Proposal for Tutorial on
“Translation & Transliteration between Related Languages”

Proposed by
Mitesh Khapra
- Researcher, IBM India Research Laboratory
Anoop Kunchukuttan
- Research Scholar, Center for Indian Language Technology, IIT Bombay

Under the direction of:
Prof. Pushpak Bhattacharyya,
IIT Patna
Abstract

Universal techniques for Machine Translation/Transliteration (MT/MX) have proven to be challenging to develop. However, a large chunk of MT/MX requirements is among related languages owing to government, business and socio-cultural communication needs (e.g. India, European Union). The proposed tutorial will discuss how the relatedness among languages can be leveraged to improve translation/transliteration quality, achieve better generalization, share linguistic resources, and reduce resource requirements. This tutorial is aimed at Machine Translation/Transliteration researchers and developers. The tutorial will also be relevant for researchers interested in multilingual computation, especially involving Indian languages.

We introduce notions of relatedness useful for MT/MX, and principles for leveraging relatedness. We explore how vocabulary shared between related languages can help MT. Then, we move beyond bilingual MT/MX and present how pivot-based and multisource methods incorporate knowledge from multiple languages, and handle language pairs lacking parallel corpora. We present approaches to multilingual word alignment, which show improvement over bilingual alignment. Finally, we discuss sharing of language resources (data & rules) among related languages, as well as among groups of related languages; thus, introducing the notion of a ‘language group’ being an apt level of system abstraction for building MT/MX systems.

Proposed Duration: Half Day
Tutorial Outline

1. Introduction: 15 slides, 30 minutes ([1],[2],[3])
   - Motivation
   - Brief introduction to Language Typology
   - Useful notions of language relatedness
   - Principles for leveraging relatedness

2. Taking advantage of orthographic similarity and cognates: 15 slides, 40 min
   - Transliteration & Cognate Mining ([4],[5],[8],[28])
   - Integrating transliteration & translation in decoder ([7])
   - Transliteration of OOV words ([9])
   - Character-level translation ([6])

3. Multilingual word alignment: 10 slides, 20 min ([19],[20],[21])

4. Multilingual phrase alignment: 20 slides, 40 minutes
   - Use of assisting & bridge languages ([12],[13],[27])
   - Pivot-based Methods ([10],[11],[14],[15])
   - Multi-source translation ([16],[17],[18])
   - Combining pivot-based SMT and transliteration methods

5. Sharing language resources: 8 slides, 15 min ([22],[23],[24])
   - Sharing among related languages
   - Sharing for translation between two groups of related languages

6. Conclusion & Future Directions: 4-5 slides, 10 min

7. Tools & Resources: 10 min ([24],[25],[26],[27])
   - Moses transliteration & transliteration mining system
   - System combination tools
   - Transliteration and Script conversion with the Indic NLP library

8. Question and Answer session: 10 min

Expected Duration: 180 minutes
Proposer Profiles:

Mitesh Khapra  
Researcher,  
IBM India Research Laboratory,  
Bangalore.  
mikhapra@in.ibm.com  

Mitesh Khapra obtained his Ph.D. from the Indian Institute of Technology, Bombay in the area of Natural Language Processing with a focus on **reusing resources for multilingual computation**. His areas of interest include Statistical Machine Translation, Text Analytics, Crowdsourcing, Argument Mining and Deep Learning. He is currently working as a researcher at IBM Research India where he is focusing on mining arguments from large unstructured text. He has co-authored papers in top NLP and ML conferences such as **ACL, NAACL, EMNLP, AAAI and NIPS**.

To view the complete publication list and presenter profile, please visit:  
[http://dblp.uni-trier.de/pers/hd/k/Khapra:Mitesh_M=](http://dblp.uni-trier.de/pers/hd/k/Khapra:Mitesh_M=)

Anoop Kunchukuttan  
Ph.D Scholar, Center for Indian Language Technology,  
Dept of Computer Science & Engineering, IIT Bombay  
anoopk@cse.iitb.ac.in  

Anoop Kunchukuttan is a research scholar at the Indian Institute of Technology Bombay. He is advised by Prof. Pushpak Bhattacharyya on his research work involving **machine translation and transliteration among related languages**. He has also investigated other NLP problems - multiword extraction, grammar correction, crowdsourcing and information extraction. He has co-authored papers in top NLP conferences such as **ACL, NAACL, CONLL, LREC, ICON**.

Prior to joining the Ph.D program, Anoop received his Masters degree in Computer Science and Engineering from IIT Bombay in 2008. He has worked in the **software industry for 4.5 years**, during which he led the development of large scale systems for information extraction and retrieval over medical text.

To view the complete publication list and presenter profile, please visit:  
[www.cse.iitb.ac.in/~anoopk](http://www.cse.iitb.ac.in/~anoopk)
References: