Automatic Creation of a Sentence Aligned Sinhala-Tamil Parallel Corpus

Riyafa Abdul Hameed, Nadeeshani Pathirennehelage, Anusha Ihalapathirana, Maryam Ziyad Mohamed, Surangika Ranathunga, Sanath Jayasena, Gihan Dias, Sandareka Fernando
Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka
{riyafa.12,pnadeeshani.12,anusha.12,maryamzi.12,surangika,
sanath,gihan,sandarekaf}@cse.mrt.ac.lk

Abstract

A sentence aligned parallel corpus is an important prerequisite in statistical machine translation. However, manual creation of such a parallel corpus is time consuming, and requires experts fluent in both languages. Automatic creation of a sentence aligned parallel corpus using parallel text is the solution to this problem. In this paper, we present the first ever empirical evaluation carried out to identify the best method to automatically create a sentence aligned Sinhala-Tamil parallel corpus. Annual reports from Sri Lankan government institutions were used as the parallel text for aligning. Despite both Sinhala and Tamil being under-resourced languages, we were able to achieve an F-score value of 0.791 using a hybrid approach that makes use of a bilingual dictionary.

1 Introduction

Sentence and word aligned parallel corpora are extensively used for statistical machine translation (Al-Onaizan et al., 1999; Callison-Burch, 2004) and in multilingual natural language processing (NLP) applications (Kaur and Kaur, 2012). In recent years, parallel corpora have become more widely available and serve as a source for data-driven NLP tasks for languages such as English and French (Hallebeek, 2000; Kaur and Kaur, 2012).

A parallel corpus is a collection of text in one or more languages with their translation into another language or languages that have been stored in a machine-readable format (Hallebeek, 2000). A parallel corpus can be aligned either at sentence level or word level. Sentence and word alignment of parallel corpus is the identification of the corresponding sentences and words (respectively) in both halves of the parallel text.

Sentence alignment could be of various combinations including one to one where one sentence maps to one sentence in the other corpus, one to many where one sentence maps to more than one sentences in the other corpus, many to many where many sentences map to many sentences in the other corpus or even one to zero where there is no mapping for a particular sentence in the other corpus.

For statistical machine translation, the more the number of parallel sentence pairs, the higher the quality of translation (Koehn, 2010). However, manual alignment of a large number of sentences is time consuming, and requires personnel fluent in both languages. Automatic sentence alignment of a parallel corpus is the widely accepted solution for this problem. Already many sentence alignment techniques have been implemented for some languages pairs such as English-French (Gale and Church, 1993; Brown et al., 1991; Chen, 1993; Braune and Fraser 2010; Lamraoui and Langlais, 2013), English-German (Gale and Church, 1993) English-Chinese (Wu, 1994; Chuang and Yeh, 2005)

Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing,
pages 124–132, Osaka, Japan, December 11-17 2016.

and Hungarian-English (Varga et al., 2005; Tóth et al., 2008). However, none of these techniques have been evaluated for Sinhala and Tamil, the two official languages in Sri Lanka.

This paper presents the first ever study on automatically creating a sentence aligned parallel corpus for Sinhala and Tamil. Sinhala and Tamil are both under-resourced languages, and research implementing basic NLP tool such as POS taggers and morphological analysers is at its inception stage (Herath et al., 2004; Hettige and Karunananda, 2006; Anandan et al., 2002). Therefore, not all the aforementioned sentence alignment techniques are applicable in the context of Sinhala and Tamil. With this limitation in mind, an extensive literature study was carried out to identify the applicable sentence alignment techniques for Sinhala and Tamil. We implemented six such methods, and evaluated their performance using a corpus of 1300 sentences based on the precision, recall, and F-measure using annual reports of Sri Lankan government departments as the source text. The highest F-measure value of 0.791 was obtained for Varga et al.'s (2005) Hunalign method, the hybrid method that combined the use of a bilingual dictionary with the statistical method by Gale and Church (1993).

The rest of the paper is organized as follows. Section 2 identifies related work in this area. Section 3 describes how different techniques were employed in the alignment process, and section 4 presents the results for these techniques. Section 5 contains a discussion of these results while section 6 presents the conclusion and future work.

2 Related Work

Automatic sentence alignment techniques can be broadly categorized into three classes: statistical, linguistic, and hybrid methods. Statistical methods use quantitative measures (such as sentence size, sentence character number) to create an alignment relationship; linguistic methods use linguistic knowledge gained from sources such as morphological analyzers, bilingual dictionaries, and word list pairs, to relate sentences; hybrid methods combine the statistical and linguistic methods to achieve accurate statistical information (Simões, 2004).

2.1 Statistical Methods

Gale and Church (1993), and Brown et al. (1991) have introduced statistical methods for aligning sentences that have been successfully used for European languages, including English-French, English-German, English-Polish, English-Spanish (McEnery et al., 1997), English-Dutch and Dutch - French (Paulussen et al, 2013).

These methods have also been used with Non-European languages such as English - Chinese (McEnery and Oakes, 1996), Italian-Japanese (Zotti et al, 2014), English-Arabic (Alkahtani et al, 2015), and English-Malay (Yeong et al, 2016). The general idea of these methods is that the closer in length two sentences are, the more likely they align. Brown et al.'s (1991) method aligns sentences based on sentence length measured using word count. Here anchor points are used for alignment. Gale and Church use the number of characters as the length measure. While the parameters such as mean and variance for Gale and Church's (1993) method are considered language independent for European languages, tuning these for non-'European language pairs has improved results (Zotti et al, 2014).

Both these methods have given good accuracy in alignment; however they require some form of initial alignment or anchor points.

Method by Chuang and Yeh (2005) exploits the statistically ordered matching of punctuation marks in the two languages English and Chinese to achieve high accuracy in sentence alignment compared with using the length-based methods alone.

2.2 Linguistic Methods

Linguistic methods exploit the linguistic characteristics of the source and target languages such as morphology and sentence structure to improve the alignment process. However linguistic methods are not used independently but have been introduced in conjunction with statistical methods, forming hybrid methods as described in the next section.

2.3 Hybrid Methods

Statistical methods such as that of Brown et al., (1991), and Gale and Church (1991) require either corpus-dependent anchor points, or prior alignment of paragraphs to obtain better accuracy. Hybrid

methods make use of statistical as well as linguistic features of the sentences obtaining better accuracy in documents with or without these types of prior alignments. Hence hybrid methods are widely used to achieve higher accuracy in alignment. The methods by Wu (1994), Chen (1993), Moore (2002), Varga et al. (2005), Sennrich and Volk (2011), Lamraoui and Langlais (2013), Braune and Fraser (2010), Tóth et al. (2008) and Mújdricza-Maydt et al. (2013) are some of them.

The method used by Wu (1994) is a modification of Gale and Church's (1993) length-based statistical method for the task of aligning English with Chinese. It uses a bilingual external lexicon with lexicon cues to improve the alignment accuracy. Dynamic programming optimization has been used for the alignment of the lexicon extensions. However, the computation and memory costs grow linearly with the number of lexical cues.

The method by Chen (1993) is a word-correspondence-based model that gives a better accuracy than length based methods, however, it was reported to be much slower than the algorithms of Brown et al., (1991) and Gale and Church (1993).

Moore's (2002) method aligns the corpus using a modified version of Brown et al.'s (1991) sentence-length-based model in the first pass. It then uses the sentence pairs that were assigned the highest probability of alignment to train a modified version of IBM Translation Model 1 (one of the five translation models that assigns a probability to each of the possible word-by-word alignments—developed by Brown et al. (1993)). The corpus is realigned, augmenting the initial alignment model with IBM Model 1, to produce an alignment based both on sentence length and word correspondences. It uses a novel search-pruning technique to efficiently find the sentence pairs that will be aligned with the highest probability without the use of anchor points or larger previously aligned units like paragraphs or sections. This is an effective method that gets a relatively high performance especially in precision. Nonetheless, this method has the drawback that it usually gets a low recall especially when dealing with sparse data (Trieu et al., 2015).

Hunalign sentence alignment method by Varga et al. (2005) uses a hybrid algorithm based on a length-based method that makes use of a bilingual dictionary. The similarity score between a source and a target sentence consists of two major components, which are token-based score and length-based score. The token-based score depends on the number of shared words in the two sentences while the length-based alignment is based on the character count of the sentence.

Varga et al.'s (2005) method uses a dictionary-based crude translation model instead of a full IBM translation model as used by Moore (2002). This has the very important advantage that it can exploit a bilingual lexicon, if one is available, and tune it according to frequencies in the target corpus. Moore's (2002) method offers no such way to tune a pre-existing language model. Moreover, the focus of Moore's (2002) algorithm on one-to-one alignments is less than optimal, since excluding one-to-many and many-to-many alignments may result in losing substantial amounts of aligned material if the two languages have different sentence structuring conventions (Varga et al., 2005).

Bleualign sentence aligner by Sennrich and Volk (2011) is based on the BLEU (bilingual evaluation understudy) score, which is an algorithm for evaluating the quality of text that has been machine-translated from one natural language to another. Instead of computing an alignment between the source and target text directly, this technique bases its alignment search on a Machine Translation (MT) of the source text.

The YASA method by Lamraoui and Langlais (2013) also operates a two-step process through the parallel data. Cognates are first recognized in order to accomplish a first token-level alignment that (efficiently) delimits a fruitful search space. Then, sentence alignment is performed on this reduced search space. The speed of the YASA aligner and memory use is comparatively better than Moore's (2002) aligner (Lamraoui and Langlais, 2013).

Though the method by Braune and Fraser (2010) is four times slower than Moore's (2002) method, it supports one to many and many to one alignments as well. It uses an improved pruning method and in the second pass, the sentences are optimally aligned and merged. This method uses a two-step clustering approach in the second pass of the alignment.

The method by Tóth et al. (2008) exploits the fact that Named Entities cannot be ignored from any translation process, so a sentence and its translation equivalent contain the same Named Entities.

The method by Mújdricza-Maydt et al. (2013) uses a two-step process to align sentences. Machine alignments known as "wood standard" annotations, produced using state-of-the-art sentence aligners in a first step, are used in a second step, to train a discriminative learner. This combination of arbitrary

amounts of machine aligned data and an expressive discriminative learner provides a boost in precision. All features used in the second step, with the exception of the POS agreement feature, are language-independent.

According to Gale and Church (1993) a considerably large parallel corpus having a small error percentage can be built without lexical constraints. According to the authors, lexical constraints might slow down the program and make it less useful in the first pass. Linguistic methods can produce better results if the performance of the system is not a concern. Hybrid methods such as that of Moore's (2002) that do not require particular knowledge about the corpus or the languages involved are faster as they tend to build the bilingual dictionary for aligning using the input to the aligner based on previous word-correspondence-based models.

Furthermore, results of some of the above methods such as Hunalign (Varga et al, 2005), Bleualign (Sennrich and Volk, 2011) and Gargantua (Braune and Fraser, 2010) could be improved by applying linguistic factors such as word forms, chunks and collocations (Navlea and Todiraşcu, 2010). Some have used morphologically processed (lemmatized and morphologically tagged) data and have used taggers (POS tagger) because it significantly increases the value of the data (Bojar et al, 2014).

2.4 Indic Languages

Automatic alignment of sentences has been attempted for few Indic language pairs from the South Asian subcontinent including Hindi-Urdu (Kaur and Kaur, 2012) and Hindi-Punjabi (Kumar and Goyal, 2010). This research used the method proposed by Gale and Church (1993) citing the close linguistic similarities between languages of these pairs, causing parallel sentences to be of similar lengths.

3 Methodology

3.1 Data Source

The parallel corpus used in aligning sentences is from annual reports published by different government departments in Sri Lanka. These government reports have been manually translated from Sinhala to Tamil by translators with different levels of experience in translation and Sinhala-Tamil competency. Thus the quality of the translations compared to other sources such as those from the Parliament of Sri Lanka is comparatively low with a considerable number of omissions and mistranslations.

These annual reports are in pdf format. Text was automatically extracted from the pdf documents, and converted to Unicode to ensure uniformity. The text thus obtained was segmented into sentences using a custom tokenization algorithm implemented specifically for Tamil and Sinhala.

Although there are some tokenizers for Sinhala[1] and Tamil, they could not be used for this purpose, since the abbreviations used in our input text are different from those in the existing tokenizers. Therefore we created a list of manually extracted abbreviations. Splitting documents into sentences was done by using delimiters such as " **.**, **?** , **!** ". Splitting into sentences using full stops is misleading at abbreviations, decimal digits, e-mails, URLs etc., because full stops at these places are not actual sentence boundaries. Therefore splitting into sentences at these points was avoided by means of regular expression checks. However issues such as omissions of punctuation marks result in the need for complex alignments (one to many, many to many).

For example[2] the following sentences in Sinhala specify five cities (Kuruwita, Rathnapura, Balangoda, Godakawela, Opanayake) followed by the sentence "The Active Committee representing the Operations Co-ordination Centers for Language Associations in Vavuniya was established".

(කුරුවිට,රත්නපුර,බලංගොඩ, ගොඩකවෙල, ඕපනායක).
වවුනියාව භාෂා සංගම් මෙහෙයුම් මධ්‍යස්ථාන ක්‍රියාකාරී කමිටුව ස්ථාපිත කරන ලදී.

However due to the omission of the period in the corresponding Tamil text, the above is identified as one single sentence in Tamil requiring the alignment to map one Tamil sentence to many Sinhala sentences.

(குருவிட்ட இரத்தினபுரி பலாங்கொடை கொடகவெல ஓபநாயக) வவுனியாவிலும் மாவட்ட மொழிச்சங்க செயற்பாட்டு குழு உருவாக்கப்பட்டது.

[1] https://github.com/madurangasiriwardena/corpus.sinhala.tools
[2] Text extracted from English, Sinhala and Tamil Annual Reports of a Government Department

The bilingual dictionary used for alignment was obtained from the trilingual dictionary[3] combined with the glossaries obtained from the Department of Official languages[4], Sri Lanka. The number of words in the lexicon obtained has around 90000 words, but it does not have all the commonly used words in the languages and mostly has the spoken forms of words in Sinhala, which are not used in the written official documents.

3.2 Sentence Alignment

Depending on the similarities and dissimilarities between the languages and the quality of the data source, different techniques discussed in section 2 have given different results for the alignment for different language pairs. For example, a method like that of Chuang and Yeh (2005) would work well for parallel text where punctuations are consistent, while that of Varga et al. (2005) would work better for languages that lack etymological relations. Thus the objective of this research is to experiment with these techniques for Sinhala-Tamil, and identify the best technique.

However, not all methods described in section 2 can be used in the context of Sinhala and Tamil. For example, methods by Tóth et al. (2008) and Mújdricza-Maydt et al. (2013) cannot be used because NER systems and comprehensive POS taggers are not fully developed for Sinhala (Dahanayaka and Weerasinghe, 2014; Manamini et al., 2016) and Tamil (Pandian et al., 2008; Vijayakrishna and Devi, 2008). Also methods that align using the punctuations in the two languages similar to that of Chuang and Yeh (2005) cannot be used in this case because when extracting text from pdf, some punctuations are lost, and also the translators of the original text have not been consistent with the use of punctuations.

Constrained by the available resources, we compared methods by Gale and Church (1993), Moore (2002), Varga et al. (2005), Braune and Fraser (2010), Lamraoui and Langlais (2013), and Sennrich and Volk (2011). These methods have shown promising results for languages that show close linguistic relationships, which is also the case with Sinhala and Tamil. These close linguistic relationships include similarities in word or sentence length, similarities in sentence structure and in languages that use the character set, similarities between words. Linguistic similarities between Sinhala and Tamil include word and sentence length similarities and sentence structure similarity with both Sinhala and Tamil following a Subject-Object-Verb structure.

The mean and variance for the number of Tamil characters per Sinhala was found and these values were used for the Gale and Church's (1993) method. Default values were used for the other methods during the evaluation.

For Moore's (2002) method, a bilingual word dictionary is built using the IBM Model 1. However, this dictionary may lack significant vocabulary when the input corpus contains sparse data, as pointed out by Trieu and Nguyen (2015). The output files from this method contain all the sentences from the input files that align 1-to-1 with probability greater than the "threshold" according to the statistical model computed by the aligner. For evaluation using this method we used a threshold of 0.8 instead of the default value of 0.5.

Around 1300 sentences were extracted from pdf files and were aligned using these methods. This corpus is publicly available[3] for the benefit of Sinhala and Tamil language computing. The same sentences were manually aligned with the help of a human translator. Then the automatically aligned sentences were compared with the manually aligned sentences to obtain the precision and recall values.

4 Evaluation

The evaluation for sentence alignment was done by using data that was manually aligned. The reason for this approach instead of getting the human translator to evaluate the automatically aligned sentences was to ensure that the manual evaluation was independent from the automatically produced output, as the automated alignments may influence the human aligner. Furthermore this approach also facilitated the comparison of the performance of multiple methods. Table 1 shows the precision, recall, and F-measure obtained for the six methods.

[3] http://www.trilingualdictionary.lk/
[4] http://www.languagesdept.gov.lk/

	Gale and Church (1993) (modified)	**Varga et al.'s (2005) (Hunalign)**	Sennrich and Volk's (2011) (BLEUalign)	Moore's (2002)	Braune and Fraser's (2010)	Lamraoui and Langlais's(YASA) (2013)
Precision	77.24%	**81.67%**	76.91%	94.56%	81.52%	80.62%
Recall	72.52%	**76.73%**	69.78%	67.56%	65.71%	76.53%
F-measure	74.8%	**79.1%**	73.2%	78.8%	72.8 %	78.5%

Table 1: Evaluation Results

5 Discussion

Most of the above methods (Gale and Church, 1993; Brown et al., 1991; Chen and S.F, 1993; Braune and Fraser, 2010) have been first used for English and French sentence alignment. Both these languages have many similarities, which include the sentence structure and the sentence length. The sentence structure of these languages is of the form subject-verb-object and the sentence length is quite close.

The same similarities can also be found in Sinhala and Tamil languages. Sinhala and Tamil languages have the same sentence structure, Subject-Object-Verb. Also the average sentence lengths of the two languages are quite close. Considering 700 sentences, average length of Sinhala is 113.76 and for Tamil it is 130.53. Therefore statistical methods have given good results in our case. The lexical components used in the hybrid methods suggested above are also language independent. Thus the hybrid methods are also applicable for Sinhala and Tamil.

We used Gale and Church (1993) method even though we could not align the paragraphs before aligning the sentences, due the dissimilarities among the text converted from pdfs. The length of Tamil sentences was comparatively higher than Sinhala sentences and the correlation between Sinhala and Tamil was comparatively low, hence we cannot consider mean and variance as language independent as suggested by Gale and Church (1993). Therefore we calculated the mean and variance for Sinhala and Tamil using 700 sentences. Gale and Church (1993) introduced 1 as mean and 6.8 as variance for English and French Languages. For Sinhala and Tamil, we figured out mean is 1.152 and variance is 1.860. Even after changing the parameters for Sinhala and Tamil in the Gale and Church (1993) method, we obtained a comparatively low precision because this method does not only look at one to one alignments but also one to zero, many to one, one to many or many to many alignments. Also according to Gale and Church (1993), in this method one to zero alignment is never handled correctly. Most misalignments arise due to one to zero, many to one to many or many to many alignments, resulting in methods that consider only one to one alignments to have better precision values. Given the nature of the source documents used in this research, there were a significant non one-to-one alignments and incorrect translations, which affected the precision value. However, as this method omits only a few sentences, it obtains high recall and F-Score than some of the other methods.

Since the text used for alignment in our case has considerably sparse data, the dictionary built in the Moore's (2002) method lacks significant vocabulary. Furthermore because of the fact that Moore's (2002) method only considers one to one alignment, the recall obtained by this method is very low while the precision is very high. In our case, even though there are alignments that are not one to one, the high precision of Moore's method has shown that it is possible to align a considerable number of sentences only by using one to one alignments. According to Moore (2002), in practice one to one alignments are the only alignments that are currently used for training machine translation systems.

The YASA aligner by Lamraoui and Langlais (2013) has proven to be robust to noise by having a good precision and recall for the parallel corpus of Sinhala and Tamil. Also the Braune and Fraser's (2010) method is known to work better especially for corpora where the sentences do not align one to one that often. However, our source text has a number of one to one alignments (as was proved by the alignment in Moore's (2002) method) along with other forms of alignments, which could be the reason for the low recall of this method.

Even though the method by Varga et al. (2005) has given the highest F-score, the results for this method could be improved using a better dictionary that includes all or most of the words that are used in the annual reports.

A factor significantly affecting the results of the alignment process was the quality of the source documents. Compared to other documents such as parliamentary documents, news articles and subtitles commonly used in evaluating alignment, the annual reports we considered were of comparatively less quality including significant omissions and inconsistencies and high complexity with significant many to one, one to many, and many to many alignments. The data set considered comprised of nearly 7% many to one, one to many or many to many alignments and nearly 15% one to zero or zero to one alignments indicating improper or incomplete translations.

6 Conclusion

We have addressed the problem of the lack of sentence aligned Sinhala-Tamil parallel corpus large enough to be useful in a multitude of natural language processing tasks. We have experimented with a number of alignment techniques developed for other language pairs, introducing necessary modifications for Sinhala and Tamil, where applicable.

The results generated have been satisfactory, indicating that better results could be obtained with more language resources such as morphological analyzers, POS taggers and named entity recognizers, which are currently not fully developed for Sinhala. This research is carried out as part of a major project to build a machine translation system between Sinhala and Tamil. POS taggers and named entity recognizers are being developed as part of this larger project. With the availability of these resources, methods utilizing these resources could also be introduced for Sinhala and Tamil in the near future, to obtain improved results.

Future work in improving the automatic generation of the Sinhala-Tamil parallel corpus includes experimenting with more techniques that have worked for other language pairs. The suitability of techniques that specifically use language resources such as POS taggers and morphological analysers could also be evaluated with the availability of such resources of better quality. Additionally the identified techniques could be evaluated with documents from different domains, whereas in this research evaluation has been done only with annual reports.

Acknowledgement

This research is part of a larger project at the Department of Computer Science and Engineering, University of Moratuwa, on developing a machine translation system for Sinhala and Tamil languages. We would like to extend our gratitude to the project team, and Mrs. Lalitha Peiris in particular, who did the manual translation of text. We would also like to thank the Department of Official Languages for providing us with the language resources.

References

Alkahtani, Saad, Wei Liu, and William J. Teahan. "A new hybrid metric for verifying parallel corpora of Arabic-English." *arXiv preprint arXiv:1502.03752* (2015).

Al-Onaizan, Yaser, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky. "Statistical machine translation." In *Final Report, JHU Summer Workshop*, vol. 30. 1999.

Anandan, P., K. Saravanan, RanjaniParthasarathi, and T. V. Geetha."Morphological analyzer for Tamil."In *International Conference on Natural language Processing*. 2002.

Bojar, Ondrej, VojtechDiatka, PavelRychlý, PavelStranák, VítSuchomel, Ales Tamchyna, and Daniel Zeman. "HindEnCorp-Hindi-English and Hindi-only Corpus for Machine Translation." In *Language Resources and Evaluation Conference*, pp. 3550-3555. 2014.

Braune, Fabienne, and Alexander Fraser. "Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora." In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 81-89.Association for Computational Linguistics, 2010.

Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer."Aligning sentences in parallel corpora."In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pp. 169-176.Association for Computational Linguistics, 1991.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. "The mathematics of statistical machine translation: Parameter estimation." *Computational linguistics* 19, no. 2 (1993): 263-311.

Callison-Burch, Chris, David Talbot,and Miles Osborne. 2004. Statistical machine translation with word-and sentence-aligned parallel corpora. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pages 176–183, Barcelona

Chen, Stanley F. "Aligning sentences in bilingual corpora using lexical information."In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pp. 9-16.Association for Computational Linguistics, 1993.

Chuang, Thomas C., and Kevin C. Yeh."Aligning parallel bilingual corpora statistically with punctuation criteria."*Computational Linguistics and Chinese Language Processing* 10, no. 1 (2005): 95-122.

Dahanayaka, J. K., and A. R. Weerasinghe."Named entity recognition for Sinhala language."In *2014 International Conference on Advances in ICT for Emerging Regions (ICTer)*, pp. 215-220. IEEE, 2014.

Gale, William A., and Kenneth W. Church. "A program for aligning sentences in bilingual corpora." *Computational linguistics* 19, no. 1 (1993): 75-102.

Hallebeek, Jos. "English parallel corpora and applications." *Cuadernos de FilologiaInglesa* 9, no. 1 (2000).

Herath, DulipLakmal, and A. R. Weerasinghe. "A Stochastic Part of Speech Tagger for Sinhala." In *Proceedings of the 06th International Information Technology Conference*, pp. 27-28. 2004.

Hettige, Budditha, and Asoka S. Karunananda."A Morphological analyzer to enable English to Sinhala Machine Translation."In *2006 International Conference on Information and Automation*, pp. 21-26.IEEE, 2006.

Kaur, Mandeep and NavdeepKaur. 2012. "Development And Analysis Of Hindi-Urdu Parallel Corpus". *International Journal Of Computing And Corporate Research* 2 (6).

Koehn, Philipp. *Statistical machine translation*.Cambridge University Press, 2009.

Kumar, Pardeep, and Vishal Goyal."Development of Hindi-Punjabi Parallel Corpus Using Existing Hindi-Punjabi Machine Translation System and Using Sentence Alignments."*Development* 5, no. 9 (2010).

Lamraoui, F. and Langlais, P., 2013. Yet another fast, robust and open source sentence aligner. Ttime to reconsider sentence alignment. *XIV Machine Translation Summit*.

Manamini, S. A. P. M., A. F. Ahamed, R. A. E. C. Rajapakshe, G. H. A. Reemal, S. Jayasena, G. V. Dias, and S. Ranathunga. "Ananya-a Named-Entity-Recognition (NER) system for Sinhala language."In 2016 *Moratuwa Engineering Research Conference (MERCon)*, pp. 30-35.IEEE, 2016.

McEnery, Tony, Andrew Wilson, Fernando Sanchez-Leon, and Amalio Nieto-Serrano. "Multilingual resources for European languages: contributions of the CRATER project." *Literary and Linguistic Computing* 12, no. 4 (1997): 219-226.

McEnery, Tony, and Michael Oakes."Sentence and word alignment in the CRATER project."*Using corpora for language research* (1996): 211-231.

Moore, Robert C. "Fast and accurate sentence alignment of bilingual corpora."In *Conference of the Association for Machine Translation in the Americas*, pp. 135-144.Springer Berlin Heidelberg, 2002.

Újdricza-Maydt, Éva, HuiqinKörkel-Qu, Stefan Riezler, and Sebastian Padó."High-precision sentence alignment by bootstrapping from wood standard annotations."*The Prague Bulletin of Mathematical Linguistics* 99 (2013): 5-16.

Navlea, Mirabela, and AmaliaTodiraşcu."Linguistic Resources for Factored Phrase-Based Statistical Machine Translation Systems." In *Proceedings of the Workshop on Exploitation of Multilingual Resources and Tools for Central and (South) Eastern European Languages, 7th International Conference on Language Resources and Evaluation*, pp. 41-48. 2010.

Pandian, S., Krishnan AravindPavithra, and T. Geetha. "Hybrid three-stage named entity recognizer for Tamil." *INFOS2008, March Cairo-Egypt. Available at: http://infos2008. fci. cu. edu. eg/infos/NLP_08_P045-052. pdf* (2008).

Paulussen, Hans, LieveMacken, Willy Vandeweghe, and Piet Desmet. "Dutch parallel corpus: A balanced parallel corpus for Dutch-English and Dutch-French." In *Essential Speech and language technology for Dutch*, pp. 185-199.Springer Berlin Heidelberg, 2013.

Thomas, Jenny, and Mick Short, eds. *Using corpora for language research*. London: Longman, 1996.

Sennrich, Rico, and Martin Volk."Iterative, MT-based sentence alignment of parallel texts."In *18th Nordic Conference of Computational Linguistics*. 2011.

Sim˜oes, Alberto Manuel Brand˜ao. 2004. Parallel corpora word alignment and applications. Master's thesis, Escola de Engenharia - Universidadedo Minho.

Tóth, Krisztina, RichárdFarkas, and AndrásKocsor."Sentence Alignment of Hungarian-English Parallel Corpora Using a Hybrid Algorithm."*ActaCybern.* 18, no. 3 (2008): 463-478.

Trieu, Long Hai, and Thai Phuong Nguyen."A New Feature to Improve Moore's Sentence Alignment Method."*VNU Journal of Science: Computer Science and Communication Engineering* 31, no. 1 (2015).

Varga, Dániel, LászlóNémeth, PéterHalácsy, AndrásKornai, Viktor Trón, and Viktor Nagy.Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing 2005 conference*, pages 590–596, Borovets, Bulgaria, 2005.

Vijayakrishna, R., and SobhaLalitha Devi. "Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields." In *International Joint Conference on Natural Language Processing*, pp. 59-66. 2008.

Wu, Dekai. "Aligning a parallel English-Chinese corpus statistically with lexical criteria."In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 80-87.Association for Computational Linguistics, 1994.

Yeong, Yin-Lai, Tien-Ping Tan, and SitiKhaotijah Mohammad."Using Dictionary and Lemmatizer to Improve Low Resource English-Malay Statistical Machine Translation System."*Procedia Computer Science* 81 (2016): 243-249.

Zotti, Patrizia, and Riccardo Apolloni Yuji Matsumoto."Sentence Alignment of a Japanese-Italian Parallel Corpus.Towards a web-based Interface." *Available at: http://www.anlp.jp/proceedings/annual_meeting/2014/pdf_dir/P1-6.pdf* (2014).