Issues and challenges of NLP in relation to Canada's Aboriginal languages

UQÀM
Fatiha Sadat*, Tan Ngoc Le* and David Huggins Daines ♠
*Department of Computer Science, UQAM, Montréal, QC, Canada
♠ Nuance communications

Introduction

- NLP, a sub-field of AI, is a multidisciplinary field that aims to create tools and linguistic resources for various applications.
- These resources include emotion and sentiment analysis, speech analysis, machine translation, information extraction, prediction, etc.
- Our concern in this research program is related to **endangered languages** and the **preservation and revitalization of North American indigenous languages.**

Motivations

- **Linguistic resource construction**
  - Preprocessing schemes
  - Morphological analyzer
- **Automatic hybrid machine translation**
  - Rule-based + Zero-shot NMT
- Intermediate outcomes explanations and providing reasoning for the proposed solutions.
- **Intelligent tutorial system**
  - Learning and teaching the indigenous languages
- **Other NLP applications** such as
  - Sentiment analysis towards some topics such as climate changes
  - Question-answering / dialogue system

Challenges

- **Polysynthetic languages**: typically have "sentence-words" and highly inflected languages
- Studying a very rich and complex morphology and learning distributed word representations
- extremely low resource languages
- handling the out-of vocabulary, using multiple modalities, etc.
- **Rule-based systems**
- Study the achievement of **NMT** when using **extremely low resource languages**

Approaches

- Study on **Inuktitut** (Hansard corpus) and other languages (**Innu, Cree**)
- Zero-shot NMT + rules
- Integrating External Knowledge

Multi-task Transfer and Lifelong Learning: transferring and learning continually

- Multi-sentence Understanding
- **Hybridization**: an architecture that can capture all of the above components

Conclusion

- Promising long-term research program