (2009)	Spanish- English
	Bharadwaj and Varma (2011)	English-Hindi
	Chu *et al*. (2013)	Chinese
Maximum Entropy Model	Smith *et al*., 2010	Spanish-English, Bulgarian-English, German-English
Sentence Similarity	Fung and Cheung (2004a)	English-Chinese
	Utiyama and Isahara (2003)	Japanese-English
	Fung *et al*. (2010)	Chinese
	Abdul and Schwenk (2011)	Arabic-English, French-English
Conditional Random Field	Smith *et al*. (2010)	Spanish-English, Bulgarian-English, German-English
	Blunsom and Cohn (2006)	French-English, Romanian- English
	Och and Ney (2003)	German-English, French-English
LEXACC	Stefanescu *et al*. (2012)	English, Estonian, German, Greek, Lithuanian, Latvian, Romanian, Slovene
	Stefanescu and Ion (2013)	English-German, English-Romanian and English-Spanish

Improving SMT Accuracy

In "Statistical Machine Translation" (Koehn, 2009; Och and Ney, 2003; Brown *et al*., 1993), the translation model is trained in unsupervised manner from parallel corpora. The translation model consists of translation pairs as well as the feature scores. The accuracy of SMT is hampered due to inaccurate translation pairs and feature scores. Inaccuracy arises due to paucity of parallel corpora. Accuracy can be improved by:

- Increasing the amount of parallel corpora
- Filtering the noise translation pairs from translation model
- Estimating new features from comparable corpora for the translation pairs

Parallel corpora are not easily available for number of languages and domains. So, increasing the quantity of parallel corpora is not an easy task.

Filtering the noisy translation pairs can no doubt increase the accuracy but it can also lead to removal of some good translation pairs. This further will decrease the coverage of translation model.

Comparable features (Irvine and Callison, 2013) such as similarity scores obtained from comparable corpora can be combined with original features to differentiate between good and bad translation pairs. BLE can be used to justify the accuracy issues in SMT. Different similarities like topical, contextual (Rapp, 1999), orthographic and temporal can be individually used or combined together for bilingual lexicon extraction. SMT quality and coverage issues were discussed simultaneously with BLE by Irvine and Callison (2013); Pal *et al*. (2014); Marton *et al.* (2009); Ganitkevitch and Callison-Burch (2014). For six languages with limited resources, a comparable corpus was used to validate the performance and scope of phrase-based Machine Translation models developed with small bilingual corpora. The results of the experiments show that each of these approaches increases the performance of the BLEU score on its own. Nevertheless, the findings suggest that having low frequency word translations increases efficiency more than translations for OOVs (out-of-vocabulary) (Callison-Burch *et al*., 2006) alone. The results showed improvement for lesser data for parallel training. Richardson *et al.* (2013) illustrated that the implementation of contextual features can dramatically improve the efficiency of transliteration. In addition, even for out-of-domain source terms that have an unknown distribution of the subject, their extended model may produce a considerable improvement of accuracy.

Chu *et al*. (2014a) made the use of paraphrases along with BLE to rectify the problem of accuracy. Paraphrases can also be used as training data to improve the accuracy of SMT. Paraphrase can be generated from parallel corpus and thus can reduce the problem of data sparseness also.

Results and Discussion

The findings of the systematic literature review are organized in accordance with the research questions which are mentioned in Table 3. A total of 188 papers were reviewed in this survey. The survey focused on the proposed technique through which parallel data could be extracted from the given nonparallel data. Different ways of "parallel data extractions" used by researchers are mentioned in this survey. Out of 188 papers, 34% literature review is done on the works under the term "Machine Translation" whereas 9% of papers are found on "Statistical Machine Translation". Furthermore, 20 and 15% of the papers are found on "Comparable Corpora" and "Parallel Corpora" respectively. Additionally, 10% of papers are found on "Bilingual Lexicon Extraction" and 6% of papers contributed towards "Parallel Sentence Extraction" which are published in esteemed journals, conferences and workshops depicted in Fig. 3.

The research papers are collected from databases like IEEE, ACL, ACM, Springer, Science Direct and some journals which are indexed in Google Scholar and Citeseer. A total of 188 papers were selected for writing this review. Out of the total papers included, 12% of research articles are printed in IEEE, 16% in ACM, 9% in Springer and 4% in Science Direct. ACL contributed 26% in writing this review. 34% of papers were the ones that were accepted in some conferences and workshops but are indexed in Google Scholar. The contribution of papers from different resources is clearly depicted in Fig. 2.

The survey performed is based on questions framed which are mentioned in Table 3. We will provide insight into these questions with justifications as per the literature we reviewed.

RQ1: What is the current status of data extraction in respect of parallel and comparable data?

It has been discovered from the literature survey that for extracting the parallel data, there is the requirement of many things such as comparable data, a seed dictionary, bilingual lexicons and some alignment tools. Even after that, the parallel data is retrieved from some parallel fragments and sentences. For the realization of this process, different databases were searched for relevant works in these fields. Around 188 papers were studied after following inclusion/exclusion criteria mentioned in Research Process and also in Fig. 4. Different researchers used various techniques for finding comparable data, aligning the data, extracting lexicons, fragments and sentences in different years. Very little work came to light from the period 1990-2000 as shown in Fig. 7. With the passage of time and improvements in technology, a noticeable amount of research was carried out from the period 2001-2020 as shown in Fig. 7.

It is observed from Fig. 8. that most of the work was done from 2010 to 2014. Fig. 9. shows the comparative analysis for the period 2010-2014. The comparative analysis is based on the selected research papers published concerning the keywords used in the literature survey. It’s noticeable from Fig. 9 that 25 papers were published in the said period on “Machine Translation”, 7 papers on “Statistical Machine Translation”, 21 on “Comparable Corpora” and 14 on “Parallel Corpora”.

RQ2: What are the various datasets used in the process of extraction?

The literature survey showed the usage of multiple kinds of datasets by various authors for the mining of parallel data from comparable data. Different researchers worked with a variety of European and Asian language pairs. The textual data was taken from numerous online as well offline resources. Data was taken from Wikipedia (Stefanescu and Ion, 2013; Chu *et al*., 2014; Adafre and Rijke, 2006; Smith *et al*., 2010; Mohammadi and Ghasem , 2010), Bilingual newspapers (Zhao and Vogel, 2002; Munteanu and Marcu, 2005; Tillmann, 2009; Do *et al*., 2010), E-books like from Gyan Nidhi, PSEB E-books, Bible and social media, Bilingual websites etc. Table 8 elaborates about different datasets used by researchers in this survey of 188 papers. The table also depicts the size of various datasets. Also, Fig. 6 shows the percentage of different datasets used by various researchers mentioned in this survey. It's evident from Fig. 6 that datasets of news and Wikipedia are largely used by researchers. While conducting this survey it's seen that 39% of researchers used news or newspapers as a source of data for comparable corpora. News is easily available in bilingual forms. Also, Wikipedia acted as a major source of comparable data, contributing nearly 28%. Wikipedia has its translator for various languages. In the literature survey its evident that Wikipedia is widely used by researchers in their work because of its easy availability.

RQ3: What type of machine translation approaches are used for translation in different language domains?

The systematic literature survey also focused on different works done in the field of “Machine Translation”. There are different approaches in MT such as Rule-Based, Corpus-Based and Hybrid. All these approaches are further subdivided into Interlingual, Statistical, direct, etc. Figure 1. presents the division of various approaches of MT. Based on these approaches, several MT Systems were created by various researchers. In this literature survey, we gathered 188 papers in contrast with some important keywords mentioned in Fig. 3. With a focus on these 188 papers, MT systems created by several authors were studied. Table 1 depicts about various MT systems created for Indian languages along with some key features. Table 1 also reveals about the language pairs used by researchers in creating the MT Systems. Whereas Table 2 reveals the MT Systems created by researchers for European and Asian languages (other than India). In this table also some prominent features used in the creation of MT Systems by authors have been mentioned with the MT approach followed. We saw from both Table 1 and 2 that despite work done in the field of MT, still a lot is to accomplish in the field of “Statistical Machine Translation” without human intervention. This literature survey focuses on the “Statistical Machine Translation approach” and various methods to mine the parallel data from comparable data under SMT.

RQ4: What kind of parallel data can be mined by using CC?

Parallel data is the collection of texts in two or more languages with exact translations. “Parallel Corpora” is the aligned text where one is the source language and the other is the target language. A target language is the one in which translation is made. PC is of huge requirement when translations are done in the context of SMT. But PC are still a scarce resource due to their non-availability in good quantity and quality (Ali *et al*., 2010; Srivastava and Bhat, 2013; Post *et al*., 2012). So, CC are exploited to get parallel data from it in the form of lexicons, fragments and sentences. Bilingual datasets can be easily created from textual dumps of different languages, available through Wikipedia, Bilateral articles, etc. This data can further be filtered, aligned and cleaned to form "comparable corpora". Different techniques are used for the mining of bilingual lexicons, parallel sentences and fragments which are mentioned under heading "Parallel Data Extraction" of this literature survey. Figure 10. presents a mind map that covers numerous aspects, properties, extraction methods of lexicons, sentences and fragments which are derived from this literature paper. "Bilingual lexicons", "Parallel Fragments" and "Parallel Sentences" collaborate to form "Parallel Corpora". All the mining techniques are elaborated in sections namely "Parallel Resources" and "Parallel Sentence and Fragment Extraction". Also, Table 5 and 7 give an insight into various techniques used by authors for "Bilingual Lexicon Extraction" and "Parallel Sentence Extraction" respectively.

RQ5: What are the different kinds of parallel sentence and fragment extraction techniques followed?

The systematic literature survey aims towards the ways of mining parallel data from the available non-parallel data. As there is an unavailability of a good amount of parallel data so the concentration moves towards the mining of comparable data. From Comparable data, "Parallel Sentences" and "Parallel Fragments" could be easily mined. This survey focuses on 188 papers that are gathered after implying the Inclusion/Exclusion Criteria mentioned in Research Process. After exploring 188 papers, the survey report manages to collaborate various extraction techniques of "Parallel Sentences" shown in Table 7. Also, the paper gives more clear insight into the mining procedures followed by numerous authors for different language pairs in terms of "Parallel Sentence and Fragment Extraction".

RQ6: What are the different ways to extract Bilingual Lexicons from comparable data?

Many bilingual Natural Language Processing (NLP) tasks, such as "statistical machine translation", rely heavily on bilingual lexicons. Meanwhile, automatic construction of bilingual lexicons is desirable because manual construction is extremely tedious and costly. So, one approach is to mine bilingual lexicons from "parallel corpora". As earlier clarified in Introduction of this study, "parallel corpora" is not available in a good amount and better quality. Extracting "bilingual lexicons" from CC is an appealing option since "comparable corpora" are much more commonly accessible than "parallel corpora". Lexicons also act as an integral part for the building of PC. Table 5 presents various extraction techniques of "bilingual lexicons" used in 188 papers that are included in the survey. Furthermore, detailed invasion in BLE is provided in section named "Bilingual Lexicon Extraction" of this literature review.

Table 8: Datasets used by various authors along with their size and language pair

Dataset used	Data size	Language pair	Citation
Multilingual newspapers	French:333M words. English:527M words	French-English	Abdul and Schwenk (2011)
Wikipedia	Dutch:18 sentences. English:65 sentences	English-Dutch	Adafre and De Rijke (2006)
Sampark	100,200,500 and 1000 sentences	Hindi-Punjabi	Ahmad *et al*. (2011)
News	-	English-German English-Greek English-Latvian	Aker *et al*. (2012)
Ahadeeth	6000 sentences	English-Urdu	Ali *et al*. (2010)
Ahadeeth	20173 sentence pairs	English-Urdu	Ali *et al*. (2013)
News	315 sentences	English-Hindi	Ananthakrishnan *et al*. (2006)
News	100000 sentences	English-Chinese	Bai *et al*. (2008)
Random Sentences	500 Sentences	English-Sanskrit	Bahadur *et al*. (2012)
Verbmobil corpus	23k Sentences	Farsi-English, English-German, Farsi-German	Bakhshaei *et al*. (2010)
Wikipedia	1600 words	English-Hindi	Bharadwaj and Varma (2011)
News	Czech:1.1M English:1.2M	English-Czech	Bojar (2007)
News	French-English:183,000	French-English	Bouamor and Sajjad (2018)
Wikipedia	French-English:193,543 Romanian-English: 136,681	French-English, Romanian-English	Bouamor *et al*. (2013)
Wikipedia	French:799,010 words English:12,81,645 words	French-English	Bouamor *et al*. (2013)
News	10,000 Sentence pairs	English	Brockett (2005)
Random Sentences	500k Sentences	Bengali-Hindi	Chatterji *et al*. (2011)
Medical Corpora	French:6,02,484 words English:6,08,320 words	French-English	Chiao and Zweigenbaum (2002)
News	50 topics each	Hindi-English, Marathi-English	Chinnakotla *et al*. (2007)
Wikipedia	680k sentences	Chinese-Japanese	Chu *et al* (2013a)
Bilingual articles	Chinese:420k sentences Japanese:5M sentences	Chinese-Japanese	Chu *et al* (2013b)
Wikipedia	Chinese:2.1M sentences Japanese:3.5M sentences	Chinese-Japanese	Chu *et al* (2014b)
Wikipedia, Gyan Nidhi,			

Table 8: Continue

Emille, Tdil, Newspaper, etc.	English:3.97M words Punjabi: 4.28M words	English-Punjabi	Deep *et al.* (2018)
TDT3	Chinese: 110,000 sentences English: 290,000 sentences	English-Chinese	Fung and cheung (2004a)
News	Chinese:64M words English: 70M words	Chinese-English	Fu *et al.* (2013)
News	Arabic: 1 M sentences French: 5 M sentences	Arabic-French	Gahbiche-Braham *et al.* (2011)
News	52 sentences	English-Turkish	Hakkani *et al.* (1998)
Academic Monograph	5k Sentences	Chinese-Korean	Kang *et al.* (2005)
Wikipedia, Gyan Nidhi, Emille, Tdil, Newspaper, etc.	3 lakh sentences	Hindi-Punjabi	Kumar and Goyal (2010b)
Patent data	22M sentences	English-Chinese	Lu *et al.* (2010)
Wikipedia	12530 Sentence Pairs	English-Persian	Mohammadi and Ghasem Aghaee (2010)
Wikipedia	200000 articles	English-German English-Spanish English-Romanian	Stefanescu and Ion (2013)
Bilingual news	1.35M sentences	Spanish-English Portuguese-English	Tillmann (2009)
Bilingual news	17310 news pair	English-Chinese	Zhao and Vogel (2002)

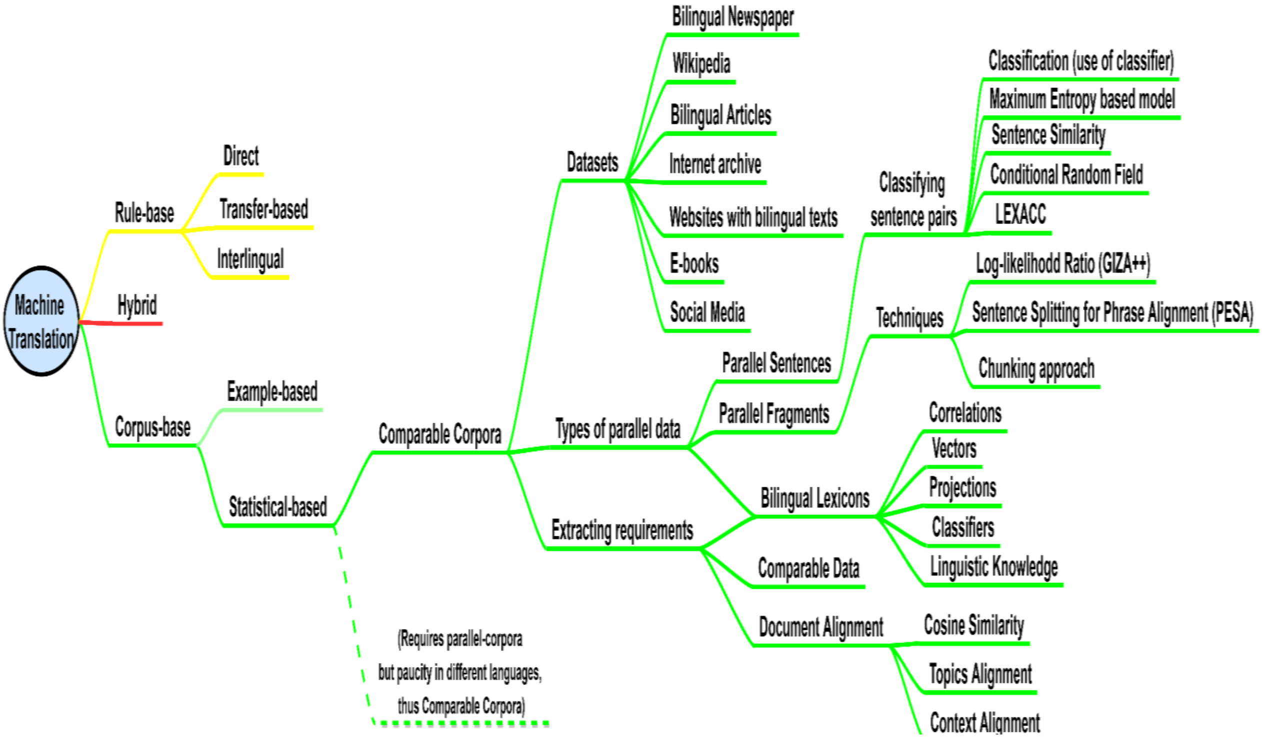

Fig. 10: Mind Map showcasing the machine translation

Conclusion

To summarize, the systematic literature survey is conducted on 188 research papers which are collected from various databases such as ACM, ACL, IEEE, Springer, ScienceDirect, etc. which are elaborated in Fig. 2. The papers were also taken from conference proceedings and workshops in the context of keywords mentioned under subheading “Vital Keywords”. From the set of 5 digital libraries, set of 1270 papers were searched. After implementing inclusion/exclusion criteria on these 1270 papers, later 188 papers were collected for writing this literature survey. The results are presented in the form of Tables, figures, pie charts, flow diagrams, mind map, bar graphs, etc., Fig. 10 presents a mind map that gives a clear picture of different aspects involved in machine translation. The contribution of different researchers in the field of parallel data mining from CC is found in this study. It seems that PC is a scarce resource. It is a major hurdle in the development of statistical machine translation for different kinds of language pairs. But there is a large amount of comparable and non-parallel corpora resources available which can be used to extract the parallel data. The work has described different kinds of “Parallel Data Extraction” such as “parallel sentence extraction”, “Parallel Fragment Extraction” and “bilingual lexicons extraction” which can be easily extracted through CC. The paper also proposed architecture for mining parallel data with the help of bilingual lexicons, fragments and sentences under "Statistical Machine Translation". Thus, it is perceived that data mined through CC can be of abundant importance in "parallel corpus" formation for language pairs with a shortage of PC resources.

Acknowledgement

The author thanks the Central University of Punjab, Bathinda for giving the opportunity of carrying out this study. I would also like to thank Dr. Satwinder Singh, HOD and Associate Professor, Computer Science and Technology, Central University of Punjab for guiding me throughout and giving valuable inputs for improving this review article.

Author's Contribution

Dilshad Kaur: Collected study material, synthesized and organized the relevant literature, analyzed and drafted the manuscript.

Dr. Satwinder Singh: Reviewed, revised the contents of manuscript, supervised the work and suggested valuable improvements in the manuscript.

Ethics

The authors confirm that this article has not been published in any other journal. The corresponding author confirms that all the authors have read and approved the review article and there are no ethical issues involved.

References

Abdul Rauf, S. and SchwenFik, H. (2011). Parallel sentence generation from comparable corpora for improved smt. (25) :341-375. doi.org/10.1007/s10590-011-9114-9

Abdul-Rauf, S., Schwenk, H., & Nawaz, M. (2017). Parallel fragments: Measuring their impact on translation performance. Computer Speech & Language, 43, 56-69. doi.org/10.1016/j.csl.2016.12.002

Adafre, S. F., & De Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. In Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources. https://aclanthology.org/W06-2810.pdf

Agic, Ž., & Vulic, I. (2020). JW300: A wide-coverage parallel corpus for low-resource languages. https://www.repository.cam.ac.uk/bitstream/handle/1810/296987/P19-1310.pdf?sequence=3

Ahmad, R., Kumar, P., Rambabu, B., Sajja, P., Sinha, M. K., & Sangal, R. (2011). Enhancing throughput of a machine translation system using mapreduce framework: An engineering approach. ICON.

Aker, A., Feng, Y., & Gaizauskas, R. (2012, December). Automatic bilingual phrase extraction from comparable corpora. In Proceedings of COLING 2012: Posters (pp. 23-32). https://aclanthology.org/C12-2003.pdf.

Ali, A., Hussain, A., & Malik, M. K. (2013). Model for english-urdu statistical machine translation. World Applied Sciences, 24, 1362-1367. doi.org/10.5829/idosi.wasj.2013.24.10.760

Ali, A., Siddiq, S., & Malik, M. K. (2010). Development of parallel corpus and english to urdu statistical machine translation. Resource, 9, 10. doi.org/10.1.1.658.7049&rep=rep1&type=pdf

Alp, N. D., & Turhan, C. (2008, April). English to Turkish Example-Based Machine Translation with Synchronous SSTC. In Fifth International Conference on Information Technology: New Generations (itng 2008) (pp. 674-679). IEEE. doi.org/10.1109/ITNG.2008.64

Ananthakrishnan, R., Kavitha, M., Jayprasad, J. H., Shekhar, R. S. C., & Bade, S. M. S. (2006). MaTra: a practical approach to fully-automatic indicative English-Hindi machine translation. In Symposium on Modeling and Shallow Parsing of Indian Languages (MSPIL'06) (pp. 1-8). http://196.1.113.56/downloads/papers/Matra.pdf

Anju, E. S., & Manoj Kumar, K. V. (2014). Malayalam to English machine translation: An EBMT system. IOSR Journal of Engineering (IOSRJEN), 4(1), 18-23. doi.org/10.9790/3021-04111823.

Archana, G. P., Jithesh, V. S., Remya, L. B., & Sherly, E. (2015, August). Building a parallel corpora: Translation issues and remedial case. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2414-2417). IEEE. doi.org/10.1109/ICACCI.2015.7275980

Ashrafi, S. S., Kabir, M. H., Anwar, M. M., & Noman, A. K. M. (2013). English to Bangla machine translation system using context-free grammars. International Journal of Computer Science Issues (IJCSI), 10(3), 144.

Azpeitia, A., Etchegoyhen, T., & Garcia, E. M. (2017, August). Weighted set-theoretic alignment of comparable sentences. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora (pp. 41-45). https://aclanthology.org/W17-2508.pdf

Babhulgaonkar, A. R., & Bharad, S. V. (2017, October). Statistical machine translation. In 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM) (pp. 62-67). IEEE. doi.org/10.1109/ICISIM.2017.8122149

Babych, B., Eberle, K., Geiß, J., Ginestí-Rosell, M., Hartley, A., Rapp, R., ... & Thomas, M. (2012, April). Design of a hybrid high quality machine translation system. In Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra) (pp. 101-112). https://aclanthology.org/W12-0114.pdf

Bahadur, P., Jain, A. K., & Chauhan, D. S. (2012). EtranS-A complete framework for English to Sanskrit machine translation. In International Journal of Advanced Computer Science and Applications (IJACSA) from International Conference and workshop on Emerging Trends in Technology. doi.org/10.14569/SpecialIssue.2012.020107.

Bai, M. H., Chen, K. J., & Chang, J. S. (2008). Improving word alignment by adjusting Chinese word segmentation. In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I. https://aclanthology.org/I08-1033.pdf

Bakhshaei, S., Khadivi, S., & Riahi, N. (2010, December). Farsi-german statistical machine translation through bridge language. In 2010 5th International Symposium on Telecommunications (pp. 557-561). IEEE. doi.org/10.1109/ISTEL.2010.5734087

Bandyopadhyay, S. (2000). ANUBAAD-the translator from English to Indian languages. In Proceedings of the 7th State Science and Technology Congress, (SSTC'00), Calcutta, India (pp. 1-9). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.8386&rep=rep1&type=pdf#page=91

Barkade, M. V. M., & Prakash, R. D. (2010). English to sanskrit machine translator lexical parser. doi.org/10.1.1.302.7550

Basavaraddi, M. C. C. S., & Shashirekha, D. H. (2014). A typical machine translation system for English to Kannada. Int J Sci Eng Res, 5(4). https://www.ijser.org/researchpaper/A-Typical-Machine-Translation-System-for-English-to-Kannada.pdf

Batra, K. K., & Lehal, G. S. (2010). Rule based machine translation of noun phrases from Punjabi to English. International Journal of Computer Science Issues (IJCSI), 7(5), 409. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.8120&rep=rep1&type=pdf#page=432

Bharadwaj, R. G., & Varma, V. (2011, March). Language independent identification of parallel sentences using wikipedia. In Proceedings of the 20th international conference companion on World wide web (pp. 11-12). doi.org/10.1145/1963192.1963199

Bharati, A., Chaitanya, V., Kulkarni, A. P., & Sangal, R. (1997). Anusaaraka: Machine translation in stages. VIVEK-BOMBAY-, 10, 22-25.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

Blunsom, P., & Cohn, T. (2006, July). Discriminative word alignment with conditional random fields. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp. 65-72). https://aclanthology.org/P06-1009.pdf

Bojar, O. (2007, June). English-to-Czech factored machine translation. In Proceedings of the second workshop on statistical machine translation (pp. 232-239). https://aclanthology.org/W07-0735.pdf

Bonelli, E. T. (2010). Theoretical overview of the evolution of corpus linguistics. The Routledge handbook of corpus linguistics, 14-28.

Bouamor, D., Popescu, A., Semmar, N., & Zweigenbaum, P. (2013a, October). Building specialized bilingual lexicons using large scale background knowledge. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 479-489). https://aclanthology.org/D13-1046.pdf

Bouamor, D., Semmar, N., & Zweigenbaum, P. (2013b, August). Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 759-764). https://aclanthology.org/P13-2133.pdf

Bouamor, H., & Sajjad, H. (2018, May). H2@ bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In Proc. Workshop on Building and Using Comparable Corpora. http://lrec-conf.org/workshops/lrec2018/W8/pdf/book_of_proceedings.pdf#page=52

Brockett, C. (2005). Support vector machines for paraphrase identification and corpus construction. Proceedings of the third International Workshop on Paraphrasing (IWP), 1–8.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational linguistics, 19(2), 263-311. https://dl.acm.org/doi/abs/10.5555/972470.972474

Callison-Burch, C., Koehn, P., & Osborne, M. (2006, June). Improved statistical machine translation using paraphrases. In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference (pp. 17-24). https://aclanthology.org/N06-1003.pdf

Cao, G., Gao, J., & Nie, J. Y. (2007). A system to mine large-scale bilingual dictionaries from monolingual web pages.

Chand, S. (2016, September). Empirical survey of machine translation tools. In 2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) (pp. 181-185). IEEE. doi.org/10.1109/ICRCICN.2016.7813653

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27. doi.org/10.1145/1961189.1961199

Chang, P. C., Galley, M., & Manning, C. D. (2008, June). Optimizing Chinese word segmentation for machine translation performance. In Proceedings of the third workshop on statistical machine translation (pp. 224-232). https://aclanthology.org/W08-0336.pdf

Chatterji, S., Sonare, P., Sarkar, S., & Basu, A. (2011). Lattice based lexical transfer in Bengali Hindi machine translation framework. In Proceedings of ICON-2011: 9th international conference on natural language processing.

Chiao, Y. C., & Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes. https://aclanthology.org/C02-2020.pdf

Chinnakotla, M. K., Ranadive, S., Damani, O. P., & Bhattacharyya, P. (2007, September). Hindi to English and Marathi to English cross language information retrieval evaluation. In Workshop of the Cross-Language Evaluation Forum for European Languages (pp. 111-118). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-540-85760-0_14

Chu, C., Nakazawa, T., & Kurohashi, S. (2013a, August). Chinese–Japanese parallel sentence extraction from quasi–comparable corpora. In Proceedings of the Sixth Workshop on Building and Using Comparable Corpora (pp. 34-42). https://aclanthology.org/W13-2505.pdf

Chu, C., Nakazawa, T., & Kurohashi, S. (2014a, April). Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 296-309). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-54903-8_25

Chu, C., Nakazawa, T., & Kurohashi, S. (2014b, May). Constructing a Chinese—Japanese Parallel Corpus from Wikipedia. In LREC (pp. 642-647). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.726.1498&rep=rep1&type=pdf

Chu, C., Nakazawa, T., Kawahara, D., & Kurohashi, S. (2012). Exploiting shared Chinese characters in Chinese word segmentation optimization for Chinese-Japanese machine translation. In Proceedings of the 16th Annual conference of the European Association for Machine Translation (pp. 35-42). https://aclanthology.org/2012.eamt-1.7.pdf

Chu, C., Nakazawa, T., Kawahara, D., & Kurohashi, S. (2013). Chinese-japanese machine translation exploiting chinese characters. ACM Transactions on Asian Language Information Processing (TALIP), 12(4), 1-25. doi.org/10.1145/2523057.2523059

Dagan, I., & Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. Computational linguistics, 20(4), 563-596. doi.org/10.5555/203987.203991

Darbari, H. (1999, September). Computer Assisted Translation System-An Indian Perspective. In Machine Translation Summit VII (pp. 80-85).

Deep, K., Kumar, A., & Goyal, V. (2018). Development of Punjabi-English (PunEng) Parallel Corpus for Machine Translation System. International Journal of Engineering & Technology, 7(2), 690-693. doi.org/10.14419/ijet.v7i2.8892

Déjean, H., Gaussier, É., & Sadat, F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In COLING 2002: The 19th International Conference on Computational Linguistics. https://aclanthology.org/C02-1166.pdf

Dhore, M. L., & Dixit, S. X. (2011). English to Devnagari translation for UI labels of commercial web based interactive applications. International Journal of Computer Applications, 35(10), 0975-8887.

Do, T. N. D., & Besacier, L. (2010). A fully unsupervised approach for mining parallel data from comparable corpora. In Proceedings of the 14th Annual conference of the European Association for Machine Translation. https://aclanthology.org/2010.eamt-1.6/

Fu, X., Wei, W., Lu, S., Chen, Z., & Xu, B. (2013, October). Phrase-based parallel fragments extraction from comparable corpora. In Proceedings of the Sixth International Joint Conference on Natural Language Processing (pp. 972-976). https://aclanthology.org/I13-1129.pdf

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In Third Workshop on Very Large Corpora. https://aclanthology.org/W95-0114.pdf

Fung, P., & Cheung, P. (2004b). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics (pp. 1051-1057). https://aclanthology.org/C04-1151.pdf

Fung, P., & Cheung, P. (2004a). Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and e. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (pp. 57-63). https://aclanthology.org/W04-3208.pdf

Fung, P., & McKeown, K. (1994). Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. arXiv preprint cmp-lg/9409011. https://arxiv.org/abs/cmp-lg/9409011

Fung, P., & Yee, L. Y. (1998, August). An IR approach for translating new words from nonparallel, comparable texts. In 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1 (pp. 414-420). https://www.aclweb.org/anthology/P98-1069.pdf

Fung, P., Prochasson, E., & Shi, S. (2010). Trillions of comparable documents. In Proceedings of the 3rd workshop on Building and Using Comparable Corpora (pp. 26-34). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.186.641&rep=rep1&type=pdf#page=34

Gahbiche-Braham, S., Bonneau-Maynard, H., & Yvon, F. (2011, June). Two ways to use a noisy parallel news corpus for improving statistical machine translation. In Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web (pp. 44-51). https://aclanthology.org/W11-1207.pdf

Gale, W. A., & Church, K. (1991). Identifying word correspondences in parallel texts. In Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991. https://aclanthology.org/H91-1026.pdf

Ganitkevitch, J., & Callison-Burch, C. (2014, May). The Multilingual Paraphrase Database. In LREC (pp. 4276-4283). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1067.1500&rep=rep1&type=pdf

Garera, N., Callison-Burch, C., & Yarowsky, D. (2009, June). Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009) (pp. 129-137). https://aclanthology.org/W09-1117.pdf

Garje, G. V., Kharate, G. K., & Kulkarni, H. (2014). Transmuter: an approach to rule-based English to Marathi machine translation. International Journal of Computer Applications, 98(21). doi.org/10.5120/17309-7782

Gaussier, E., Renders, J. M., Matveeva, I., Goutte, C., & Déjean, H. (2004, July). A geometric view on bilingual lexicon extraction from comparable corpora. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04) (pp. 526-533). https://aclanthology.org/P04-1067.pdf

Goyal, V., & Lehal, G. S. (2010). Web based Hindi to Punjabi machine translation system. journal of emerging technologies in web intelligence, 2(2), 148-151. doi.org/10.4304/jetwi.2.2.148-151

Goyal, V., & Lehal, G. S. (2011, March). Hindi to Punjabi machine translation system. In International Conference on Information Systems for Indian Languages (pp. 236-241). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-19403-0_40

Goyal, V., Kumar, A., & Lehal, M. S. (2020). Document Alignment for Generation of English-Punjabi Comparable Corpora from Wikipedia. International Journal of E-Adoption (IJEA), 12(1), 42-51. doi.org/10.4018/ijea.2020010104.

Gupta, R., Pal, S., & Bandyopadhyay, S. (2013, August). Improving mt system using extracted parallel fragments of text from comparable corpora. In Proceedings of the Sixth Workshop on Building and Using Comparable Corpora (pp. 69-76). https://www.aclweb.org/anthology/W13-2509.pdf

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., & Klein, D. (2008, June). Learning bilingual lexicons from monolingual corpora. In Proceedings of ACL-08: Hlt (pp. 771-779). https://aclanthology.org/P08-1088.pdf

Hakkani, D. Z., Tür, G., Oflazer, K., Mitamura, T., & Nyberg, E. H. (1998, October). An English-to-Turkish interlingual MT system. In Conference of the Association for Machine Translation in the Americas (pp. 83-94). Springer, Berlin, Heidelberg. doi.org/10.1007/3-540-49478-2_8

Harastani, R., Daille, B., & Morin, E. (2013, October). Ranking translation candidates acquired from comparable corpora. In Proceedings of the Sixth International Joint Conference on Natural Language Processing (pp. 401-409). https://aclanthology.org/I13-1046.pdf

Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162. doi.org/10.1080/00437956.1954.11659520.

Hazem, A., & Morin, E. (2013a, October). Word co-occurrence counts prediction for bilingual terminology extraction from comparable corpora. In Proceedings of the Sixth International Joint Conference on Natural Language Processing (pp. 1392-1400). aclweb.org/anthology/I13-1196.pdf

Hazem, A. and Morin, E. (2013b). A comparison of smoothing techniques for bilingual lexicon extraction from comparable corpora. Workshop on Building and Using Comparable Corpora: 24-33. https://aclanthology.org/W13-2504.pdf

Hewavitharana, S., & Vogel, S. (2013). Extracting parallel phrases from comparable data. In Building and using comparable corpora (pp. 191-204). Springer, Berlin, Heidelberg. https://aclanthology.org/W11-1209.pdf

Hong, G., Li, C. H., Zhou, M., & Rim, H. C. (2010, August). An empirical study on web mining of parallel data. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) (pp. 474-482). https://aclanthology.org/C10-1054.pdf

Hutchins, W. J., & Somers, H. L. (1992). An Introduction fo Machine Translation. https://www.computer.org/csdl/magazine/co/1992/12/rz118/13rRUB6SpOK

Irvine, A., & Callison-Burch, C. (2013, August). Combining bilingual and comparable corpora for low resource machine translation. In Proceedings of the eighth workshop on statistical machine translation (pp. 262-270). https://aclanthology.org/W13-2233.pdf

Isabelle, P., & Foster, G. (2006). Machine Translation: Overview. Encyclopedia of Language & Linguistics, 404–422. doi.org/ 10.1016/b0-08-044854-2/00936-6.

Iyer, S. (2015). Literature survey on comparable corpora. CFILT Resources 2015.

Jagarlamudi, J., & Kumaran, A. (2007, September). Cross-lingual information retrieval system for Indian languages. In Workshop of the Cross-Language Evaluation Forum for European Languages (pp. 80-87). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-540-85760-0_10

Jain, R., Sinha, R. M. K., & Jain, A. (2001). ANUBHARTI: using hybrid example-based approach for machine translation. STRANS-2001, IIT Kanpur, 20-32.A

Jiang, L., Yang, S., Zhou, M., Liu, X., & Zhu, Q. (2009, August). Mining bilingual data from the web with adaptively learnt patterns. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (pp. 870-878). https://aclanthology.org/P09-1098.pdf

Jindal, S., Goyal, V., & Bhullar, J. S. (2018a). Building English-Punjabi Parallel corpus for Machine Translation. International Journal of Engineering, Science and Mathematics, 7(3), 223-229. doi.org/10.5120/ijca2017916036.

Jindal, S., Goyal, V., & Bhullar, J. S. (2018b). English to Punjabi statistical machine translation using moses (Corpus Based). Journal of Statistics and Management Systems, 21(4), 553-560. doi.org/10.1080/09720510.2018.1471265

Josan, G. S., & Lehal, G. S. (2008, August). A Punjabi to Hindi machine translation system. In Coling 2008: Companion volume: Demonstrations (pp. 157-160). https://aclanthology.org/C08-3004.pdf

Kang, B. K., Chen, Y. R., Chang, B. B., & Yu, S. W. (2005). Translating multi word terms into Korean from Chinese documents. In 2005 International Conference on Natural Language Processing and Knowledge Engineering (pp. 449-454). IEEE. doi.org/10.1109/NLPKE.2005.1598779

Kaur, A., & Rani, J. (2015, December). A web based Punjabi to Hindi Statistical Machine Translation System. In 2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/7453298/

Kenning, M. M. (2010). What are parallel and comparable corpora and how can we use them. In The Routledge handbook of corpus linguistics (pp. 487-500). Routledge. doi.org/10.4324/9780203856949-35

Khan, S., & Mishra, R. B. (2011). Translation rules and ANN based model for English to Urdu machine translation. INFOCOMP Journal of Computer Science, 10(3), 36-47. http://infocomp.dcc.ufla.br/index.php/INFOCOMP/article/view/345

Khosla, S., & Acharya, H. (2018). A survey report on the existing methods of building a parallel corpus. International Journal of Advanced Research in Computer Science, 9(4). doi.org/10.26483/ijarcs.v9i4.6171.

Koehn, P. (2009). Statistical machine translation. Cambridge University Press. doi.org/10.1017/cbo9780511815829

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Herbst, E. (2007, June). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions (pp. 177-180). https://www.aclweb.org/anthology/P07-2045.pdf

Krishnamurthy, P. (2015). Development of Telugu-Tamil Transfer-Based Machine Translation system: With Special reference to Divergence Index. In Proceedings of the 1st Deep Machine Translation Workshop (pp. 48-54).

Kulkarni, R. C. (2013). Extraction of Parallel Corpora from Comparable Corpora. Department of Computer Science & Engineering, Indian Institute of Technology, India. https://www.cfilt.iitb.ac.in/resources/surveys/ComparableCorporaSurvey.pdf

Kumar, A., & Goyal, V. (2012, May). Practical approach for developing Hindi-Punjabi parallel corpus. In LREC 2012 Workshop on Indian Language and Data: Resources and Evaluation (pp. 65-69). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.433.2162&rep=rep1&type=pdf#page=73

Kumar, A., & Goyal, V. (2018). Hindi to Punjabi machine translation system based on statistical approach. Journal of Statistics and Management Systems, 21(4), 547-552. doi.org/10.1080/09720510.2018.1466963.

Kumar, P., & Goyal, V. (2010a, December). Development of Hindi-Punjabi parallel corpus using existing Hindi-Punjabi machine translation system. In Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia: 114-118. doi.org/10.1145/1963564.1963583

Kumar, P., & Goyal, V. (2010b). Development of Hindi-Punjabi parallel corpus using existing Hindi-Punjabi machine translation system and using sentence alignments. International Journal of Computer Applications, 5(9), 15-19. doi.org/10.5120/941-1319

Lakshmi, S., & Shambhavi, B. R. (2020). Extraction of Bilingual Dictionary from Comparable Corpora for Resource Scarce Languages. Journal of Computational and Theoretical Nanoscience, 17(1), 54-60. doi.org/10.1166/jctn.2020.8629

Lee, Y. S., Sinder, D. J., & Weinstein, C. J. (2002). Interlingua-based English–Korean two-way speech translation of Doctor–Patient dialogues with CCLINC. Machine Translation, 17(3), 213-243. doi.org/10.1023/b:coat.0000010801.30299.10

Lehal, M. S., Kumar, A., & Goyal, V. (2018). Review of techniques for extraction of bilingual lexicon from comparable corpora. International Journal of Engineering & Technology, 7(2.30), 15-20. doi.org/ 10.14419/ijet.v7i2.30.13456.

Lehal, M. S., Kumar, A., & Goyal, V. (2019). Comparative analysis of similarity measures for extraction of parallel data. International Journal of Control and Automation, 12(6), 408-417.

Li, B., & Gaussier, E. (2010, August). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) (pp. 644-652). https://www.aclweb.org/anthology/C10-1073.pdf

Li, B., Gaussier, E., & Aizawa, A. (2011a, June). Clustering comparable corpora for bilingual lexicon extraction. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 473-478). https://aclanthology.org/P11-2083.pdf

Li, M. H., Klyuev, V., & Wu, S. H. (2011b, September). A novel approach to sentence alignment from comparable corpora. In Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems (Vol. 2, pp. 618-623). IEEE. https://ieeexplore.ieee.org/abstract/document/6072842/

Ling, W., Xiang, G., Dyer, C., Black, A. W., & Trancoso, I. (2013, August). Microblogs as parallel corpora. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 176-186). https://www.aclweb.org/anthology/P13-1018.pdf

Liu, X., Duh, K., & Matsumoto, Y. (2013, August). Topic models+ word alignment= a flexible framework for extracting bilingual dictionary from comparable corpus. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning (pp. 212-221). https://aclanthology.org/W13-3523.pdf

Lu, B., Jiang, T., Chow, K. & Tsou, B. K. (2010). Building a large english-chinese parallel corpus from comparable patents and its experimental application to smt. *Workshop on Building and Using Comparable Corpora, LREC 2010*: 42-49.

Malik, A. A., & Habib, A. (2013). Urdu to English machine translation using bilingual evaluation understudy. International Journal of Computer Applications, 82(7). doi.org/10.5120/14126-1040.

Marton, Y., Callison-Burch, C., & Resnik, P. (2009, August). Improved statistical machine translation using monolingually-derived paraphrases. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (pp. 381-390). https://aclanthology.org/D09-1040.pdf

Maskara, S. & Bhattacharyya, P. (2018). Recent works on Parallel Sentence Extraction from Comparable Corpora. *CFILT Resources 2018.*

McCarthy, M., & O'Keeffe, A. (2010). Historical perspective: What are corpora and how have they evolved?. In The Routledge handbook of corpus linguistics (pp. 3-13). Routledge. doi.org/10.4324/9780203856949

Mishra, V., & Mishra, R. B. (2010). ANN and Rule based model for English to Sanskrit Machine Translation. INFOCOMP Journal of Computer Science, 9(1), 80-89. http://infocomp.dcc.ufla.br/index.php/infocomp/article/view/294

Mohammadi, M., & Ghasem Aghaee, N. (2010, March). Building bilingual parallel corpora based on wikipedia. In 2010 Second International Conference on Computer Engineering and Applications (Vol. 2, pp. 264-268). IEEE. https://ieeexplore.ieee.org/abstract/document/5445653/

Motazedi, Y., & Shamsfard, M. (2009, October). English to persian machine translation exploiting semantic word sense disambiguation. In 2009 14th International CSI Computer Conference (pp. 253-258). IEEE. https://ieeexplore.ieee.org/abstract/document/5349401/

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics* 31(4):477-504. doi.org/10.1162/089120105775299168.

Munteanu, D. S., & Marcu, D. (2006, July). Extracting parallel sub-sentential fragments from non-parallel corpora. In Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics (pp. 81-88). https://www.aclweb.org/anthology/P06-1011.pdf

Munteanu, D. S., Fraser, A., & Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004 (pp. 265-272). https://aclanthology.org/N04-1034.pdf

Nie, J. Y., Simard, M., Isabelle, P., & Durand, R. (1999, August). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 74-81). https://dl.acm.org/doi/abs/10.1145/312624.312656

Ning, S., Yan, X., Nuo, Y., Zhou, F., Xie, Q., & Zhang, J. P. (2020). Chinese-Khmer Parallel fragments Extraction from Comparable Corpus Based on Dirichlet Process. Procedia Computer Science, 166, 213-221. doi.org/10.1016/j.procs.2020.02.049

Nirenburg, S., & Wilks, Y. (2000). Machine translation. Fortieth Anniversary Volume: Advancing into the 21st Century, 159–188. doi.org/10.1016/s0065-2458(00)80018-2

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics, 29(1), 19-51. doi.org/10.1162/089120103321337421.

Otero, P. G. (2007). Learning bilingual lexicons from comparable English and Spanish corpora. Proceedings of MT Summit XI, 191-198. http://gramatica.usc.es/~gamallo/artigos-web/gamalloSUMMIT2007.pdf

Padhya, D., & Sheth, J. (2019). A review of machine translation systems for Indian languages and their approaches. Emerging Trends in Expert Applications and Security, 103-110. doi.org/10.1007/978-981-13-2285-3_13

Pal, S., Lohar, P., & Naskar, S. K. (2014, April). Role of paraphrases in pb-smt. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 242-253). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-54903-8_21

Pekar, V., Mitkov, R., Blagoev, D., & Mulloni, A. (2006). Finding translations for low-frequency words in comparable corpora. Machine Translation, 20(4), 247-266. doi.org/10.1007/s10590-007-9029-7

Pirkola, A., Hedlund, T., Keskustalo, H., & Järvelin, K. (2001). Dictionary-based cross-language information retrieval: Problems, methods and research findings. Information retrieval, 4(3), 209-230. doi.org/10.1023/a:1011994105352

Pisharoty, D., Sidhaye, P., Utpat, H., Wandkar, S., & Sugandhi, R. (2012). Extending capabilities of English to Marathi machine translator. International Journal of Computer Science Issues (IJCSI), 9(3), 375. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.401.9953&rep=rep1&type=pdf

Post, M., Callison-Burch, C., & Osborne, M. (2012, June). Constructing parallel corpora for six indian languages via crowdsourcing. In Proceedings of the Seventh Workshop on Statistical Machine Translation (pp. 401-409). https://www.aclweb.org/anthology/W12-3152.pdf

Premjith, B., Kumar, M. A., & Soman, K. P. (2019). Neural Machine Translation System for English to Indian Language Translation Using MTIL Parallel Corpus. Journal of Intelligent Systems, 28(3), 387-398. doi.org/10.1515/jisys-2019-2510.

Prochasson, E., & Fung, P. (2011, June). Rare word translation extraction from aligned comparable documents. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 1327-1335). https://aclanthology.org/P11-1133.pdf

Qi, X., Zhou, H., & Chen, H. (2002). An interlingua-based Chinese-English MT system. Journal of Computer Science and Technology, 17(4), 464-472. doi.org/10.1007/bf02943286.

Qian, L., Wang, H., Zhou, G., & Zhu, Q. (2012, December). Bilingual lexicon construction from comparable corpora via dependency mapping. In Proceedings of COLING 2012 (pp. 2275-2290). https://www.aclweb.org/anthology/C12-1139.pdf

Quirk, C., Udupa, R., & Menezes, A. (2007). Generative models of noisy translations with applications to parallel fragment extraction. Proceedings of the Machine Translation Summit XI, 377-384. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/mt-mtsummit2007_compcorp.pdf

Rahimi, R., Shakery, A., & King, I. (2016). Extracting translations from comparable corpora for Cross-Language Information Retrieval using the language modeling framework. Information Processing & Management, 52(2), 299-318. doi.org/10.1016/j.ipm.2015.08.001.

Raju, B. N., & Raju, M. B. (2016, February). Statistical Machine Translation System for Indian Languages. In 2016 IEEE 6th International Conference on Advanced Computing (IACC) (pp. 174-177). IEEE.

Rapp, R. (1995). Identifying word translations in non-parallel texts. arXiv preprint cmp-lg/9505037. https://arxiv.org/abs/cmp-lg/9505037

Rapp, R. (1999, June). Automatic identification of word translations from unrelated English and German corpora. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics (pp. 519-526). https://www.aclweb.org/anthology/P99-1067.pdf

Rathod, S. G. (2014). Machine translation of natural language using different approaches. International journal of computer applications, 102(15). doi.org/ 10.5120/17893-8899.

Rathod, S. G., & Sondur, S. (2012). English to sanskrit translator and synthesizer (etsts). International Journal of Emerging Technology and Advanced Engineering, 2(12), 379-383. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.414.2272&rep=rep1&type=pdf

Reddy, M. V., & Hanumanthappa, M. (2011). English to Kannada/Telugu name transliteration in Clir: a statistical approach. International Journal of Machine Intelligence, 3(4). doi.org/10.1162/089120103322711578.

Resnik, P and Smith, N.A. (2003). The web as a parallel corpus. Computational Linguistics, 29(3): 349-380. doi.org/10.1162/089120103322711578

Richardson, J., Nakazawa, T., & Kurohashi, S. (2013, October). Robust transliteration mining from comparable corpora with bilingual topic models. In Proceedings of the Sixth International Joint Conference on Natural Language Processing (pp. 261-269). https://aclanthology.org/I13-1030.pdf

Saini, S., & Sahula, V. (2015, February). A survey of machine translation techniques and systems for Indian languages. In 2015 IEEE International Conference on Computational Intelligence & Communication Technology (pp. 676-681). IEEE. https://ieeexplore.ieee.org/abstract/document/7078789/

Sangal, R. (2004). Architecture of shakti machine translation system. IIIT Hyderabad.

Sankaran, B., Razmara, M., & Sarkar, A. (2012). Kriya-an end-to-end hierarchical phrase-based mt system. The Prague Bulletin of Mathematical Linguistics, 97, 83. doi.org/10.2478/v10108-012-0004-y.

Sebastian, M. P., & Kumar, G. S. (2010). English to Malayalam translation: a statistical approach. In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India (pp. 1-5).

Sindhu, D. V., & Sagar, B. M. (2016, December). Study on machine translation approaches for Indian languages and their challenges. In 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT) (pp. 262-267). IEEE. https://ieeexplore.ieee.org/abstract/document/7955227/

Singh, S., & Kaur, S. (2018). A systematic literature review: Refactoring for disclosing code smells in object oriented software. Ain Shams Engineering Journal, 9(4), 2129-2151. doi.org/ 10.1016/j.asej.2017.03.002

Sinha, R. M. K., & Jain, A. (2003). AnglaHindi: an English to Hindi machine-aided translation system. MT Summit IX, New Orleans, USA, 494-497. https://aclanthology.org/2003.mtsummit-systems.15.pdf

Sinha, R. M. K., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R., & Jain, A. (1995, October). ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. In 1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century (Vol. 2, pp. 1609-1614). IEEE. https://ieeexplore.ieee.org/abstract/document/538002/

Sinhal, R. A., & Gupta, K. O. (2014). A pure EBMT approach for English to Hindi sentence translation system. International Journal of Modern Education and Computer Science, 6(7), 1. doi.org/10.5815/ijmecs.2014.07.01.

Smith, J., Quirk, C., & Toutanova, K. (2010, June). Extracting parallel sentences from comparable corpora using document level alignment. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 403-411). https://www.aclweb.org/anthology/N10-1063.pdf

Snover, M., Dorr, B., & Schwartz, R. (2008, October). Language and translation model adaptation using comparable corpora. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (pp. 857-866). https://aclanthology.org/D08-1090.pdf

Srivastava, J., & Sanyal, S. (2012, November). A hybrid approach for word alignment in english-hindi parallel corpora with scarce resources. In 2012 International Conference on Asian Language Processing (pp. 185-188). IEEE. https://ieeexplore.ieee.org/abstract/document/6473727/

Srivastava, R., & Bhat, R. A. (2013, November). Transliteration systems across indian languages using parallel corpora. In Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation (PACLIC 27) (pp. 390-398). https://aclanthology.org/Y13-1040.pdf

Ştefănescu, D., & Ion, R. (2013). Parallel-Wiki: A collection of parallel sentences extracted from Wikipedia. In Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2013) (pp. 24-30).

Stefanescu, D., Ion, R., & Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In Proceedings of the 16th Annual conference of the European Association for Machine Translation (pp. 137-144). https://aclanthology.org/2012.eamt-1.37.pdf

Sun, X., Ren, F., & Huang, D. (2009, September). Extended super function based Chinese Japanese machine translation. In 2009 International Conference on Natural Language Processing and Knowledge Engineering (pp. 1-8). IEEE. https://ieeexplore.ieee.org/abstract/document/5313817/

Tamura, A., Watanabe, T., & Sumita, E. (2012, July). Bilingual lexicon extraction from comparable corpora using label propagation. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 24-36). https://aclanthology.org/D12-1003.pdf

Tan, B., & Zhou, X. Y. (2010, April). Automatic construction of web-based English/Chinese parallel corpora. In 2010 Third International Symposium on Intelligent Information Technology and Security Informatics (pp. 114-117). IEEE. https://ieeexplore.ieee.org/abstract/document/5453637/

Terumasa, E. (2007, September). Rule based machine translation combined with statistical post editor for japanese to english patent translation. In Proceedings of the MT Summit XI Workshop on Patent Translation (Vol. 11, pp. 13-18). sn. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.597.6991&rep=rep1&type=pdf

Tillmann, C. (2009, August). A beam-search extraction algorithm for comparable data. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 225-228). https://www.aclweb.org/anthology/P09-2057.pdf

Tripathi, S., & Sarkhel, J. K. (2010). Approaches to machine translation. ISSN: 0975-2404. http://nopr.niscair.res.in/handle/123456789/11057.

Tognini Bonelli, E. and Sinclair, J. (2006) 'Corpora', in K. Brown (ed.) Encyclopedia of Language and Linguistics, second edition. Amsterdam: Elsevier, pp. 206–19. doi.org/10.1016/b0-08-044854-2/00940-8

Upadhyay, P., Jaiswal, U. C., & Ashish, K. (2014). Transish: Translator from sanskrit to english-a rule based machine translation. International Journal of Current Engineering and Technology E-ISSN, 2277-4106.

Uszkoreit, J., Ponte, J., Popat, A., & Dubiner, M. (2010, August). Large scale parallel document mining for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) (pp. 1101-1109). https://aclanthology.org/C10-1124.pdf

Utiyama, M., & Isahara, H. (2003, July). Reliable measures for aligning Japanese-English news articles and sentences. In Proceedings of the 41st annual meeting of the association for computational linguistics (pp. 72-79). https://www.aclweb.org/anthology/P03-1010.pdf

Véronis, J. (2000). From the Rosetta stone to the information society. In Parallel text processing (pp. 1-24). Springer, Dordrecht. doi.org/10.1007/978-94-017-2535-4_1.

Vijayanand, K., Choudhury, S. I., & Ratna, P. (2002, December). Vaasaanubaada: automatic machine translation of bilingual Bengali-Assamese news texts. In Language Engineering Conference, 2002. Proceedings (pp. 183-188). IEEE. https://ieeexplore.ieee.org/abstract/document/1182307/

Vulic, I., & Moens, M. F. (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012) (pp. 449-459). ACL; East Stroudsburg, PA. https://lirias.kuleuven.be/1572200?limo=0

Vulic, I., De Smet, W., & Moens, M. F. (2011). Identifying word translations from comparable corpora using latent topic models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011) (Vol. 2, pp. 479-484). ACL; East Stroudsburg, PA. https://lirias.kuleuven.be/1572215?limo=0

Warhade, S. R., Patil, S. H., & Devale, P. R. (2012). English-to-sanskrit statistical machine translation with ubiquitous application. International Journal of Computer Applications, 51(1). doi.org/10.5120/8009-1374.

Widdows, D., Dorow, B., & Chan, C. K. (2002, May). Using Parallel Corpora to enrich Multilingual Lexical Resources. In LREC (pp. 240-245). https://muchmore.dfki.de/pubs/bilingual-terms-widdows.pdf

Winiwarter, W. (2007, January). Automatic acquisition of translation knowledge using structural matching between parse trees. In First International Conference on the Digital Society (ICDS'07) (pp. 10-10). IEEE. https://ieeexplore.ieee.org/abstract/document/4063771/

Wołk, K., & Marasek, K. (2014). Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. Procedia Technology, 18, 126-132. doi.org/10.1016/j.protcy.2014.11.024.

Xu, H., Liu, D., Qian, L., & Zhou, G. (2011, November). Improving Bilingual Lexicon Construction from Chinese-English Comparable Corpora via Dependency Relationship Mapping. In 2011 International Conference on Asian Language Processing (pp. 169-172). IEEE. doi.org/10.1109/IALP.2011.22

Yang, C. C., & Li, K. W. (2004). Building parallel corpora by automatic title alignment using length-based and text-based approaches. Information processing & management, 40(6), 939-955. doi.org/10.1016/j.ipm.2003.11.002

Yu, K., & Tsujii, J. I. (2009, June). Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers (pp. 121-124). https://aclanthology.org/N09-2031.pdf

Zesch, T., & Gurevych, I. (2010). Wisdom of crowds versus wisdom of linguists–measuring the semantic relatedness of words. Natural Language Engineering, 16(1), 25-59. doi.org/10.1017/s1351324909990167

Zhang, Y., Wu, K., Gao, J., & Vines, P. (2006, April). Automatic acquisition of Chinese–English parallel corpus from the web. In European Conference on Information Retrieval (pp. 420-431). Springer, Berlin, Heidelberg. doi.org/10.1007/11735106_37

Zhang, Z., & Zweigenbaum, P. (2017, August). zNLP: Identifying parallel sentences in Chinese-English comparable corpora. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora (pp. 51-55). https://www.aclweb.org/anthology/W17-2510.pdf

Zhao, B., & Vogel, S. (2002, December). Adaptive parallel sentences mining from web bilingual news collection. In 2002 IEEE International Conference on Data Mining, 2002. Proceedings. (pp. 745-748). IEEE. https://ieeexplore.ieee.org/abstract/document/1184044/

Zhu, Z., Li, M., Chen, L., & Yang, Z. (2013, August). Building comparable corpora based on bilingual lda model. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 278-282). https://aclanthology.org/P13-2050.pdf

Zhu, Z., Li, M., Chen, L., & Zeng, W. (2012, November). Automatically Mining Parallel Corpora for Minority Languages from Web Pages. In 2012 International Conference on Asian Language Processing (pp. 97-100). IEEE. https://ieeexplore.ieee.org/abstract/document/6473705/

Zhu, Z., Li, M., Chen, L., & Zheng, S. (2011, November). Automatic Construction of Chinese-Mongolian Parallel Corpora from the Web Based on the New Heuristic Information. In 2011 International Conference on Asian Language Processing (pp. 264-267). IEEE. https://ieeexplore.ieee.org/abstract/document/6121517/