先端科学技術研究科　修士論文要旨

所属研究室 (主指導教員)	自然言語処理学 (渡辺　太郎 (教授))		
学籍番号	2011406	提出日	令和 4年 7月 25日
学生氏名	田口　智大		
論文題目	Natural Language Processing for Lower-Resource Code-Switching Languages: Case Studies on Transliteration and Resource Building for Tatar		

要旨

While recent rapid development in machine learning has boosted the study and technologies in Natural Language Processing (NLP), its core contributions have been centered around a few higher-resourced languages, and lower-resourced, socially minoritized languages are often neglected.
Given this situation, this thesis provides two case studies of applying NLP technologies to Tatar, a lower-resource code-switching language spoken in Tatarstan, Russia. The first study demonstrates machine learning-based transliteration of Tatar from the Cyrillic alphabet to the Latin orthography utilizing Byte-Pair encoding for subword tokenization and language classification for detecting words code-switching with Russian.
Since Russian words require different transliteration rules, and code-switching in Tatar can occur inside a word, it is necessary to predict a language for each segment. Existing tools employ rule-based transliteration, but they do not always distinguish code-switching and are prone to mistakes.
The present study shows that the proposed automatic transliteration outscores the existing tools in accuracy. The second study presents a more down-to-earth approach to create annotated language resources for the code-switching language. We report our new Tatar treebank NMCTT (NAIST Multilingual Corpus TaTar), and propose a new way to annotate code-switching segments and their corresponding language tags in the framework of Universal Dependencies (UD), a unified syntactic annotation for any languages of the world. The experiments show that including the proposed code-switching information in UD annotation is beneficial for tasks such as code-switching segmentation and segment-level language tagging when combined with other grammatical information available in UD.
These studies support that both macroscopic language engineering and microscopic resource building are crucial and effective for the application of NLP technologies to code-switching minority languages.