氏名	KAROL PIOTR NOWAKOWSKI
授与学位	博士（工学）
学位記番号	博甲第184号
学位授与年月日	令和２年９月７日
学位授与の要件	学位規則第４条第１項
学位論文題目	Development of a Digital Corpus and Core Language Technologies for the Ainu Language （アイヌ語の電子コーパス及び基盤言語技術の開発における研究）
論文審査委員	主査　教授　桝井　文人 准教授　プタシンスキ　ミハウ　エドムンド 教授　前田　康成 教授　阿部　良夫 准教授　鈴木　育男

学位論文内容の要旨

Technologies being developed within the field of Natural Language Processing (NLP) have an important role to play in the urgent tasks of documenting, analyzing and revitalizing endangered languages. At the same time, the rapid development and the spread of language-related technologies observed in recent decades may result in creating a technological gap between smaller languages and majority languages – which in turn will threaten the survival of the former group – if linguistic minorities are not provided with equal access to said technologies. Among such languages is Ainu, a critically endangered language isolate native to northern parts of the Japanese archipelago. In this dissertation, I present the results of my research devoted to the development of Natural Language Processing technologies for the Ainu language.

I begin with presenting the background of this research. In particular, I provide an overview of previous research in the field of Natural Language Processing for under-resourced and endangered languages. I then describe the characteristics and current situation of the Ainu language, and review some of the related research on the Ainu language, including the few existing studies in the field of Natural Language Processing. I also describe the major challenges facing Ainu language research and revitalization efforts.

After that, I report the results of a survey conducted on a group of Ainu language learners, scholars, and other people interested in the Ainu language, with the aim of evaluating their experiences and needs concerning language-related technologies. Based on the findings from the questionnaire, I outline future research goals.

A major obstacle for the application of advanced language processing technologies to minority languages, including Ainu, is the lack of high-volume digital linguistic resources, such as text and speech corpora and in particular, annotated corpora. One of the main contributions of the presented research is the compilation of a large-scale corpus of Ainu. It covers a wide range of documents in a consistent structure, allowing for the application of corpus-driven approaches to NLP, such as statistical language models and representation learning techniques.

The above-mentioned corpus is then utilized in the development of Natural Language Processing tools for Ainu, namely word segmentation models and part-of-speech taggers. For the first task, I propose a novel algorithm: MiNgMatch Segmenter. I argue that the problem of identifying word boundaries can be reduced to finding the shortest sequence of lexical n-grams matching the input text, thus reducing its computational cost and increasing efficiency of the process. I perform a series of experiments comparing the algorithm with systems utilizing state-of-the-art statistical language modelling techniques, as well as a neural model performing word segmentation as character sequence labelling. The experimental results demonstrate high performance of the proposed approach, comparable with the other best-performing models. Next, I apply a state-of-the-art generator of sequential taggers – SVMTool – and a tagger based on Artificial Neural Networks, equipped with word representations inferred from the corpus, in the task of automatic part-of-speech annotation. Evaluation results reveal that they perform better than the dictionary-based system proposed in previous research (POST-AL), especially when applied to out-of-domain data.

Furthermore, I describe a preliminary Ainu language conversational program for the Pepper robot, which serves as a proof of concept of how robots could support Ainu language education. The proposed robot can hold simple conversations, teach new words and play interactive games using the Ainu language. In a group of Ainu language experts and experienced learners whom I asked for feedback (in the form of a survey study), the majority supported the idea of developing an Ainu-speaking robot and using it in language teaching.

Advances in cross-lingual learning indicate that the problems facing Natural Language Processing for low-resource languages can be, to a certain extent, alleviated by transferring knowledge from resource-rich languages. Unsurprisingly, however, such techniques tend to yield the best results for closely related languages, whereas the Ainu language is a language isolate, with no known cognates. Nevertheless, given the similarity of phonological systems and some grammatical constructions between Ainu and Japanese, it may still be beneficial to use the existing Japanese resources as a starting point in the development of language processing technologies for Ainu. In this thesis, I describe two preliminary experiments in cross-lingual knowledge transfer from Japanese to Ainu: firstly, I propose a method for generating useful word representations by using an Ainu-Japanese parallel corpus with morpho-syntactic annotations on the Japanese side. When applied to the problem of part-of-speech tagging of Ainu text, the proposed method contributes positively to the performance of a neural tagger. Secondly, I investigate the performance of Japanese models for Speech Synthesis and Speech Recognition in generating and detecting speech in the Ainu language. I perform human evaluation of the robot's speech in terms of intelligibility and pronunciation, as well as automatic evaluation of Speech Recognition. Experiment results suggest that cross-lingual transfer from Japanese has a potential to facilitate the development of speech technologies for Ainu, especially in the case of Speech Recognition. I also discuss main areas for improvement.

論文審査結果の要旨

世界的絶滅危機言語であるアイヌ語を対象として、主要言語に対して整備されているものと同等の自然言語処理技術の開発と整備を行った。具体的には、電子化コーパス（言語資源）の構築計算機処理のための表記の書き換えと標準化処理、トークン処理、品詞タグ付け処理、単語分割処理、音声対話処理を開発し、評価実験によって各処理の有効性を検証した。各処理の実現においては、辞書知識、n-gramモデルなどの統計処理、SVMやニューラルネットワークなどの機械学習技術などがアイヌ語の解析にどのように寄与するかを詳細に比較分析することで訓練データが十分に確保できないデメリットを回避しながら性能を確保することに成功した。

これを要するに、著者は、自然言語処理において長らく課題となっていた絶滅危機言語に関する大規模な言語資源の構築と計算機環境を活用した自動解析、自動翻訳の可能性に関する新知見を得たものであり、自然言語処理において少数の言語資源のみで高精度な言語処理推論機構を実現し得る可能性の拡大、及び絶滅危機言語の保存や分析を対象とする言語学や社会学において自然言語処理が貢献する範囲拡大に対して貢献するところ大なるものがある。よって著者は、北見工業大学博士（工学）の学位を授与される資格があるものと認める。