Statistical Machine Translation between Related Languages

Instructors: Pushpak Bhattacharyya, Mitesh Khapra, and Anoop Kunchukuttan

Prerequisites: Basic knowledge of statistical machine translation

Abstract:

Language-independent Statistical Machine Translation (SMT) has proven to be very challenging. The diversity of languages makes high accuracy difficult and requires substantial parallel corpus as well as linguistic resources (parsers, morph analyzers, etc.). An interesting observation is that a large chunk of machine translation (MT) requirements involve related languages. They are either : (i) between related languages, or (ii) between a lingua franca (like English) and a set of related languages. For instance, India, the European Union and South-East Asia have such translation requirements due to government, business and socio-cultural communication needs.

Related languages share a lot of linguistic features and the divergences among them are at a lower level of the NLP pipeline. The objective of the tutorial is to discuss how the relatedness among languages can be leveraged to bridge this language divergence thereby achieving some/all of these goals: (i) improving translation quality, (ii) achieving better generalization, (iii) sharing linguistic resources, and (iv) reducing resource requirements.

We will look at the existing research in SMT from the perspective of related languages, with the goal to build a toolbox of methods that are useful for translation between related languages. This tutorial would be relevant to Machine Translation researchers and developers, especially those interested in translation between low-resource languages which have resource-rich related languages. It will also be relevant for researchers interested in multilingual computation.

We start with a motivation for looking at the SMT problem from the perspective of related languages. We introduce notions of language relatedness useful for MT. We explore how lexical, morphological and syntactic similarity among related languages can help MT. Lexical similarity will receive special attention since related languages share a significant vocabulary in terms of cognates, loanwords, etc.

Then, we look beyond bilingual MT and present how pivot-based and multi-source methods incorporate knowledge from multiple languages, and handle language pairs lacking parallel corpora. We present some studies concerning the implications of languages relatedness to pivot-based SMT, and ways of handling language divergence in the pivot-based SMT scenario. Recent advances in deep learning have made it possible to train multi-language neural MT systems, which we think would be relevant to training between related languages.

We will summarize the tutorial by pointing out how the toolbox addresses the following goals we set out: (i) improving translation quality, (ii) achieving better generalization, (iii) sharing linguistic resources, and (iv) reducing resource requirements. We will conclude by emphasizing how the toolbox can be used to design translation system architectures customized to a set of related languages.

Time permitting, we will briefly describe a toolkit for Indian language NLP, which can be used to leverage similarities between Indian languages (http://anoopkunchukuttan.github.io/indic_nlp_library).

Outline:

- Introduction
- Motivation
- Important questions
- Useful notions of language relatedness
- Leveraging lexical similarity for translation:
- Phonetic and Orthographic Similarity
- Transliteration & Cognate Mining
- Integrating transliteration & translation in decoder
- Transliteration of OOV words
- Translation using Transliteration (character-level translation)
- Neural character-level translation
- Leveraging morphological and syntactic similarity for translation:
- Morphological Isomorphism
- Common source reordering solution for a set of related languages
- Synergy among languages:
- Pivot-based Methods
- Combining pivot-based and character-level SMT
- Multi-source translation
- Multi-lingual word alignment
- Multilingual translation with Neural MT

About the Instructors:

Dr. Pushpak Bhattacharyya

Professor, Dept. of Computer Science & Engineering, Indian Institute of Technology Bombay

Mumbai, India.

e-mail: pb@cse.iitb.ac.in

Website: https://www.cse.iitb.ac.in/~pb

Dr. Pushpak Bhattacharyya is Vijay and Sita Vashee Chair Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology Bombay (IITB) where he heads the Center for Natural Language Processing. He is also the Director of Indian Institute of Technology Patna.

Dr. Bhattacharyya obtained his Ph.D from IIT Bombay. His areas of interest cover a broad spectrum of problems in Natural Language Processing like machine translation, cross-lingual search, sentiment analysis - specially with reference to Indian languages.

Dr. Bhattacharyya has published extensively in top quality conferences and journals (about 200). He has also written a textbook on machine translation. He has advised 12 PhDs in NLP and ML, and is currently supervising 10 PhD students. He has also advised close to 125 masters students and above 40 bachelor degree students for their research work.

Dr. Mitesh Khapra

Researcher, IBM India Research Laboratory,

Bangalore, India.

e-mail: mikhapra@in.ibm.com

Website: http://researcher.watson.ibm.com/researcher/view.php?person=in-mikhapra

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 NLP and ML conferences such as ACL, NAACL, EMNLP, AAAI and NIPS.

Anoop Kunchukuttan

Ph.D Scholar, Center for Indian Language Technology,

Dept. of Computer Science & Engineering, Indian Institute of Technology Bombay

Mumbai, India.

e-mail: anoopk@cse.iitb.ac.in

Website: www.cse.iitb.ac.in/~anoopk

Anoop Kunchukuttan is a senior Ph.D student 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 NLP conferences such as ACL, NAACL, CONLL, LREC, ICON. He has worked in the software industry for about 5 years, during which he led the development of large scale systems for information extraction and retrieval over medical text. He completed his M.Tech in Computer Science & Engineering from IIT Bombay.