A Toolkit for Analysis of Ainu Language

Michal Ptaszynski †, Fumito Masui †, and Yoshio Momouchi ‡

†Department of Computer Science, Kitami Institute of Technology, 090-8507, Japan
{ptaszynski,masui}@cs.kitami-it.ac.jp
‡Professor Emeritus at Hokkai-Gakuen University, 064-0926, Japan

1 Introduction

Ainu language is a language of Ainu people living in northern Japan. It is critically endangered. However, it has a large heritage including myths, stories and poetry. In ongoing numerous linguistic research these resources have been analyzed, till now by hand. We present a toolkit for the most necessary linguistic analysis to help Ainu language researchers and translators.

2 Toolkit Description

We based the toolkit on a handcrafted dictionary developed especially to reflect unique features of Ainu parts of speech model. The dictionary contains such information as tokens, parts of speech (POS), translation (in Japanese), reference to the story it appears in, and usage examples. The toolkit consists of a number of tools. **Tokenizer** uses a Dictionary Lookup method performed according to the Longest Match Principle. **POS tagger** [1], based on CON-POST (*Contextual Part of Speech Tagging*) method was trained on dictionary examples using higher order Hidden-Markov Model (HMM), in which a given word is analyzed with respect to two or more words preceding or succeeding it (trigrams and longer). POS can be printed in one of three most anticipated POS naming standards for Ainu language. **Morphological analyzer** uses recursive analysis to extract and analyze each compound word and applies the HMM model for further word sense disambiguation. **Translation support tool** provides translations of tokens. It uses CON-ToT (*Contextual Token Translation*) method, where the translation is selected specifically for the word selected in POS tagging. The translations contain additional information (subject, object, etc.), useful in linguistic analysis, but hindering readability of the output. We added an option to turn off this information. The simplified forms could be useful in training a machine translation system for Ainu language in the future. Finally, a rule-based **shallow parser** applies POS information to divide a sentence into clauses and further into noun phrases and verb phrases.

References

1. Ptaszynski, M., Momouchi, Y.: Part-of-Speech Tagger for Ainu Language Based on Higher Order Hidden Markov Model. Exp. Syst. Appl., 39(14), 11576-11582 (2012)