Tutorial: Machine Learning for Machine Translation

Presenter: Dr. Pushpak Bhattacharyya, Professor, Department of Computer Science and Engineering, IIT Bombay.

Part of: ICON, 2013

Date: 18th December, 2013

Place: New Delhi

Duration: Full day (9.30 AM to 6.30PM)

Abstract:

Current times are witnessing an intense activity on Statistical Machine Translation (SMT) driven by parallel corpora. Machine Translation (MT) is the oldest branch of Natural Language Processing (NLP) which has been called the forcing function of even computer science.

In this tutorial we will delve into machine learning (ML) techniques applied to MT for various tasks starting from alignment to decoding to evaluation. ML makes use of the parallel corpora which is gradually becoming abundant in different languages. We will start with a brief history of MT, touching on the paradigms of interlingua and transfer based approaches. The foundational platform of Vauquois Triangle will be described, which provides the Analysis-Transfer-Generation (ATG) technique for MT. Then we will move onto the basic task of word alignment which makes use of Expectation Maximization (EM) to produce word mappings. IBM models 1, 2 and 3 will be described in this context. Next topic will be phrase based machine translation (PBMT), the ruling paradigm of SMT these days. However, PBMT often proves inadequate for free word order, morphologically rich languages. This motivates the discussion on Factor Based SMT (FBSMT). We discuss how introduction of factors reduces the need for parallel corpora for a given level of accuracy. FBSMT is deemed to be the future of SMT of Indian Languages (IL).

PBSMT and FBSMT give rise to more advanced techniques of SMT, namely, Hierarchical SMT (HSMT), Tree based SMT (TSMT) and Hybrid SMT (HySMT). These will be touched upon. Discussions on Decoding, comparable to parallel corpora and Evaluation will form the next part of the tutorial. We will conclude with observations on future of SMT, especially for Indian Languages and comparison with EBMT (example based machine translation).

Schedule:

9.30-10.00: Introduction- MT perspective, Vauquois Triangle, MT paradigms

10.00-11.30: Word Alignment- EM, Viterbi Alignment, IBM Models 1-3

11.30-12.00: Tea

12.00-13.30: Phrase Based SMT followed by Tree Based SMT; The Moses tool, The Joshua tool

13.30-14.30: Lunch

14.30-16.00: Factor Based SMT and SMT of Indian Languages with Hybrid Approaches

16.00-16.30: Tea

16.30-18.00: Complexity of Decoding, MT Evaluation, Comparable to Parallel Corpora

18.00-18.30: Comparison with Example Based MT (EBMT), Conclusions

There will demonstrations all along interspersed with lectures.

CV of the presenter:

Dr. Pushpak Bhattacharyya is a Professor of Computer Science and Engineering at the Indian Institute of Technology Bombay (IITB), where he has served for last 24 years, teaching, researching and executing national and international projects.

Dr. Bhattacharyya was educated at IIT Kharagpur (B.Tech), IIT Kanpur (M.Tech) and IIT Bombay (PhD). During PhD, he was a visiting scholar at MIT, Cambridge, USA. Subsequently he has been visiting professor at Stanford University (2004), University of Grenoble (2005, 2009 and 2011) and distinguished lecturer in University of Houston, USA (2012).

Prof. Bhattacharyya was the organizing chair of the Computational Linguistics Conference at IIT Bombay Dec, 2012. He also has been Associate Editor, ACM Transaction on Asian Language Information Processing (TALIP, 2010- till date). He has been PC member and area chair in ACL, COLING, IJCNLP, EMNP, GWC, ICON and so on.

Prof. Bhattacharyya has published extensively in top quality conferences and journals (about 200). He has guided 12 PhDs in AI and NLP and is currently guiding 10 PhD students. He has also advised close to 125 masters students and above 40 bachelor degree students for their research work. The research grants he has got from international and national agencies- government and industries included- have been substantial, with 15 completed and 8 ongoing projects in various areas of machine translation and search.

Prof. Bhattacharyya has been recipient of a number of prestigious awards and honors: Speaker as expert in multilingual computation in the prestigious Dagstuhl Seminar, Germany (2012), Yahoo Faculty Award (2011), Manthan Award (2009; given by Ministry of IT, India and Digital India Foundation), IIT Bombay's P. K. Patwardhan Award for Technology Development (2008), IBM Faculty Award (2007), Microsoft Distinguished Research Grant in a focused Area (2007) and United Nations Research Grant (1996).

Personal home page URL: http://www.cse.iitb.ac.in/~pb/