Software to Extract Parallel Data from English-Punjabi Comparable Corpora

Manpreet Singh Lehal[1*], Dr. Ajit Kumar[2], Dr. Vishal Goyal[3]
[1]Department of Computer Science, Lyallpur Khalsa College, Jalandhar
[2]Associate Professor, Department of Computer Science, Multani Mal Modi College, Patiala
[3]Department of Computer Science, Punjabi University, Patiala
Email: mslehal@lkc.ac.in {ajit8671, vishal.pup}@gmail.com

Abstract

Machine translation from English to Indian languages is always a difficult task due to the unavailability of a good quality corpus and morphological richness in the Indian languages. For a system to produce better translations, the size of the corpus should be huge. We have employed three similarity and distance measures for the research and developed a software to extract parallel data from comparable corpora automatically with high precision using minimal resources. The software works upon four algorithms. The three algorithms have been used for finding Cosine Similarity, Euclidean Distance Similarity and Jaccard Similarity. The fourth algorithm is to integrate the outputs of the three algorithms in order to improve the efficiency of the system.

1. Introduction

Machine translation from English to Indian languages is always a difficult task due to the unavailability of a good quality corpus and morphological richness in the Indian languages. For a system to produce better translations, the size of the corpus should be huge. In addition to that, the parallel sentences should convey similar meanings, and the sentences should cover different domains. Modelling the system with such a corpus can assure good translations while testing the model. Since English - Punjabi language pair is an under-resourced pair, this study provides a breakthrough in acquiring English - Punjabi Corpus for performing the task of machine translation. We have employed Statistical methods for the research and developed a software to extract parallel data from comparable corpora automatically with high precision using minimal resources.

We generate an English-Punjabi Comparable Corpora which is used as input data. We have used the articles from Wikipedia which are stored in the dump. The articles of English and Punjabi languages are extracted, aligned and refined. We also received access to the database of Indian Language Technology Proliferation and Deployment Centre (TDIL) and used the noisy parallel sentences. Sentences were also collected from Gyan Nidhi corpus and reports of college activities. Thus, our data is not restricted to one particular domain.

We employ three similarity measures of Cosine Similarity, Jaccard Distance and Euclidean Distance to find the similarity of two English Corpora. Firstly, the algorithms are performed individually and then the integrated approach is used by combining the results of all the three similarity measuring algorithms to reach better output levels. The software works upon four algorithms. The three algorithms have been used for finding Cosine Similarity, Euclidean Distance Similarity and Jaccard Similarity. The fourth algorithm is to integrate the outputs of the three algorithms in order to improve the efficiency of the system. The codes for similarity algorithms have been implemented in python using Scikit Learn. The sentences are first converted into vectors using tf-idf vectorization and then the algorithms are employed.

Every similarity measure has its own limitations when used individually. Combining the scores of three similarity measures complements the features and give better results. Only those

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 10–12
Patna, India, December 18 - 21, 2020.

translation pairs are selected which are similarly paired in all the three algorithms. The translation pairs which do not occur in the output of one or two algorithms are discarded.

We run the three similarity algorithms and obtain similarity scores. Threshold values are fixed by getting the average of all the similarity scores obtained for each algorithm. In case of Euclidean Distance and Jaccard Distance, the translation pairs having similarity scores below the threshold values are selected and in case of Cosine similarity translation pairs with similarity scores above the threshold value are selected. The remaining translation pairs are filtered out. This refines the output to a great extent.

The results obtained from the three algorithms are integrated aiming at the improvement of the output translation pairs and finding the best pairs. Only those translation pairs are selected which are similarly paired in all the three algorithms. The translation pairs which do not occur in the output of one or two algorithms are discarded.

With the integrated approach, we are able to achieve a precision level of 93 percent and accuracy is 86 percent. The results make it clear that the integrated approach improve the results to a great extent and thus, validate the usage of this approach.

There are three components of the web interface of the software: Punjabi Input, English Input and Aligned Data. The input data in form of sentences or paragraphs is copied in relevant language boxes on the left side and submitted. It gives the translation pair output in the Aligned data box on the right-side. The tuning button is used to identify translations at the required level of similarity. It can be increased or decreased to find exact parallel sentences as well as translations pairs similar at phrase level.

References

Afli, H., Barrault, L. and Schwenk, H. (2014) ‘Multimodal Comparable Corpora for Machine Translation’.

Artetxe, M. and Schwenk, H. (2019) ‘Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond’, Transactions of the Association for Computational Linguistics. doi: 10.1162/tacl_a_00288.

Bharadwaj, R. G. and Varma, V. (2011) ‘Language independent identification of parallel sentences using Wikipedia’, Proceedings of the 20th international conference companion on World wide web - WWW ’11, p. 11. doi: 10.1145/1963192.1963199.

Cettolo, M., Federico, M. and Bertoldi, N. (2010) ‘Mining Parallel Fragments from Comparable Texts’, Iwslt-2010, pp. 227–234.

Chu, C., Nakazawa, T. and Kurohashi, S. (2013) ‘Chinese – Japanese Parallel Sentence Extraction from Quasi – Comparable Corpora’, pp. 34–42.

Deep, K., Kumar, A. and Goyal, V. (2018) ‘Development of Punjabi-English (PunEng) Parallel Corpus for Machine Translation System’, International Journal of Engineering & Technology. doi: 10.14419/ijet.v7i2.10762.

Dwivedi S.K , Sukhadeve, P. P. (2010) ‘Machine Translation System in Indian Perspectives’, 6(10), pp. 1082–1087.

Dzmitry, B., Kyunghyun, C. and Yoshua, B. (2014) ‘Neural machine translation by jointly learning to align and translate’, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Eisele, A. and Xu, J. (2010) ‘Improving Machine Translation Performance Using Comparable Corpora’, in Proceedings of the 3rd Workshop on Building and Using Comparable Corpora. Workshop on Building and Using Comparable Corpora (BUCC-3), Applications of Parallel and Comparable Corpora in Natural Language Engineering and the Humanities, La Valletta, Malta, pp. 35–41.

Fu, X. et al. (2013) ‘Phrase-based Parallel Fragments Extraction from Comparable Corpora’, Proceedings of the Sixth International Joint Conference on Natural Language Processing, (October), pp. 972–976. Available at: http://aclweb.org/anthology/I13-1129.

Fung, P. et al. (2004) ‘Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM’, EMNLP 2004 - Conference on Empirical Methods in Natural Language Processing, pp. 57–63. Available at: http://www.aclweb.org/anthology-new/W/W04/W04-3208.pdf.

Goyal, V., Kumar, A. and Lehal, M. S. (2020) ‘Document Alignment for Generation of English-Punjabi Comparable Corpora from Wikipedia’, International Journal of E-Adoption. doi: 10.4018/ijea.2020010104.

HEWAVITHARANA, S. and VOGEL, S. (2016) ‘Extracting parallel phrases from comparable data

for machine translation', Natural Language Engineering. doi: 10.1017/s1351324916000139.

Jindal, S., Goyal, V. and Singh, J. (2017) 'Building English-Punjabi Parallel corpus for Machine Translation', International Journal of Computer Applications. doi: 10.5120/ijca2017916036.

Kumar, A. and Goyal, V. (2018) 'Hindi to Punjabi machine translation system based on statistical approach', Journal of Statistics and Management Systems. Taylor & Francis, 21(4), pp. 547–552.

Kvapil\'\iková, I. et al. (2020) 'Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining', in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 255–262.

Lehal, M. S., Kumar, A. and Goyal, V. (2018) 'Review of techniques for extraction of bilingual lexicon from comparable corpora', International Journal of Engineering and Technology(UAE). doi: 10.14419/ijet.v7i2.30.13456.

Lehal, M. S., Kumar, A. and Goyal, V. (2019) 'Comparative analysis of similarity measures for extraction of parallel data', International Journal of Control and Automation.

Munteanu, D. M. D. (2002) 'Processing comparable corpora with Bilingual Suffix Trees', Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10, pp. 289–295.

Pascale Fung, E. P. and S. S. (2010) 'Trillions of comparable documents', Proceedings of the 3rd workshop on building and using comparable corpora: from parallel to non-parallel corpora, (May), pp. 26–34.

Quirk, C., Udupa, R. and Menezes, A. (2007) 'Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction', Machine Translation Summit XI, (2000). Available at: http://www.mt-archive.info/MTS-2007-Quirk.pdf.

Rauf, S. A. and Schwenk, H. (2011) 'Parallel sentence generation from comparable corpora for improved SMT', Machine translation. Springer, 25(4), pp. 341–375.

Riesa, J. and Marcu, D. (2012) 'Automatic Parallel Fragment Extraction from Noisy Data', in Proc. NAACL, pp. 538–542.

Tillmann, C. (2009) 'A Beam-Search Extraction Algorithm for Comparable Data', Acl-2009, (August), p. 4. doi: 10.3115/1667583.1667653