Mr. Rajen K. Chatterjee

EB2C24, San Bartolameo, via della Manpensada 90, 38123, Trento, Italy
Mobile: +39-331 479 4752 rajen.k.chatterjee@gmail.com DOB: 24th Jun., 1986

Summary

I am a final year PhD student at the University of Trento, and working on a European sponsored project at Fondazione Bruno Kessler, Italy. The ultimate goal of my research is to develop technologies that will ease communication within a multi-lingual environment. Applications directly benefiting from these technologies include business globalization, cross-lingual information retrieval, real-time translation among many others. More specifically, I am developing automatic post-editing systems based on deep learning that detect and correct errors in a machine translated document. Overall, I have six years of experience in machine translation with several publications at top tier conferences.

Skill Set

- Programming Languages: C, C++, Python, Java, OCTAVE, MATLAB
- Deep Learning Framework: Theano, Tensorflow
- Machine Translation Framework: MOSES
- Scripting Languages: JSP, JavaScript, jQuery, AJAX
- Operating System: Linux, Windows, Mac

Research Experience

July.'17 – Oct.'17 Amazon Cambridge, UK
Applied Scientist Intern, Alexa Machine Learning Team

Project: Improving Natural Language Understanding of Alexa

Oct.'14 – Present Fondazione Bruno Kessler Trento, TN, Italy
PhD student, Human Language Technology Group

Project: Human in the loop for advanced machine translation (QT-21 EU Project)

Goal: Leverage human post-edited feedback to build automatic post-editing (APE) system capable to automatically correct the recurring machine translation errors

Product: Flexible APE systems, which are portable across language pairs and domains, and are capable to model and customize to users' style and need

Methods: Applying divide and conquer strategy using factored translation models
Use of distributional semantics to address the problem of limited and sparse data
Exploring various deep learning architectures like ConvNets, Autoencoder, RNN, Inception, Residual networks to build neural APE system.

Publications: Exploring the Planet of the APEs, **ACL, 2015**
Online APE for MT in a Multi-Domain Translation Environment, **EACL, 2017**
Guiding Neural MT Decoding with External Knowledge, **WMT, 2017**

Multi-source Neural Automatic Post-Editing, **WMT, 2017**

Oct.'11 – Sept.'14 Indian Institute of Technology Mumbai, MH, India

Research Engineer, Natural Language Processing Group

Project: Developing Multilingual Resources for Indian Languages (collaboration with Xerox Research India)

Goal: Develop low cost and reliable solutions to generate parallel corpora for building machine translation systems

Product: TransDoop: A Map-Reduce based Crowdsourced Translation for Complex Domains

Methods: A Map-Reduce-like architecture to translation crowdsourcing, where sentence translation is decomposed into:
(a) translation of constituent phrases of the sentence;
(b) validation of quality of the phrase translations; and
(c) composition of complete sentence translations from phrase translation

Publications: TransDoop**:** A Map-Reduce based Crowdsourced Translation for Complex Domains. **ACL, 2013**

Project: English - Indian language machine translation

Goal: Build machine translation systems among low resource English and Indian languages to address the problem of language divergence and morphological issues

Product: Sata-Anuvadak : Tackling Multiway Translation of Indian Languages

Methods: Sata-Anuvadak is a compendium of 110 Statistical Machine Translation systems built from parallel corpora of 11 Indian languages based on phrase based translation model.
Language divergence is addressed by reordering the words (based on Supertag grammar) in the input sentence to match target language word order.
Morphological complexity is addressed by factored translation models based on divide and conquer strategy.
Explored various factors like POS tag, lemma, suffix, synset ID,
dependency labels, supertags in factor based SMT.

Publications: Sata-Anuvadak**:** Tackling Multiway Translation of Indian Languages. **LREC, 2014**

Achievements:

- Winner of the Automatic Post-Editing task @Second Conference on Machine Translation, 2017
- Young author distinguished paper award @Italian Conference on Computational Linguistics 2016
- Secured 1st rank in MSc IT and BSc IT

Activities:

- Organizing committee member:
  - Conference on Machine Translation, 2015/16/17
  - International Conference on Computational Linguistics, 2012
- Program committee member:
  - European Chapter of the Association for Computational Linguistics, 2017
  - Conference on Machine Translation, 2015/16/17
  - International Conference on Computational Linguistics, 2016

Education

2007-2009 K. C. College Mumbai, MH, India

Master of Science in Information Technology (M. Sc. I.T.)

Secured 1st Rank in College (72.20%, Distinction)

2004-2007 Elphinstone College Mumbai, MH, India

Bachelor of Science in Information Technology (B. Sc. I.T.)

Secured 1st Rank in College (79.10%, Distinction)

External Courses:

- Lisbon Machine Learning Summer School (LxMLS, 2015)
- Machine Learning Course on coursera (by Andrew Ng)

Applications Developed:

2013-14 Sata-Anuvadak I.I.T. Bombay

- This is a compendium of 110 Statistical Machine Translation systems built from parallel corpora of 11 Indian languages
- Learning of the translation models are based on phrase based model, and the decoding is based on log linear model
- Use of source side reordering helped to mitigate the problem of divergence between English and Indian languages
- Further, addition of the transliteration model significantly improved the translation quality
- Resources:
  - MOSES toolkit, Python

2012-13 TransDoop I.I.T. Bombay

- This is an automated tool for collecting human translation via Crowdsourcing.
- At the core, it is a three-stage pipelined architecture, where initially a document is split into sentences and then each sentence is split into phrases and saved in the database.
  - In stage 1, these phrases are floated as micro task to get their translation via third party crowdsourcing platform like Amazon Mechanical Turk (AMT). AMT then distributes these phrase translation task to the crowd. Once crowd solves this task the results are then retrieved back from AMT.

- In stage 2, we float these phrases along with their translation, which was obtained in stage 1, back to the crowd for rating. Once the rating is done, we retrieve back the result from AMT and move to stage 3.
- In stage 3, we select the translation pair with high rating and stich the translations together to form target sentence.

- Resources:
  - J2EE, Python, MySQL