Abstract

We present POST-AL, a part-of-speech tagger for Ainu language. This system uses a hand-crafted dictionary and performs three tasks: tokenization, part of speech tagging, and token translation (to Japanese). The system is evaluated on 13 Ainu stories called "yukar". The system could be useful in a number of fields related to the research on Ainu language, such as content analysis or translation, which till now have been done mostly manually.

Introduction

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

Purpose of this research:
- To contribute to the reviving of Ainu language.
- To create part of speech tagger for Ainu language.
- To contribute in the field of any kind of language-related research.

Previous Research on Ainu language

- collections of Ainu epic stories and myths (Chiri, 1978; Kayano, 1998; Plutsikus and Majewicz, 2004).
- grammar descriptions (Chiri, 1974; Murasaki, 1979; Refsing, 1986; Kindaichi, 1993; Sato, 2008).
- NLP-related Studies:
  - attempt to transform Ainu language dictionary into an online database (Bugaenko, 2010).
  - annotating Ainu “yukar” stories for machine translation system (Momouchi et al., 2008).
  - a system for translation of Ainu topological names (Momouchi and Kobayashi, 2010).

Dictionary Construction

- POS tagger is annotated with POS and token tags.
- System Description
  - Tokenization:
    - POS Tagging:
      - Statistical Part of Speech Tagging.
      - S-POST (Contextual Token Translation) based on higher order HMM trained on dictionary examples.
  - Token translation:
    - RAN-TOT (Random Token Translation) based on the list of words selected from the same POS (S-POST extension).

Evaluation

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

Yukar 10: Pon Okikurmy yuyoyueru “Kutnisu kutukutun”
(“The Kutnisu kutukutun” story told by Small Okikurmy himself)

Score Calculation

F-score for all parts of POST-AL

Results

Tokenization

- DL-LSM was slightly better (98.22%) than statistical (97.94%).

POS tagging

- Contextual POS tagging was much better (98.62%) than statistical (90.11%).

Token translation

- Contextual token translation was much better (98.36%) than statistical (90.11%).

Conclusions and Future Work

- POST-AL is the first POS tagger for Ainu language.
- The system performs three main tasks: tokenization, part-of-speech tagging, and token translation. The results were around 97-98%.
- Output of POST-AL can be presented in one of three POS standards for Ainu language, with either vertical or horizontal view.

In the near future we plan to:
- compare different tokenization approaches (ex. Huang et al., 2007).
- add other dictionaries (Nakagawa, 1998; Tamura, 1998).
- add English translations (Batchelor, 1905) to make the tool usable also for non-Japanese speaking researchers.
- perform a robust evaluation of the annotations with the help of several experts and Ainu native speakers.
- Bootstrap the system for even better performance.
- Apply POST-AL to machine translation research.

References