{tag} {/tag}International Journal of Computer Applications
Foundation of Computer Science (FCS), NY, USA

Volume 180
-
Number 8

Year of Publication: 2017

Authors:

Shishpal Jindal, Vishal Goyal, Jaskarn Singh Bhullar

10.5120/ijca2017916036

{bibtex}2017916036.bib{/bibtex}

Abstract

Objective

Parallel corpus is the key resource for English Punjabi machine translation. At wide level there is no availability of English-Punjabi corpora. There is a primary requirement of parallel corpus for the training of statistical machine translation.

Methods/Analysis

In this paper, authors focus on building English-Punjabi corpus at large scale. It posed difficulties and the intensive labor to develop the corpus. We are intricate on the collection as well as the flow of work for the construction of parallel corpus. Now after getting the raw text, we need to refine the corpus in such a way that every source language sentence should have corresponding target language sentence.

Findings

The paper attempts to explore existing tools as well as building new tools. One of the goals is alignment of bilingual corpus. The alignment algorithms are used to tune the sentences. The accuracy depends on the type of corpus.

Novelty/Improvement

A cautious endeavor has been made to capture different types of texts.

References

1. P. Baker, A. Hardie, T. McEnery, R. Xiao, K. Bontcheva, H. Cunningham, R. Gaizauskas, O. Hamza, D. Maynard, V. Tablan, C. Ursu, B. D. Jayaram, M. Leisher, “Corpus linguistics and South Asian languages: corpus creation and tool development”, Literary Linguist. Comput. Vo. 19, pp. 509–524, 2004.
2. G. N. Jha, “The TDIL program and the Indian language corpora initiative (ILCI)”, Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC’10). European Language Resources Association, 2010.
3. N. Choudhary, “Web-drawn Corpus for Indian languages: a case of Hindi”, Proceedings of the ICISIL, vol. 139, pp. 218–223. 2011.
4. M. Shrivastava, P. Bhattacharyya, “Hindi POS tagger using naive stemming: harnessing morphological information without extensive Linguistic knowledge”, Proceedings of the International Conference on NLP (ICON08), 2008.
5. S. Dandapat, S. Sarkar, A. Basu, “Automatic part-of-speech tagging for Bengali: an approach for morphologically rich languages in a poor resource scenario”, Proceedings of the Association for Computational Linguistic, pp 221–224, 2007.
6. A. Bharati, D. M. Sharma, L. Bai, R. Sangal, “Anncorra: Annotating Corpora”, LTRC, IIIT, Hyderabad, 2006.
7. S. Baskaran, K. Bali, M. Choudhury, T. Bhattacharya, P. Bhattacharyya, G. N. Jha, S. Rajendran, K. Saravanan, L. Sobha, B. M. Subbarao, “A Commonparts-of-speech tag set framework for Indian languages”, Proceedings of the 6th International Language Resources and Evaluation (LREC’08), 2008.
8. V. Goyal, G. S. Lehal, “Hindi morphological analyzer and generator”, Proceedings of the 1st International Conference on Emerging Trends in Engineering and Technology, 2008.
9. T. Bögel, M. Butt, A. Hautli, S. Sulger, “Developing a finite-state morphological analyzer for Urdu and Hindi”, Proceedings of the 6th International Workshop on Finite-State Methods and Natural Language Processing, 2007.
10. V. Goyal and G. S. Lehal, “N-Grams Based Word Sense Disambiguation: A Case Study of Hindi to Punjabi Machine Translation System”, International Journal of Translation, Vol. 23(1), pp. 99-113, 2011.
11. V. Goyal and G. S. Lehal, “Advances in Machine Translation Systems”, Language In India, Vol. 9, pp. 138-150, 2010.
12. V. Goyal and G. S. Lahal, “Hindi Morphological Analyzer and Generator”, IEEE Computer Society Press, Washington, DC, USA 1156-1159, 2008.
13. P. Brown, S. A. D. Pietra, V. J. D. Pietra, R. L. Mercer, “The Mathematics of Statistical

Machine Translation: Parameter Estimation", Computational Linguistics, Vol. 19 (2), pp. 263-311, 1993.

14. V. B. Dang, and B. Ho, "Automatic Construction of English-Vietnamese Parallel Corpus through Web Mining", Proceedings of the International Conference on Innovation and Vision for the Future, pp..261-266, 2007.
15. www.sikhiwiki.org/index.php/Guru_Granth_Sahib
16. www.pseb.ac.in
17. www.christos-c.com/bible
18. www.tdil.mit.gov.in

Index Terms

Computer Science Information Systems

Keywords

Bilingual corpora, Machine-translation, English, Punjabi, NLP.