Invited Presentation

Machine Translation in Low Resource Setting

Pushpak Bhattacharyya
Computer Science and Engineering Department
Indian Institute of Technology Bombay

Abstract

AI now and in future will have to grapple continuously with the problem of low resource. AI will increasingly be ML intensive. But ML needs data often with annotation. However, annotation is costly.

Over the years, through work on multiple problems, we have developed insight into how to do language processing in low resource setting. Following 6 methods—individually and in combination—seem to be the way forward:

1. Artificially augment resource (e.g. subwords)

2. Cooperative NLP (e.g., pivot in MT)

3. Linguistic embellishment (e.g. factor based MT, source reordering)

4. Joint Modeling (e.g., Coref and NER, Sentiment and Emotion: each task helping the other to either boost accuracy or reduce resource requirement)

5. Multimodality (e.g., eye tracking based NLP, also picture+text+speech based Sentiment Analysis)

6. Cross Lingual Embedding (e.g., embedding from multiple languages helping MT, close to 2 above)

The present talk will focus on low resource machine translation. We describe the use of techniques from the above list and bring home the seriousness and methodology of doing Machine Translation in low resource settings.