Indian Journal of Science and Technology, Vol 9(45), DOI: 10.17485/ijst/2016/v9i45/106520, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Fast and Efficient Framework for Creating Parallel Corpus

B. Premjith*, S. Sachin Kumar, R. Shyam, M. Anand Kumar and K. P. Soman

Centre for Computational Engineering and Networking (CEN),Amrita School of Engineering, Amrita University, Amrita Vishwa Vidyapeetham, Amritanagar, Coimbatore – 641 112, Tamilnadu, India; b_premjith@cb.amrita.edu, sachinnme@gmail.com, shyam.neezhoor@gmail.com, m_anandkumar@cb.amrita.edu, kp_soman@amrita.edu

Abstract

Objectives: A framework involving Scansnap SV600 scanner and Google Optical character recognition (OCR) for creating parallel corpus which is a very essential component of Statistical Machine Translation (SMT). **Methods and Analysis:** Training a language model for a SMT system highly depends on the availability of a parallel corpus. An efficacious approach for collecting parallel sentences is the predominant step in an MT system. However, the creation of a parallel corpus requires extensive knowledge in both languages which is a time consuming process. Due to these limitations, making the documents digital becomes very difficult and which in turn affects the quality of machine translation systems. In this paper, we propose a faster and efficient way of generating English to Indian languages parallel corpus with less human involvement. With the help of a special type of scanner called Scansnap SV600 and Google OCR and a little linguistic knowledge, we can create a parallel corpus for any language pair, provided there should be paper documents with parallel sentences. **Findings:** It was possible to generate 40 parallel sentences in 1 hour time with this approach. Sophisticated morphological tools were used for changing the morphology of the text generated and thereby increase the size of the corpus. An additional benefit of this is to make ancient scriptures or other manuscripts in digital format which can then be referred by the coming generation to keep up the traditions of a nation or a society. **Novelty:** Time required for creating parallel corpus is reduced by incorporating Google OCR and book scanner.

Keywords: Google OCR, Machine Translation, Parallel Corpus, Statistical Machine Translation, Scansnap SV600 Scanner

1. Introduction

Roughly, there are more than 6900 spoken languages around the globe[1]. Each part of the world has its own local language and it becomes very difficult to exchange information between people living in different parts of the world. According to the Census, 2001 in India, we have 122 major languages and 1599 other languages[2]. The disparities in the linguistic information in many languages are even more striking. So communication between people speaking different languages becomes a tough task. Even though English is renowned as the global language, people tend to speak their own natural languages. Communication in local languages gives one the freedom to think in one's own language and to use the vocabulary which is familiar. At this juncture, Nelson Mandela's words have great significance. Mandela once said, "If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language that goes to his heart". This shows the importance of sharing information in local or natural language. Here comes the importance of translation between languages. Machine Translation system is very helpful in translating online news, scientific materials, tourism information, movie information etc.

Today English is widely used in all the official as well as unofficial communications. However, the impact of local language is still strong. Moreover many people who are non-native speakers of English are illiterate as far as English is concerned. So a translation method

*Author for correspondence

from English to a local language and vice versa is very important. Despite the success rate of human translation, machine translations are more preferable as the former one is a time consuming process. There are different approaches towards machine translation – Rule Based Machine Translation (RBMT)[3], SMT[4] and Hybrid machine translation[5] which combines the advantages of both RBMT and SMT. RBMT requires deep linguistic knowledge about both target and source language. The key intuition behind SMT is the availability of bilingual as well as monolingual corpus. Since SMT systems are based on the statistical features extracted from the data, reasonable quantity of quality corpus is very essential. So the optimal translation probability estimation heavily relies on the quantity and quality of the corpus. However, large corpora are not easily available[6]. The availability of a large quality corpus is essential for machine translation using neural networks[7–9].

Unavailability of quality bilingual corpus for English and Indian languages is the main reason behind the underperformance of English to Indian languages machine translation system. Compared to European language pairs, digital records of parallel sentences in English - Indian languages are very less. Creating such a massive parallel corpus for a machine translation system is a herculean task. It also requires a huge volume of man power with deep knowledge in language and time to create a parallel corpus. There are certain parallel corpuses such as Technology Development for Indian Languages (TDIL) corpus[10] available for English – Indian languages which are not freely available. Amazon's Mechanical Turk[11], crowd-souring marketplace over the internet has been used for creating parallel corpus for many languages such as Hindi, Telugu, Spanish, Urdu, Chinese and Creole[12,13]. But the quality is questionable. For SMT, we need a good quality corpus from which translation model learns the problem better. Therefore, while collecting more and more parallel corpus, it is necessary to find effective ways for making better use of available parallel training data. In this paper, we propose a way of creating effective parallel corpus for English and Malayalam by using a special scanner and Google OCR. This same approach can also be used for the creation of other resource poor languages. Further, this method aids to make the digital version of sacred writings and other primordial texts. Therefore, in addition to construct a corpus for a machine translation system, this method brings a sociological relevance to the customs and traditions of different geographical locations around the world. Apart from creating a parallel corpus, this approach can also be used for the creation of comparable corpus[14].

2. Proposed Method

The proposed method is an effective and efficient way of creating parallel corpus for English and Malayalam. The method utilizes the efficiency of Google OCR[15] for converting the images into text format. Large varieties of books are scanned with a Scansnap SV6000 scanner, shown in Figure 1[16]. This scanner is able to scan any document very easily without making any damages to the original document (as we do with typical scanners). It can scan documents of size A3, A4, B5, B4 etc. Google OCR gives good results for the scanned images. Multiple images can be automated by the procedure of OCR and a little modification is required. Manual work is required for aligning the parallel sentences. Steps and basic block diagram for corpus creation are given below.

Figure 1. Scansnap SV6000 scanner.

Generation of Unicode representation of hardcopies of books undergoes the following steps and the block diagram is shown in Figure 2.

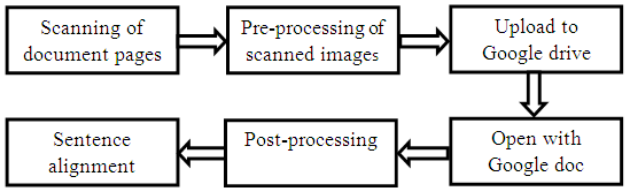


Figure 2. Block diagram of steps for creating parallel corpus.

1. Scanning pages
2. Skew correction and cropping of images
3. Upload to Google drive
4 .Open the image as Google doc

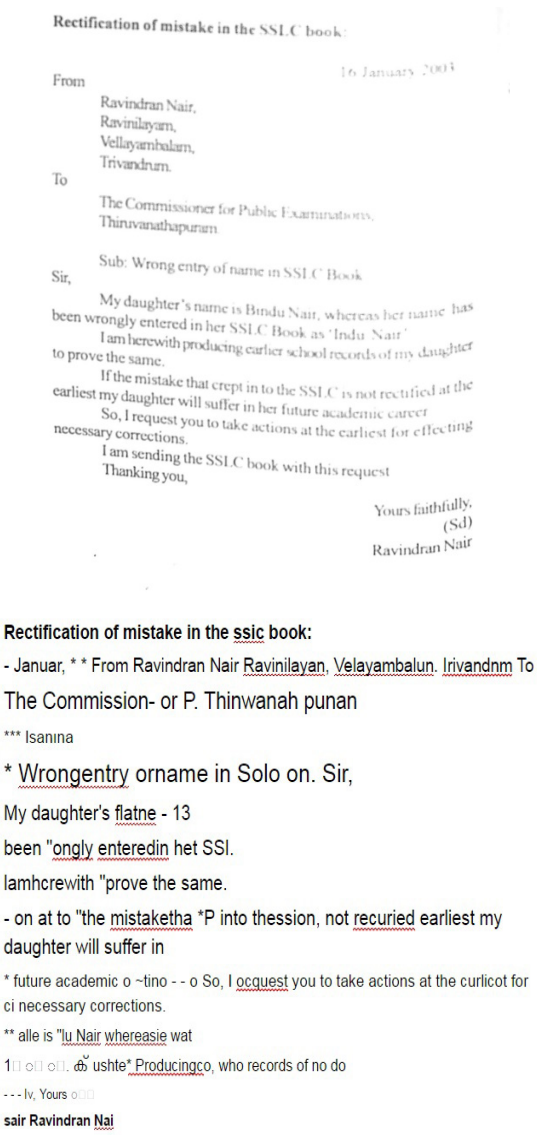
5 .Post-processing
5. Sentence alignment.

Parallel sentences were collected from sources such as open available encyclopedia, Kerala[17] and Tamil Nadu[18] school text books (both English and Malayalam medium) from standard 1 to standard 10, Ente Nadodikkathakal – My Folk tales (Bilingual)[19], The art of letter writing[20] and a few English-Malayalam speaking course books. Lengthy sentences collected from these sources are split to obtain short sentences. Such changes should be made in sentences from both languages. Short sentences or simple sentences from both languages are able to emanate meanings very lucid. This increases the number of sentences in the corpus. In order to avoid copy right issues, resultant sentences can be modified using morphological tools[21–23].Nouns, Verbs and other morphological information can be altered to increase the size of bilingual corpus. This 'corpus' is created strictly in view of research purpose. Transformed sentences look like encrypted copies of original sentences extracted from the documents. It is quite impossible to construct original sentences (from the above mentioned books) from the modified sentences for it is impractical to distil the changes we have made. Therefore this parallel corpus creation will never affect the business of respective publishers.

3. Experiments and Results

A lot of documents collected from various resources which contain both English and Malayalam sentences were scanned for this work. Though identification and collection of documents with such sentence pairs was a massive task, numerous books were searched and picked out many parallel sentences. Before the scanning begins, a few arrangements have to be made. Since the scanner light is very bright, documents have to be placed in a proper height. The problem with bad lighting causes bad generation of OCR and an example is shown in Figure 3. So the document to be scanned must be lighted fairly such that all the letters are clearly visible.

Another important thing which is to be taken into consideration is placement of documents. If alignment of the document is not proper and if the document is tilted or shaken during scanning, we will not get good OCR of the respective document. These issues are shown in Figure 4 and Figure 5. Another important thing to be taken into account is the quality of the text document. When we


Rectification of mistake in the SSLC book

16 January 2003

From

Ravindran Nair,
Ravinilayam,
Vellayambalam,
Trivandrum.

To

The Commissioner for Public Examinations,
Thiruvanathapuram

Sub: Wrong entry of name in SSLC Book

Sir,

My daughter's name is Bindu Nair, whereas her name has been wrongly entered in her SSLC Book as 'Indu Nair'.
I am herewith producing earlier school records of my daughter to prove the same.
If the mistake that crept in to the SSLC is not rectified at the earliest my daughter will suffer in her future academic career.
So, I request you to take actions at the earliest for effecting necessary corrections.
I am sending the SSLC book with this request
Thanking you,

Yours faithfully,
(Sd)
Ravindran Nair

Rectification of mistake in the ssic book:

- Januar, * * From Ravindran Nair Ravinilayan, Velayambalun. Irivandnm To

The Commission- or P. Thinwanah punan

*** Isanina

* Wrongentry orname in Solo on. Sir,

My daughter's flatne - 13

been "ongly enteredin het SSI.

lamhcrewith "prove the same.

- on at to "the mistaketha *P into thession, not recuried earliest my daughter will suffer in

* future academic o ~tino - - o So, I ocquest you to take actions at the curlicot for ci necessary corrections.

** alle is "lu Nair whereasie wat

1□ o□ o□, ☙ ushte* Producingco, who records of no do

- - - Iv, Yours o□□

sair Ravindran Nai

Figure 3. Bad lighting gives bad OCR.

scan low quality documents such as old manuscripts, we get highly noisy images and it becomes difficult to obtain good OCR and this issue is shown in Figure 6. In Figure 7, it is shown that a text document with proper alignment and lighting will give accurate results with Google OCR. Even if we scan documents with at most care, some preprocessing has to be done to the scanned documents since Google OCR gives very good result for good quality images. So before uploading to Google drive, preprocessing like skew correction and cropping out of unnecessary things are done on the scanned pages. If needed, to improve the image quality further, brightness, contrast and sharpness can also be adjusted. Software tools such as scan tailor, Google Picasa and Adobe Photoshop can be

employed to do this preprocessing step.OCR of an image or a pdf is obtained by uploading it to Google drive and opens it as Google doc. Since Google allows creation of one document at a time, multiple uploading and multiple OCR creation was automated. Even though we upload good quality images, there is a possibility of getting small errors in the OCR. Errors such as spelling corrections, removing unwanted spaces and unwanted characters and adding missing spaces and missing characters can be easily rectified manually. From our experience this method gives about 98% accuracy in generating the OCR of both English and Malayalam image documents for a quality input document. The last step of creating the parallel corpus is the most tedious job as it requires lot of time and concentration. Initially we split larger sentences into shorter sentences. This process will increase the number of sentences in the corpus. Moreover, short sentences convey meaning simply compared to lengthy complex sentences. Examples of splitting long sentences into short ones are given below. This requires small modifications in sentences in both languages.

Example for splitting the sentences into shorter sentences:

English:

Before splitting:

People who were surrounding him clapped their hands, and threw coins at him to encourage him.

After splitting:

1. People who were surrounding him clapped their hands.
2. They threw coins at him to encourage him.

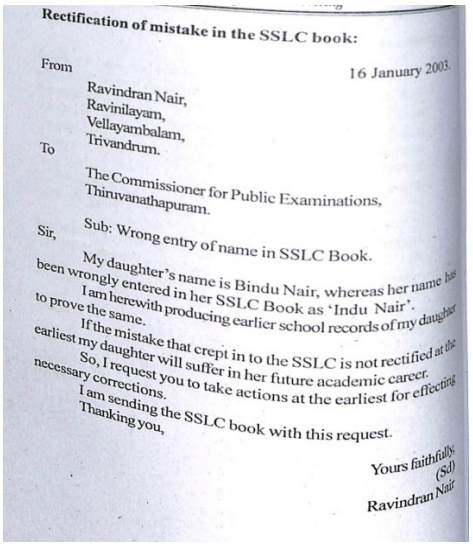
Malayalam:

Before splitting:

ചുറ്റുംകൂടിനിന്നിരുന്നജനങ്ങൾകൈ കൊ ട്ടുകയുംഅയാളെപ്രോ ത്സാഹിപ്പിക്കുവാൻനാണയത്തുട്ടുകൾഎറിഞ്ഞുകൊ ടുക്കുകയുംചെയ്തു.

After splitting:

1. അയാളുടെചുറ്റുംകൂടിനിന്നിരുന്നജനങ്ങൾകൈകൊ ട്ടി.
2. അയാളെപ്രോ ത്സാഹിപ്പിക്കുവാൻഅവർനാണയത്തുട്ടുകൾഎറിഞ്ഞുകൊ ടുത്തു.



Rectification of mistake in the SSLC book:

From

16 January 2003.

Ravindran Nair,
Ravinilayam,
Vellayambalam,
Trivandrum.

To

The Commissioner for Public Examinations,
Thiruvanathapuram.

Sir,

Sub: Wrong entry of name in SSLC Book.

My daughter's name is Bindu Nair, whereas her name has been wrongly entered in her SSLC Book as 'Indu Nair'.

I am herewith producing earlier school records of my daughter to prove the same.

If the mistake that crept in to the SSLC is not rectified at the earliest my daughter will suffer in her future academic career.

So, I request you to take actions at the earliest for effecting necessary corrections.

I am sending the SSLC book with this request.

Thanking you,

Yours faithfully,
(Sd)
Ravindran Nair

o her nam Bindu Nair, whereas

SLC Book as 'Indu Sthe uhplOdu

"sories recordso

ified8 :."ouessicismorei -

- ef. . . ! Srin he, future academic caro so". ° take a....

- for c **tions at the earliest

Figure 4. Bad alignment gives bad OCR.

By this way, we were able to generate 40 sentences per hour. This number can again be increased by changing the morphology with certain morphological tools developed by CEN department, Amrita Vishwa Vidyapeetham. We changed nouns, verbs and tenses of all the obtained sentences so as to increase the number of sentences in the corpus by 10 times. This procedure requires profound knowledge in both languages. Sentence alignment is done at this step which needs linguistic expertise in language since meaning of parallel sentences should be consistent.

A few examples for how morphological information is changed are given below.

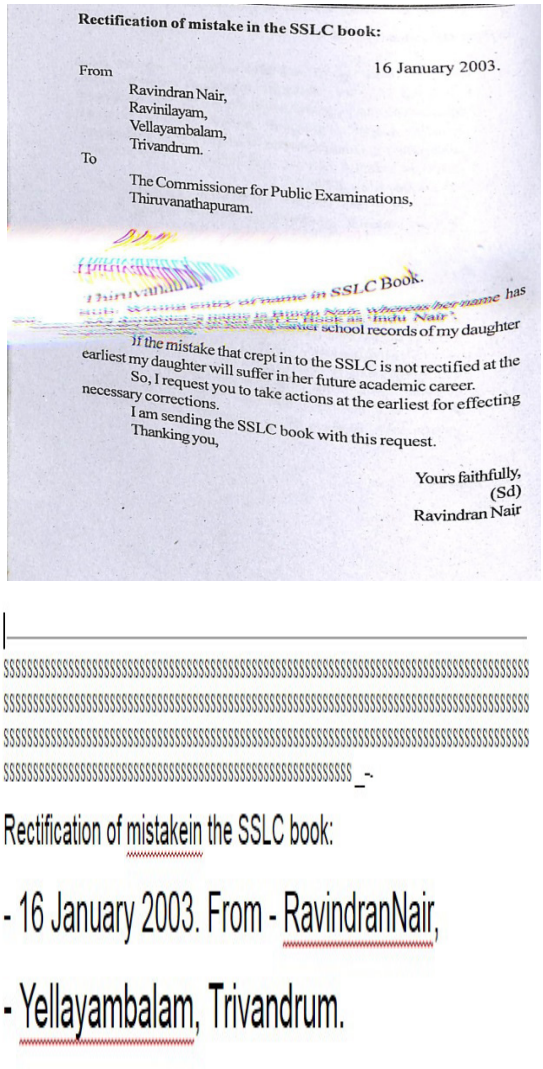


Figure 5. If the image is shaken during scanning, we will obtain poor OCR.

Example for morphological changes made to the extracted text:

English: I know that you will kill me.

Malayalam:

എനിക്കറിയാംനീഎന്നകെ ാല്ലും.

After changing morphological information,

English: I know that he will kill me.

Malayalam:

എനിക്കറിയാംഅവൻഎന്നകെ ാല്ലും.

English: I know that he is going to kill her.

Malayalam: എനിക്കറിയാംഅവൻഅവ ളകെ ാല്ലാൻപ ോകുകയാണന്ന്.

English: She knows that they see her.

Malayalam: അവൾക്കറിയാംഅവർഅവ ളകൊണുന്നുഎന്ന്.

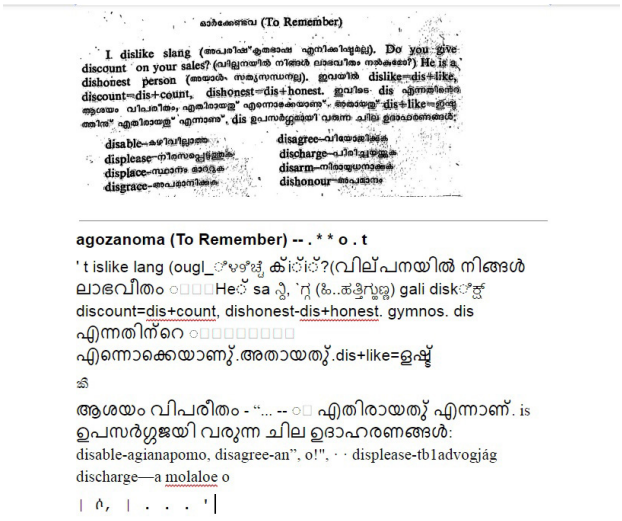
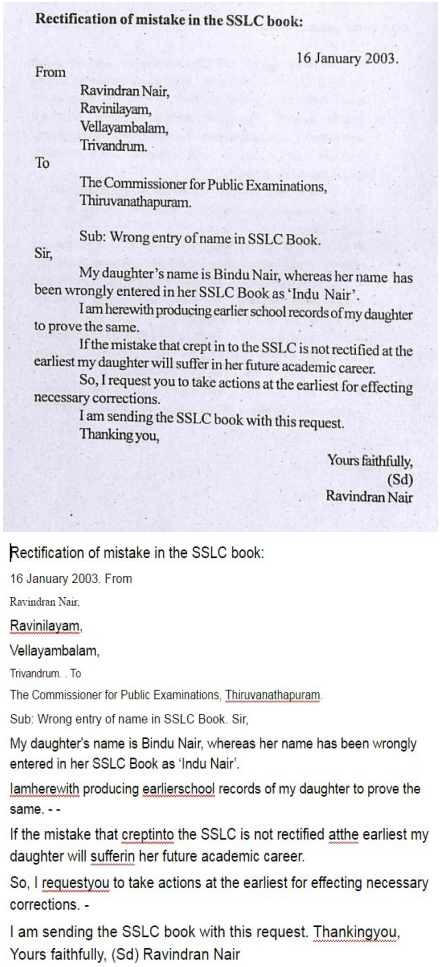


Figure 6. Poor quality images leads to poor OCR creation.

Figure 8 shows the OCR of an image document with some special fonts. Fonts in italic, fonts like handwritten letters, unconventional style fonts etc. gives very poor accuracy in OCR. Another dilemma we faced was the generation of OCR from image converted pdf. OCR of such documents provides very less accuracy (less than 10%). So in order to deal with such scenario, we convert those pdfs to images and then upload to Google drive. Conversion of pdfs to images using online tools also was automated. Developing a corpus using this approach can be applied to any language, especially languages with very less digital documents. However, in order to make a bilingual corpus, availability of parallel sentences in hard copy is very much essential. This is one problem we encounter with the creation of bilingual corpus for Indian language pairs. In the case of Encyclopedia Britannica, we met with another problem of finding respective Malayalam article for an English article which took a definite amount of time.

4. Conclusion

Unavailability of parallel corpus is a main drawback of SMT systems especially systems involving Indian lan-



Rectification of mistake in the SSLC book:

16 January 2003.

From

Ravindran Nair,
Ravinilayam,
Vellayambalam,
Trivandrum.

To

The Commissioner for Public Examinations,
Thiruvanathapuram.

Sub: Wrong entry of name in SSLC Book.

Sir,

My daughter's name is Bindu Nair, whereas her name has been wrongly entered in her SSLC Book as 'Indu Nair'.

I am herewith producing earlier school records of my daughter to prove the same.

If the mistake that crept in to the SSLC is not rectified at the earliest my daughter will suffer in her future academic career.

So, I request you to take actions at the earliest for effecting necessary corrections.

I am sending the SSLC book with this request.

Thanking you,

Yours faithfully,
(Sd)
Ravindran Nair

Rectification of mistake in the SSLC book:
16 January 2003. From
Ravindran Nair,
Ravinilayam,
Vellayambalam,
Trivandrum. . To
The Commissioner for Public Examinations, Thiruvanathapuram.
Sub: Wrong entry of name in SSLC Book. Sir,
My daughter's name is Bindu Nair, whereas her name has been wrongly entered in her SSLC Book as 'Indu Nair'.
Iamherewith producing earlierschool records of my daughter to prove the same. - -
If the mistake that creptinto the SSLC is not rectified atthe earliest my daughter will sufferin her future academic career.
So, I requestyou to take actions at the earliest for effecting necessary corrections. -
I am sending the SSLC book with this request. Thankingyou,
Yours faithfully, (Sd) Ravindran Nair

Figure 7. Properly aligned scanned documents give good OCR.

No. VOTI TO ξ\(YOKHΞ\ Λ; N. BSE 'gy € ﻫ C E N → - <-

We ク○○のエンCのルee/%

& Fuustuvest. | of

- Booting Future . . . Stipend. . . . . Donel o

Project exp . . [Donel ₹ ! Research . . . . . | Donel

Getting UOb c co-c to to o o ========== I,100% a - a ---------- IDonel %, Knowledge . . . [Donel
Publications . . . . . - - c_ - _0 - - - - - - [Donel

M.Tech. . [Well Donel

cENTRE FOR EXCELLENCE IN

COMPUTATIONAL ENGINEERING & NETworking (CEN)

AΛΛEITA

VISHWA VIDYAP $ U N I v E R FITA Established under Section 3 of the YGCAct 1956

Figure 8. OCR for some special type of fonts gives poor accuracy.

guages. Generation of such corpus is considered as a very heavy task as it involves lot of human resource and the amount of time it takes to prepare a corpus with very large number of sentences. So a fast and less time consuming approach towards the generation of parallel corpus was high in demand. We propose a method through which a parallel corpus can be generated with less human involvement and time. This approach makes use of a special scanner, Scansnap SV6000, and automates the intermediate processes. One of the important things to be taken care while preparing the corpus is the one who is preparing the corpus should have sound knowledge in both languages. Morphology of the resultant sentences can be changed to increase the size of the parallel corpus.

5. References

1. How many languages are there in the world? [Internet]. 2016 [cited 2016 Jul 7]. Available from: http://www.linguisticsociety.org/content/how-many-languages-are-there-world.
2. Languages of India [Internet]. 2016 [cited 2016 Jul 7]. Available from: https://en.wikipedia.org/wiki/Languages_of_India.
3. Nirenburg S. Knowledge-based machine translation. Machine Translation. 1989 Mar 1; 4(1):5–24.
4. Koehn P. SMT. Cambridge University Press; 2009 Dec 17.

5. Sawaf H, Shihadah M, Yaghi M, inventors; AppTek, assignee. Hybrid machine translation. United States patent application US 12/606,110; 2010 Jul 15.
6. Lü Y, Huang J, Liu Q. Improving SMT performance by training data selection and optimization. In EMNLP-CoNLL. 2007 Jun 28; 34:3–350.
7. Chung J, Cho K, Bengio Y. A character-level decoder without explicit segmentation for neural machine translation [Internet]. 2016 [updated 2016 Jun 21; cited 2016 Mar 19]. Available from: arXiv: 1603.06147.
8. Firat O, Cho K, Bengio Y. Multi-way, multilingual neural machine translation with a shared attention mechanism [Internet]. 2016 [cited 2016 Jan 6]. Available from: arXiv: 1601.01073.
9. Jussà MRC, Fonollosa JA. Character-based neural machine translation [Internet]. 2016 [updated 2016 Jun 30; cited 2016 Mar 2]. Available from: arXiv: 1603.00810.
10. TDIL [Internet]. 2016 [cited 2016 Jul 7]. Available from: http://tdil.mit.gov.in/.
11. Amazon Mechanical Turk [Internet]. 2016 [cited 2016 Jul 7]. Available from: https://www.mturk.com/mturk.
12. Ambati V, Vogel S. Can crowds build parallel corpora for machine translation systems? In Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with Amazon's Mechanical Turk. Association for Computational Linguistics; 2010 Jun 6. p. 62–5.
13. Burch CC, Dredze M. Creating speech and language data with Amazon's Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with Amazon's Mechanical Turk. Association for Computational Linguistics; 2010 Jun 6. p. 1–12.
14. Ansari E, Sadreddini MH, Tabebordbar A, Sheikhalishahi M. Combining different seed dictionaries to extract lexicon from comparable corpus. Indian Journal of Science and Technology. 2014 Sep 15; 7(9):1279–88.
15. Google OCR [Internet]. 2016 [cited 2016 Jul 7]. Available from: https://support.google.com/drive/answer/176692?hl=en.
16. Scansnap SV600 [Internet]. 2016 [cited 2016 Jul 7]. Available from: http://www.fujitsu.com/global/products/computing/peripheral/scanners/scansnap/sv600/.
17. SCERT, Kerala [Internet]. 2016 [cited 2016 Jul 7]. Available from: http://www.scert.kerala.gov.in/index.php?option=com_content&view=article&id=86&Itemid=76.
18. Tamil Nadu School Text Books [Internet]. 2016 [cited 2016 Jul 7]. Available from: http://www.textbooksonline.tn.nic.in/.
19. Prabhakumar TL, Balakrishnan V. EnteNadodikkathakal – My Folk tales (Bilingual). Arshaasri Publishing Co.
20. James R. The art of letter writing (Malayalam – English). Dronacharya Publications; 2006.
21. Kumar MA, Dhanalakshmi V, Soman KP, Rajendran S. Factored SMT system for English to Tamil language. Pertanika Journal of Social Science and Humanities. 2014; 22(4):1045–61,
22. Kumar MA, Dhanalakshmi V, Soman KP, Rajendran S. A sequence labeling approach to morphological analyzer for Tamil language. International Journal on Computer Science and Engineering. 2010; 2(6):1944–5.
23. Dhanalakshmi V, Rekha RU, Kumar A, Soman KP, Rajendran S. Morphological analyzer for agglutinative languages using machine learning approaches. In Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference IEEE; 2009 Oct 27. p. 433–5.