ECONSPEAK: A Journal of Advances in Management IT & Social Sciences,
Vol. 12, Issue 6, June 2022 Impact Factor 8.049 ISSN: (2231-4571)
www.skirec.org Email Id: skirec.org@gmail.com

A preprocessing Computational aid for Sanskrit Text Translation

Susanta Kumar Satapathy
Computer Teacher
Shree Sadashiva Campus
Puri

Introduction

Sanskrit, the classical language of India, presents specific challenges for computational linguistics: exact phonetic transcription in writing that obscures word boundaries, rich morphology and an enormous corpus, among others. Recent international cooperation has developed innovative solutions to these problems and significant resources for linguistic research. Solutions include efficient segmenting and tagging algorithms and dependency parsers based on constraint programming. The integration of lexical resources, text archives and linguistic software is achieved by distributed interoperable Web services. Resources include a morphological tagger and tagged corpus . Formal and computational linguistics was dominated by English at its inception and developed in subsequent decades primarily in the environment of European languages. More recently there has been a concerted effort to undertake formal linguistic analysis of a wide variety of languages, with particular interest in those with dramatically different features, and to enrich linguistic theory to account for linguistic variety. In spite of this effort, analytic structures and procedures utilized in formal linguistics remain dominated by those invented for, and most suitable for, English and other European languages. Linguistic theory remains unduly weighted in favor of European languages even as their extension to the variety of the world's languages involves undue complication thereby revealing their inadequacy in representing language universally. Machine Translation is one of the most challenging tasks in natural language processing. Statistical machine translation (SMT) looks into the translation of natural language as a machine learning problem. Since, the advent of globalization need for cross language translator has increased. English has emerged as most popular language on World Wide Web. The developing regions still strive to access the information in local languages. Translation of English into local languages can make information flow easier. This paper is on undergoing research for design and development of a cross language system from English to Sanskrit to make the same convenient.

Significant and Objectives

The basic objective of the paper is to illustrate with examples the divergence and adaptation mechanism in English to Sanskrit translation. This paper describes the Phrase-Based Statistical Machine Translation Decoder. Our goal is to improve the translation quality by enhancing the translation table and by preprocessing the source language text. research. We discuss the major design objective for the decoder, its performance relative to other SMT decoders . Phrase-based translation has been one of the major advances in statistical machine translation in recent years and is currently one of the techniques which can claim to be state-of-the-art in machine translation. Phrase-based models are a development of the word based models as exemplified by the In phrase-based translation, contiguous segments of words in the input sentence are mapped to contiguous segments of words in the output sentence. The main contribution of this paper is to show how we have created an extensible decoder, has acceptable run time performance compared to similar systems, and the ease of use and development that has made it the preferred choice for researchers looking for a phrase-based SMT decoder. Statistical language modeling is the science (and often art) of building models that estimate the prior probabilities of word strings. Language modeling has many applications in natural language technology and other areas where sequences of discrete objects play a role, with prominent roles in speech recognition and natural language tagging (including specialized tasks such as part-of-speech tagging, word and sentence segmentation, and shallow parsing). As pointed out , the main techniques for effective language modeling have been known for at least a decade, although one suspects that important advances are possible, and indeed needed, to bring about significant breakthroughs in the application areas cited above—such breakthroughs just have been very hard . Translational model allows us to enumerate possible structural relationship between pairs of strings. However , even within the constrains of strict model , the ambiguity of natural language results in a very large number of possible target sentences for any input source sentence . Our translation system needs a mechanism to choose between them . This mechanism comes from modeling : parameterization . We design a function that allows to assign a real valued score to any pair of source and target sentences. The general forms of these model are similar to those in other machine learning problems. The general forms of these models are similar to those in other machine learning problems.

Methodology

Now that we have a model and estimates for all of our parameters, we can translate new input words and sentences . This is called decoding. We call this the decision rule . The phrase-based decoder we developed for purpose of comparing different phrase based translation models employs a beam of search algorithm , similar to the one by. The Sanskrit out sentence is generated left to right in form of partial translations or hypotheses . We start with an initial empty hypothesis . A new hypothesis is expanded from an existing hypothesis by the translation phrase as follows. A sequence of un- translated foreign words and a possible English phrase translation for them is selected . The English phrase is attached to the existing English output sequence . The foreign words are marked as translated and the probability cost of hypothesis is updated . Phrase – based statistical machine translation approaches continue to dominate the filed of machine translation . The translation service makes use of state of the art phrase based SMT systems within the framework of feature based exponential model containing the following features .

- Phrase translation probability
- Inverse phrase translation probability
- Lexical weighting probability
- Inverse lexical weighting probability
- Phrase Penalty
- Language model probability
- Simple distance based deformation model
- Word penalty
- Image

Conclusion :-

Sanskrit is the communicator of an unbroken knowledge tradition from the vedic times to the present times. Modern Indian languages can benefit from profound knowledge of the texts of Indian intellectual tradition by being able to access these texts in a cost effective manner. Therefore automatic translation from Sanskrit to Indian languages is highly desirable. And no automatic translation from Sanskrit is possible without building such analysis tools. This decoder delivers a very good baseline system. This is capable of estimating parameters over a large development corpus in a reasonable time, thus it is able to generate highly relevant parameters. We have applied the sound software engineering principles and design to the implementation of the decoder which has enabled other researchers to use and extend its functionality. We believe this has been a major factor for the widespread adoption of Moses within the SMT community. The Example-Based Machine Translation is used in situations, where on-line resources (such as parser, morphological analyzer, rich bilingual dictionary, rich parallel corpora etc) are scarce. The Sanskrit is free word order language. Thus, we maintain a grammatical and semantic meaning for every sentence obtained by the change in the ordering of the words in the original sentence.

I hope that the design of the decoder will enable it to maintain it leading edge status into the future.

REFERENCE :-

[1] Philipp Koehn, Hieu Hoang, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, ―Moses: Open Source Toolkit for Statistical Machine Translation‖, Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177–180, June 2007.

[2] Brown, P. F., J. Cocke, et al. (1990). "A statistical approach to machine translation."

[3] Och, F. J. and H. Ney (2004). "The alignment template approach to statistical machine

translation." Computational Linguistics.

[4] Koehn, P. (2004). Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Models. AMTA.

[5] CHURCH, K. AND HOVY, E. 1993. Good applications for crummy machine translation. Mach. Transl. 8, 239–258.

[6] P. Clarkson and R. Rosenfeld, ―Statistical language modeling using the CMU-Cambridge toolkit‖, in G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, Proc. EUROSPEECH, vol. 1, pp. 2707–2710, Rhodes, Greece, Sep. 1997.

[7] F. Jelinek, ―Up from trigrams! The struggle for improved language models‖, in Proc. EUROSPEECH, pp. 1037–1040, Genova, Italy, Sep. 1991.

[8] R. Rosenfeld, ―Two decades of statistical language modeling: Where do we go from here?‖, Proceedings of the IEEE, vol. 88, 2000.

[9] MITCHELL, T. M. 1997. Machine Learning. McGraw-Hill.

[10] Jelinek, F. (1998). Statistical Methods for Speech Recognition. The MIT Press.

[11] Och, F. J., Ueffing, N., and Ney, H. (2001). An efficient A* search algorithm for statistical machine translation. In Data-Driven MT Workshop.

[12] Franz Josef Och and Hermann Ney. 2003. A system atic comparison of various statistical alignment mod els. Computational Linguistics, 29(1):19–51.

[13] Talbot, D. and M. Osborne (2007). Smoothed Bloom filter language models: Tera-Scale LMs on the Cheap. EMNLP, Prague, Czech Republic.

[14] http://www.gelbukh.com/polibits/37_05.pdf

[15] Carme Armentano-Oller , Rafael C. Carrasco , Antonio M. Corbí-Bellot , Mikel L. Forcada ,

Mireia Ginestí-Rosell , Sergio Ortiz-Rojas , Juan Antonio Pérez-Ortiz , Gema Ramírez-Sánchez , Felipe Sánchez-Martínez , Miriam A. Scalco, Open-Source portuguese–spanish machine translation, Proceedings of the 7th international conference on Computational Processing of the Portuguese Language, May 13-17, 2006, Itatiaia, Brazil [doi>10.1007/11751984_6]

[16] T. Dhanabalan et al, Tamil to UNL En-Converter, Proc. of ICUKL 2002, November 2002, Goa, India

[17] T. Dhanabalan, T. V. Geetha, UNL Deconverter for Tamil, International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies, Convergences 2003, December 2--6, 2003, Alexandria, EGYPT

[18] Natural Language Processing James Allen, Pages: 23--109 Pearson Educations.

[19] Neetu Mishra et al, An unsupervised Approach to Hindi Word Sense Disambiguation Pages: 327 HCI 2009.

[20] Statistical Machine Translation ADAM LOPEZACM Computing Surveys, Vol. 40, No. 3, Article 8, Publication date: August 2008.

[21] http://sanskrit.jnu.ac.in/subanta/Paper/MSPIL.pdf

****