A SURVEY REPORT ON THE EXISTING METHODS OF BUILDING A PARALLEL CORPUS

Article · August 2018

DOI: 10.26483/ijarcs.v9i4.6171

1 author:

Sonal Khosla
Symbiosis International University
3 PUBLICATIONS **3** CITATIONS

SEE PROFILE

DOI: http://dx.doi.org/10.26483/ijarcs.v9i4.6171

ISSN No. 0976-5697

Volume 9, No. 4, July – August 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A SURVEY REPORT ON THE EXISTING METHODS OF BUILDING A PARALLEL CORPUS

Sonal Khosla
Symbiosis International (Deemed University)
Pune, India

Haridasa Acharya
Symbiosis International (Deemed University)
Pune, India

Abstract: This paper is a survey of the existing methods of building a parallel Corpus. The paper starts with a short introduction to a parallel corpus followed by the applications of a parallel corpus. Parallel corpus built in different language pairs and the method adopted is discussed and presented. The paper covers some of the methodologies of the major parallel corpus built. The survey report is restricted to corpus built aligned at sentence and document level.

Keywords: Sentence Alignment; Web Mining; Parallel Corpus; Manual; Corpus

I. INTRODUCTION

Natural language applications which are built to run on computers have tremendous importance and requires the support of a good corpus. A corpus is a representative subset of a language having an optimum and adequate size so that it is useful for any linguistic analysis. If a Corpus contains content from a single language, it is known as a Monolingual Corpus, while a Corpus containing text in two languages is known as a bilingual Corpus. A Bilingual Corpus is either comparable or parallel.

A comparable corpus contains comparable text in two languages that talks about the similar concept in source and target language. The text is collected from the same sampling frame, similar meaning and same domain, but are not exact translations of each other [1]. While a parallel corpus is a bilingual corpus that contains text in source language with its equivalent translation in the target language [2].

Literature shows that there have been no attempt to define architecture and scope of parallel corpora. Given that such is the status a comparison of methods of building corpora becomes a difficult task. In general a corpus is simply a collection of documents as source with free structure defined by the agency building the corpus and dictated by the context for which the corpus is being developed.
Not very extensive literature is available for building a parallel Corpus. Though parallel corpus for different language pairs are in abundance. The paper is an attempt to bring together the different approaches in building a parallel corpus here. The paper also discusses the applications of a parallel corpus.

II. NEED OF PARALLEL CORPUS

This section discusses the various applications and need of a parallel corpus.

Development of Computations tools

Computational tools like frequency counting systems, text summarization systems, question answering systems corpus-based language processing tools like lexical Collocator, Concordancer, local word grouper, text annotator, text editor, word tagger, sentence parser, lemmatiser, spell-checker, grammar checker, morphological parsing, word processing systems etc. are dependent on the availability of language resources [3] [4]. Tools like translation memories, machine translation systems, bilingual concordances and word processors require parallel corpus [5].

Study of Linguistic Features

A word in a language can be used in different contexts and with different meanings. A very simple example would be the usage of the word “bank” which may mean a financial bank or a river bank. Studying a word with its neighbouring words helps in disambiguating its meaning and making sense. Hence, multilingual text helps to study the change in the usage of a word from one language to another, which can be used in knowledge extraction and in the study of the patterns of language use. It can also help to study the frequency distribution in the original and the translated texts [2]. These are helpful to the educators and linguists in language teaching [6]. Corpora tagged with extra linguistic information can be a useful resource for theoretical linguistic analysis including the study of language, change in the language and its semantics, polysemy study, ambiguity study etc. as well as for designing sophisticated language processing tools like morphological processors, sentence parsers, information retrieval systems, machine translation systems, etc [4]. Another usage of multilingual texts is the comparative study of languages [7].

Machine Translation

One of the very important application of language resources is Machine translation which has gained importance over the last few years. In the 1990’s Mona Baker was the first scholar to apply corpus for translation studies [2]. Machine Translation systems are either Rule Based or Statistical [8]. Statistical MT systems are dependent on the availability and quality of a large parallel corpus [9][10] [11] that is a good representative of the language and other lexical resources like Dictionary [1][12]. During translation, a word is

replaced with its possible translation from a dictionary obtained during training of the system or through external dictionary. For learning a dictionary during the training, a large parallel corpus is required. Linguistic resources like dictionaries are helpful in improving the speed and quality of the translation process [13] [4].

Transliteration

Texts like names of places, persons etc. which do not require translation from one language to another are simply converted from one language to another and the process is known as transliteration. Training a transliteration system also requires sufficient amount of data [10]. Most of the transliteration systems learn from the training data, identifies the named entities in one language and transliterate in the other language [14]. A Machine transliteration has further usage in machine translation, question-answering systems, cross lingual information retrieval systems etc [14].

Cross Language Information Retrieval

The ability to retrieve information in a language by making a query in another language is referred to as Cross Language Information Retrieval [15]. Commonly used approaches of CLIR are:

- Document translation
- Query translation.

Due to its large resource requirements, the preferred approach is Query translation over document translation, since it does not require translation of the whole document. Even the Query translation approach is dependent on the availability of parallel texts and machine translation systems [15]. Bilingual dictionaries and morphological analyzers can be used to build cross language information retrieval systems [15][16].

Language Teaching

The corpora developed for local languages is a very handy tool for teaching languages in classrooms and gives better results than "intuition controlled classroom teaching" [6]. Students are able to learn the actual categories and proper contextual use of various lexical items. Also it enables to learn the appropriate use of words depending on its context, know the various types of sentences and other linguistic features. It thus helps in improving the linguistic skills of the language learners. These corpora can be further used as help in writing articles, books revising written texts etc. [4]. Another application of corpora in language teaching is teacher development activities, language testing, inter-language analysis [17].

A parallel corpus can be of great help to researchers in the areas such as statistical machine translation, cross lingual information retrieval and Bilingual lexicography. A sentence aligned parallel corpora is a major linguistic resource for statistical machine translation. Ref [18] list a number of potential uses of parallel corpora in NLP.

A lot of literature shows a parallel corpus been built manually or by human translations. The translations obtained through this approach has to deal with various issues. The translators can provide multiple alternate translations for each sentence. Also the chances of occurrence of human errors is more. Even if built manually, a corpus must have a structure and no agreed structure seems to be in existence.

III. PARALLEL CORPUS BUILDING APPROACHES

The Parallel corpus building approaches have been broadly divided into four types discussed in this section.

A. Sentence alignment Approach

First method would be based on deriving a parallel corpus from a Comparable Corpora through alignment of parallel sentences. The first automatic parallel text alignment was attempted by [19], which is based on the idea that long sentences will be translated into long sentences and short sentences into short ones. Their approach works remarkably well on language pairs with high length correlation, such as French and English. Alignment performance degrades when the length correlation breaks down, such as in the case of Chinese and English [20]. Even the Gale-Church algorithm may fail at regions that contain several sentences with similar lengths for language pairs with high length correlation [5].

In a parallel bilingual corpus, the correct alignment of the various textual elements (i.e. paragraphs, sentences, phrases, words) is an essential job for statistical alignment using the principle of probability theory and using probabilities estimated from the contents of the corpus. A number of different approaches to sentence alignment have been adopted such as sentence length, word co-occurrence, cognates, use of dictionaries and parts of speech etc. to produce a parallel bilingual corpus [21].

Ref [22] has built a sentence and word aligned parallel corpus for English-Inuktitut by taking 155 documents from the transcribed proceedings of 155 days from the Nunavut Legislative Assembly already in electronic format. The sentence length approach is adopted to find parallel sentences from these 155 documents by using Anchor words with a dynamic programming search to find the path with the largest number of alignments. A precision of 91.4% and recall rate of 92.3% was achieved.

Instead of the Standard length based approach, ref [23] has used a new alignment approach based on time overlaps to mine a parallel corpus for different language pairs. The methodology has been applied in discovering a corpus of 23,000 pairs of aligned subtitles covering about 2,700 movies in 29 languages. The entire database of about 308,000 files from http://www.opensubtitles.org, a free on-line collection of movie subtitles in many languages has been downloaded. The pre-processing and cleaning of these subtitles was done and then converted into XML based Corpus files after sentence alignment.

Ref [24] has proposed an automatic method of building a Chinese-Hungarian dictionary from a Chinese-English and English-Hungarian dictionary. A parallel corpus of 60 literatures and religion books are collected manually from the Web and pre-processing done on them. After a pre-processed version of the books are obtained, automatic method of sentence alignment is done on them to obtain a sentence aligned parallel corpus and a word aligned dictionary for the given pair of languages.

Ref [25] have done alignment of parallel corpora at document, sentence and vocabulary levels on English, Spanish and French. The approach adopted in language independent. The basic requirement of this approach is a set of documents written in two languages. Following steps are followed: separating the set of documents in the two languages, aligning the documents by identifying equivalent translations, aligning the sentences in the pair of translated documents and generating a bilingual vocabulary by word alignment.

Other similar work by different authors are also done. Ref [26] have built a parallel corpus of movie subtitles for Tehran English-Persian Parallel Corpus by aligning the sentences in the files using the sentence length approach. Ref [5] has built an English-Manipuri parallel corpus by aligning sentences in a comparable corpora and tested its efficiency in a MT system. A basic requirement in all these methods is the availability of comparable corpora in the desired pair of languages. Moreover, if the comparable corpora selected for building a parallel corpora are not good enough or not fully comparable, then the results will vary [5]. Mining translations in them is very challenging. And often the parallel fragments contain non parallel fragments at the beginning or end.

B. Web Mining Approach

Another method of building a parallel corpus is through mining of parallel content from the Web. Bilingual Websites are identified which may or may not contain parallel content. Some of the work has been discussed here. Seed words are used for crawling of web content, after which the retrieved pages are cleaned, tokenized and loaded into corpus query tools. The accuracy of such corpora cannot be judged and most often depends on the use it will be put to.

Ref [27] have built an English-Chinese parallel corpus that is aligned at document and sentence level. The parallel data is taken from bilingual Websites containing good quality content in the two language from multiple domains through web crawling. The process starts from document alignment till sentence boundary detection. The HTML documents are parsed and aligned. The alignment step is divided into two stages: document alignment and sentence alignment with an intermediate step of sentence boundary detection. The complete process is automated and manual alignment process was applied at later stages with the help of human translation. Although 15 million sentences were there in the corpus, but only 2 million of them were only released that were proofread by human translators. To avoid any copyright issues, anonymous operations were carried out manually to remove any proper names with certain placeholders.

Ref [28] have proposed a method of extracting parallel sentences from the Wikipedia by identifying the cross lingual links in Wikipedia and using them to generate feature vectors for each sentence pair. These feature vectors are then classified into parallel or non-parallel links. The link structure and metadata of the Wikipedia article is used to identify the parallel sentences. It is a scalable and language independent approach that can be extended to any pair of 272 languages available on the Wikipedia. The approach is based and dependent on the existence of cross lingual links between articles in different languages on the same topics. Even though Wikipedia links exist, but the content on the two links in the two languages are not same and has lot of noise. The cross lingual links between the parallel Wikipedia articles are studied to extract various statistics which are further used to generate feature vectors for each sentence pair. These feature vectors are used to classify the sentences to parallel or non-parallel and is able to achieve an accuracy of 78%. Sentence similarity is achieved using their semantics rather than syntax.

Ref [29] has built a Chinese-English parallel corpus of 160K sentence pairs by harvesting comparable patents from the Web. Initially, more than 22 million bilingual sentence pairs have been mined, out of which 7 million high quality parallel sentences have been considered.

Ref [30] have built a Spanish-Portugese Parallel Corpus aligned at sentence and word levels. The initial parallel data is manually collected from a Websource. The TCA (Translation corpus aligner) has been implemented to obtain the set of alignments which consist of m-to-n relations marked by XML tags. Manual adjustments were made wherever needed to generate a 1-1 alignment. Two versions of the corpus are built, one represents the aligned corpus in its original format, with capital letters, punctuation marks and alignment tags (XML) and the other in the GIZA++ format. The corpus consists of 17,681 sentence pairs.

Ref [31] has built a Parallel corpus by Web crawling for English-Spanish. First bilingual websites are identified manually or through search engine. Second step is the alignment of document pages. A total of 2039272 bilingual sentence pairs were extracted through sentence alignment in the bilingual web link pairs obtained through web crawling. Acquiring parallel text from the Web is a challenging problem, through monolingual approaches have been well established [31]. The problem with this approach is that quality and scale of parallel data is dependent on the initial pairs of the bilingual websites and are difficult to obtain on a large scale of low resource languages like Hindi and Marathi. The retrieval error is carried forward in the next stage of sentence alignment. Even when the manual approach was adopted the parallel links on the websites [31] did not contain parallel data. The translation were missing or were incorrect. Hence this approach is not feasible for Hindi-Marathi.

C. Manual Approach

The third method of building a parallel corpus are manual and time consuming. Ref [32] have built a parallel bilingual syntactically annotated corpus of Czech-English as part of a Project carried out at the Charles University. The target text is obtained manually by translating an existing monolingual syntactically annotated corpus, thereby reducing the annotation efforts to annotation of a text in a single language. Annotation is done at three levels: morphological layer (lowest), analytic layer (middle) and superficial semantic layer (highest). Dependency parse trees are created concentrated around the verb.

Ref [33] have built a Japanese – Chinese parallel corpus doing human translation of Japanese text into Chinese. The Corpus is morphologically and syntactically annotated with alignments at word and phrase levels. Japanese sentences from a newspaper were collected and manually obtained the Chinese sentences. Word alignment is done through GIZA++. Syntactic annotation is done with the help of tools.

Another parallel corpus built through the manual approach is by [34]. The alignment is done automatically with manual verification and POS tagging. Manual the texts in both the languages are collected from internet from different genres and domains in reasonable proportions. The first annotation done is sentence alignment. Textual structural annotation of paragraph, word and sentence have been annotated in the corpus. After alignment, human verification is done and errors are removed [34].

Ref [35] have built a Swedish Turkish Parallel Corpus from texts collected manually from fiction and technical documents. The corpus is passed through the complete Corpus Annotation procedure to obtain a parallel Corpus in XML Corpus Encoding Standard. The BLARK (Basic Language Resource Toolkit) has been used for building and annotating. The tool UPLUG tool has been used to automatically link sentences and word alignment of the source and target texts.

A bidirectional parallel corpora, COMPARA has been created for Portugese and English by [36]. The text excerpts have been taken from the beginning, middle and end of books. Digitized versions of the texts were obtained by removing extra linguistic elements like page nos, figures etc.The corpus is encoded with the IMS Workbench format. Alignment of a particular word with multiple words in the other languages. The text in both the languages were annotated with the parsers available. The corpus consists of 3 million words from 72 source texts and 75 translations. The complete method of building the corpus is manual. The corpus is encoded with the IMS corpus workbench format. The focus has been more on identifying the decisions to be made in building the corpus and highlighting the aspects of corpus compilation that are unique to a parallel corpora. Unavailability of language resources limits most of the existing approaches for sentence similarity.

ILCI (Indian Language Corpora Initiative) handled by a consortia of different participating institutions in India has built many monolingual and parallel corpora for various Indian languages. The parallel corpora is built by collecting sentence in source language and translated by human translators to the target language. Data has been taken manually from promotional materials, published and distributed by Government and/or private institutions/agencies. The sentences are POS tagged by the Bureau of Indian Standards (BIS) tagset. A Health text Parallel Corpus of 25000 sentences for Hindi-Marathi is also freely available for research purposes [37]. As in the year 2011, this corpora was not publicly available. But now it is available for download for research purposes.

Ref [38] have built a parallel corpus for different languages centered around Czech with a focus on texts related to fiction. The parallel text is manually corrected for avoiding any errors and passed through the process of pre-processing, tokenization and morphological processing. Linguistic Markup is only done for the languages that have readily available tools.

Ref [9] have used Amazon's Mechanical Turk to build parallel corpus for six Indian languages: Bengali, Hindi, Malayalam, Tamil, Telugu and Urdu. The source data is the 100 most viewed documents on Wikipedia for each language and the translations are obtained through human translators. The parallel corpus thus obtained is used to build translation models and accuracy evaluated for each translation model for each language pair. The first step was to build the bilingual dictionaries. These dictionaries were then used to build glosses of the source sentence and then compared to the manual translations obtained. The method adopted by them had the prerequisite of availability of electronic text and a lot of human effort and intervention was required.

Ref [39] have built an Arabic-Spanish-English parallel Corpus. The entire process has been done in two stages: basic processing (tokenization and segmentation) and alignment. The corpus has been annotated with POS tags and encoded in the Corpus Encoding Standard TMX (Translation Memory Exchange). Three monolingual corpora in each language is taken for aligning of sentences to produce a sentence aligned parallel corpus.
Ref [11] have attempted to build a parallel corpus for English to Urdu Statistical Machine Translation. Similar to Marathi, morpho syntactic synthesis is done from Urdu to Hindi, as in the case of Hindi, the Postpositions exist separately. The parallel corpus was collected and sentence alignment was done manually with the help of a tool developed during the study.

D. Machine Translation Approach

Ref [40] has built a parallel corpus for multiple languages (English-German, English-Spanish, English-Czech) by taking source text and obtaining target text through five MT systems (Joshua, Lucy, Metis, Apertium, MaTrEx). The parallel corpus thus obtained is annotated with meta-information. The training of the MT systems is done on a separate dataset. It was concluded by them that the various MT systems perform complementary to each other. The size of the Corpus built is 2051 sentences translated by five different MT systems in six translation directions and annotated with various metadata information provided by the translation model. The corpus is further annotated with linguistically motivated information. XLIFF (XML Localisation Interchange File Format) is used to represent and store the corpus data.

Ref [41] has proposed a standard pipeline to provide uniform linguistic annotation to corpus resources using state of the art NLP technologies so as to obtain uniformly annotated data. The proposed pipeline provides a framework to create annotated corpora that can be used to compare and analyze different approaches to MT.

Most of the methods of building a parallel corpus are dependent on the availability of electronic text in both the languages and then through different approaches the parallel

text is extracted. Other methods is manual and requires a lot of human intervention. It is clear from the work of earlier researchers that there is still no standard method of building a parallel corpus.

Table 1: Summary of existing methods of building a Parallel Corpus

		Input	Output	Language Pair
Sentence Alignment				
1	Martin, Johnson, Farley and Maclachlan (2003)	Parallel Documents	Sentence and Word aligned Parallel Corpus (text)	English-Inuktitut
2	Tiedemann (2007)	Parallel Documents	Sentence aligned (XML)	Multiple languages
3	Pilevar, Failli and Pilevar (2011)	Parallel Documents	Sentence aligned (txt)	English-Persian
4	Singh (2012)	Comparable Corpora	Sentence aligned (txt)	English-Manipuri
5	Liu (2013)	Parallel Documents	Sentence aligned (txt)	
Web Mining				
6	Tian, Wong, Chao, Quaresma and Oliveira (2014)	Bilingual Websites	Document and Sentence aligned (txt)	English-Chinese
7	Sridhar et.al. (2011)	Bilingual Websites	Sentence aligned (txt)	English-Spanish
8	Bharadwaj and Verma (2011)	Wikipedia	Sentence aligned (txt)	
9	Bin, Jiang, Chow and Benjamin (2010)	Comparable Data from the Web	Sentence aligned	Chinese-English
10	Aziz, Pardo and Paraboni (2008)		Word and Sentence aligned (XML)	Spanish-Portuguese
Manual				
11	Curin et. al. (2004)	Manually obtained translations in target language	Parallel Syntactically Annotated Corpus	Czech-English
12	Garcia (2009)		(IMS Workbench Format)	Portugese-English
13	Choudhary and Jha (2011)	Manually obtained translations in target language	Sentence aligned (in Excel files)	Different Pairs of Indian Languages including Hindi-Marathi
14	Rosen and Vavrin (2012)	Manually obtained translations in target language	Syntactically annotated and sentence aligned (txt format)	

15	Megyesi, Hein and Johanson (2006)	Manually obtained translations in target language	Sentence aligned (in XML CES)	Swedish-Turkish
16	Post, Burch and Osborne (2012)	Manually obtained translations in target language		Bengali, Hindi, Malayalam, Tamil, Telugu, Urdu
Machine Translation				
17	Avramidis et. al. (2012)	Translations obtained in target language through SMT system		

IV. CONCLUSIONS

The paper discusses the three approaches used by different authors. None of the methods seem to be having a special applicability. The choice of the approach seems to be dictated by the liking of the building team, purpose for which the corpus is being built, and to some extent the domain.From information technology point of view it is evident that none of the sources seem to have made any attempt at defining requirements of optimum sizes and standardizations of architectures and frameworks to help the builders of corpora.

REFERENCES

[1] Garje, G. V., &Kharate, G. K. (2013). Survey of machine translation systems in India. International Journal on Natural Language Computing (IJNLC), 2(4), 47-67.

[2] Shen, G. R. (2011). Corpus-based Approaches to Translation Studies. Cross-Cultural Communication, 6(4), 181-187.

[3] Jayaram, B. D., &Rajyashree, K. S. (2005). Corpora in Indian languages. Problems of Quantitative Linguistics, 323-329.

[4] Dash, N. S., &Chaudhuri, B. B. (2001, November). Why do we need to develop corpora in Indian languages? In the International Working Conference on Sharing Capability in Localization and Human Language Technologies SCALLA-2001. Bangalore.

[5] Singh, T. D. (2012). Building Parallel Corpora for SMT System: A Case Study of English-Manipuri. International Journal of Computer Applications, 52(14).

[6] Botley, S., McEnery, T., & Wilson, A. (Eds.). (2000). Multilingual corpora in teaching and research (No. 22). Rodopi.

[7] Singh, A. K., &Surana, H. (2007a, June). Can corpus based measures be used for comparative study of languages? In Proceedings of ninth meeting of the ACL special interest group in computational morphology and phonology (pp. 40-47). Association for Computational Linguistics.

[8] Eberle, K., Geiß, J., Ginestí-Rosell, M., Babych, B., Hartley, A., Rapp, R., Sharoff, S. & 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). Association for Computational Linguistics.

[9] 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). Association for Computational Linguistics.

[10] Srivastava, R., &Bhat, R. A. (2013). Transliteration Systems across Indian Languages Using Parallel Corpora. In PACLIC.

[11] Ali, A., Siddiq, S., & Malik, M. K. (2010). Development of parallel corpus and English to Urdu statistical machine translation. Int. J. of Engineering & Technology IJET-IJENS, 10, 31-33.

[12] Nair, L. R., & David Peter, S. (2012). Machine translation systems for Indian languages. International Journal of Computer Applications (0975–8887), 39(1).

[13] Sreelekha, S., Bhattacharyya, P., &Malathi, D. (2014). Lexical resources for Hindi-Marathi MT. In: The WILDRE2 2nd Workshop on Indian Language Data: Resources and evaluation.

[14] Sinha, R. M. K. (2009, August). Automated mining of names using parallel Hindi-English corpus. In Proceedings of the 7th Workshop on Asian Language Resources (pp. 48-54). Association for Computational Linguistics.

[15] Jagarlamudi, J., &Kumaran, A. (2007, September). Cross-Lingual Information Retrieval System for Indian Languages. In CLEF (pp. 80-87).

[16] 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.

[17] McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning? Handbook of research in second language teaching and learning, 2, 364-380.

[18] Steinberger, R., Eisele, A., Klocek, S., Pilos, S., &Schlu¨ter, P. (2013). DGT-TM: A freely available translation memory in 22 languages. arXiv preprint arXiv:13095226.

[19] Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. Computational linguistics, 19(1), 75-102.

[20] Ma, X. (2006, May). Champollion: A robust parallel text sentence aligner. In LREC 2006: Fifth International Conference on Language Resources and Evaluation (pp. 489-492).

[21] Liu, W., Chang, Z., Teahan, W., 2014. Experiments with compression-based methods for English-Chinese

sentence alignment. In Proceedings of Second International Conference on Statistical Language and Speech Processing (SLSP), Springer International Publishing, pp. 14–16.

[22] Martin, J., Johnson, H., Farley, B., &Maclachlan, A. (2003, May). Aligning and using an English-Inuktitut parallel corpus. In Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond-Volume 3 (pp. 115-118). Association for Computational Linguistics.

[23] Tiedemann, J. (2007). Building a multilingual parallel subtitle corpus. Proc. CLIN, 14.

[24] Liu, Z. (2013). Automated Building of Sentence-Level Parallel Corpus and Chinese-Hungarian Dictionary (Doctoral dissertation, WORCESTER POLYTECHNIC INSTITUTE).

[25] Nazar, R. (2011). Parallel corpus alignment at the document, sentence and vocabulary levels. Procesamiento del lenguaje natural, (47).

[26] Pilevar, M. T., Faili, H., &Pilevar, A. H. (2011, February). Tep: Tehran english-persian parallel corpus. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 68-79). Springer Berlin Heidelberg.

[27] Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., & Yi, L. (2014). UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. In LREC (pp. 1837-1842).

[28] 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). ACM.

[29] Bin, L. U., Jiang, T., Chow, K., & BENJAMIN K, T. (2010). Building a large English-Chinese parallel corpus from comparable patents and its experimental application to SMT. InProceedings of the 3rd Workshop on Building and Using Comparable Corpora (pp. 42-49).

[30] Aziz, W. F., Pardo, T. A., &Paraboni, I. (2008, October). Building a Spanish-Portuguese parallel corpus for statistical machine translation. In Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web (pp. 369-371). ACM.

[31] Sridhar, V. K. R., Barbosa, L., & Bangalore, S. (2011). A Scalable Approach to Building a Parallel Corpus from the Web. In INTERSPEECH (pp. 2113-2116).

[32] Cuřín, J., Čmejrek, M., Havelka, J., &Kuboň, V. (2004, March). Building a parallel bilingual syntactically annotated corpus. In International Conference on Natural Language Processing (pp. 168-176). Springer, Berlin, Heidelberg.

[33] Zhang, Y., Uchimoto, K., Ma, Q., &Isahara, H. (2005). Building an annotated Japanese-Chinese parallel corpus–a part of NICT multilingual corpora. In Second International Joint Conference on Natural Language Processing (pp. 85-90).

[34] Chang, B. (2004). Chinese-English parallel corpus construction and its application. In Proceedings of The 18th Pacific Asia Conference on Language, Information and Computation (pp. 283-290).

[35] Megyesi, B. B., Hein, A. S., &Johanson, E. C. (2006). Building a swedish-turkish parallel corpus. LREC, Genoa, Italy.

[36] Frankenberg-Garcia, A. (2009). Compiling and using a parallel corpus for research in translation. Babel: international journal of translation, 21(1), 57-71.

[37] Choudhary, N., &Jha, G. N. (2011, November). Creating multilingual parallel corpora in indian languages. In Language and Technology Conference (pp. 527-537). Springer, Cham.

[38] Rosen, A., &Vavrín, M. (2012). Building a multilingual parallel corpus for human users. In LREC (pp. 2447-2452).

[39] Samy, D., Sandoval, A. M., Guirao, J. M., &Alfonseca, E. (2006). Building a Parallel Multilingual Corpus (Arabic-Spanish-English). In Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC.

[40] Avramidis, E., Ruiz Costa-Jussà, M., Federmann, C., Melero, M., Pecina, P., & Van Genabith, J. (2012). A Richly annotated, multilingual parallel corpus for hybrid machine translation. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (pp. 2189-2193). European Language Resources Association (ELRA).

[41] Yeka, J. R., Kolachina, P., & Sharma, D. M. (2014, May). Benchmarking of English-Hindi parallel corpora. In LREC (pp. 1812-1818).