A Toolkit for Analysis of Ainu Language

Introduction

- Ainu language is a language of Ainu people, mostly living in northern Japan.
- Population of Ainu ~ about 23 thousand people.
- Number of native speakers ~ less than hundred (Hohnem, 2008).
- Ainu language is critically endangered (Moseley, 2010).

Purpose of this research:

- Create language analysis tool including POS tagger, translation support tool and parallel corpus.
- Help in linguistic and language anthropology research and support translation of Ainu texts.
- Contribute to the process of reviving Ainu language.

Previous Research on Ainu Language

- linguistic studies: • collections of Ainu epic stories and myths (Chiri, Kayano, Shigeru, 1987; Kayano, 1995; Kayano, 1998; Ptaszynski and Majewicz, 2004)
• grammar descriptions (Chiri, 1974; Murasaki, 1979; Retsing, 1986; Kidaichi, 1993; Saito, 2008).

- relation studies: • attempt to develop Ainu language dictionary into an online dictionary (Bugaeva, 2010).
• automatically gather word translations from texts (Eichzen-ya et al., 2004).
• analysis / retrieval of hierarchical Ainu-Japanese translations (Azumi and Momouchi, 2008).
• annotating Ainu “yukar” stories for machine translation system (Momouchi et al., 2008).

- Morphological Analysis

- automatically gather word translations from texts

- POS taggers

- based on higher order hidden Markov model

- Contextual token translation

- based on longest match principle

- Description

- higher examples

- POST

- enhanced base

- Help in linguistic and language anthropology

• all stories are tokenized by (Kirikae, 2003).
• one yukar is annotated with POS tags and translations (by Momouchi et al., 2008).

Yukar 10: Pon Okikirimuy yoyeyukar “Kuritsuna kutunukkon”
(The “Kuritsuna kutunukkon” story told by Small Okikirimuy himself)

Evaluation Dataset Description

- transform dictionary information to XML database: 1 token (word, morpheme, etc.).
- 2 part of speech.
- 3 meaning (in Japanese).
- 4 usage examples (not for all cases).
- 5 reference to the story it appears in (not for all cases).

Evaluation

- Score Calculation

- Calculate score as balanced

- F1 score for all parts of POST-AL

- Tokens

- Evaluation

- Token translation

- Contextual token translation

- Precision Recall F-score

- Tokenization

- POS Tagging

- POS tagging

- S-POST / CON-POST

- output

- System Description

- POS standard selection

- Simple (Japanese): no POS dependency

- English: English POS tag names

- Abbreviated: abbreviated English POS tag names

- Output

- Options

- 1. POS standard selection

- □ Nakagawa (1995) / compact

- □ Tamura (1998) / sophisticated


- □ Simple (Japanese): no POS dependency

- □ English: English POS tag names

- □ Abbreviated: abbreviated English POS tag names

- View Selection

- [default] Vertical (typical for POS taggers)

- (horizontal) (useful for language anthropologists)

- Additional options: Additional information

- References