Comparing Multiple Dictionaries to Improve Part-of-Speech Tagging of Ainu Language

Michal Ptaszynski † Karol Nowakowski ‡ Yoshio Momouchi § Fumito Masui †

† Department of Computer Science, Kitami Institute of Technology
{ptaszynski,f-masui}@cs.kitami-it.ac.jp

‡ Independent Researcher
karol_nowakowski@interia.pl

§ Professor Emeritus of Hokkai Gakuen University

Abstract

In this paper we present our research in improving POST-AL, a POS tagger for Ainu language. We focused on improving its dictionary base by comparing various Ainu language resources. We discuss the differences between the dictionaries and evaluate the system when each dictionary is applied. The experiments indicate that the size of the dictionary is less important than the way it was created. Moreover, dictionary records should be refined and unified.

1 Introduction

Ainu language is a unique language isolate spoken by the Ainu people[1], mostly living on northern parts of Japan extending to Sakhalin. However, despite the place of their inhabitation, Ainu people are genetically not related to peoples of Asia, such as modern Mongoloids (like Chinese) [1]. Similarly to the uniqueness of their genetic origin, the language spoken by the Ainu people has remained unique in its origin with no proof proposed showing its similarity to any other known world language [2].

Although the official estimate of the population of Ainu people is 23 to 25 thousand people [3, 4], the latest estimate of the number of people who can fluently use the language in conversation is less than hundred [5], and the language is considered as an endangered one. There have been numerous research on the language done from the point of view of linguistics and anthropolinguistics, aimed to describe, analyze and therefore preserve the language. Unfortunately, there have been only a few attempts have been made to to process the language computationally. Therefore the present research contributes to the task of reviving and revitalizing the Ainu language with the use of Natural Language Processing techniques.

In particular, we aimed at improving the only available tool for computer-supported Ainu language processing, namely, **POST-AL**, or ***P**art **o**f **S**peech **T**agger for **A**inu **L**anguage*, developed by Ptaszynski et al. (2012) [6]. We focused on improving the dictionary base of the POST-AL system by obtaining and comparing other available Ainu language resources.

In the following sections we firstly present the POST-AL system used in this research (section 2), describe all dictionaries used in this research (section 3), and evaluate POST-AL with the use of each different dictionary (section 4). Finally, we conclude the paper and propose some ideas for further improvements (section 5).

[1]The word "ainu" in the Ainu language means "a person".

2 POS Tagger for Ainu Language

POST-AL, or ***P**art **o**f **S**peech **T**agger for **A**inu **L**anguage*, is a tool developed by Ptaszynski et al. (2012) [6]. It is the first and so far only tool for analysis of Ainu language. POST-AL performs three main tasks: tokenization, part-of-speech (POS) tagging and token translation.

In this paper we focused on improving the POS tagging function of the system. Previously, POST-AL used a database created on only one dictionary, namely "Lexicon to Yukie Chiri's Ainu Shin-yōsyū (Ainu Songs of Gods)" by Kirikae (2003) [7]. In present research we attempt to improve the system performance by modifying and extending its original dictionary base. To do this, we either obtain or develop anew other Ainu language resources. In the following section we describe the applied dictionaries.

3 Overview of Applied Dictionaries

3.1 Lexicon to Ainu Songs of Gods

The base dictionary originally used in POST-AL was *Ainu shin-yōshū jiten* (Lexicon to Yukie Chiri's Ainu Shin-yōsyū (Ainu Songs of Gods)) by Kirikae (2003) [7] (later abbreviated to KK). It is one of the newest Ainu language dictionaries with a firm part-of-speech classification developed especially to reflect the differences between Ainu parts of speech model to models existing in other languages. Therefore except POS names like proper nouns or verbs, one can find examples rare or not existing in other languages, such as "interrogative indefinite adverb", like *hempara*, "demonstrative adverbs", like *ene* or *neno*, "postpositive adverb", like *ari, epitta* or *kama*, "nominal particles", such as *i, kur* or *p*, or "count verbs".

The dictionary contains 2,019 entries, each of it containing five types of information: token (word, morpheme, etc.), part of speech (POS), meaning (in Japanese), reference to the story it appears in, and usage examples (not for all cases).

ORIGINAL ENTRY:

```
<name>sat poro pet</name>      ←place name
      <pos>vi vi n</pos>       ←part-of-speech
      <tr>dry big river</tr>   ←translation
```

ENTRIES AFTER MODIFICATION:

```
<word>sat</word>               ←word entry
      <pos>vi</pos>
      <tr>dry</tr>             ↓ usage example
      <ex>sat poro pet:vi vi n:
            :dry big river</ex>
      <ex>sat poro:vi vi:dry big</ex>
      ⋮
<word>poro</word>
      <pos>vt</pos>
      <tr>flow</tr>
      <ex>sat poro pet:vi vi n:
            :dry big river</ex>
      <ex>ha poro pet:vt vi n:
            :flow big river</ex>
```

Figure 1: Example of modification of original Place Name Dictionary [8] for application in POST-AL (vi=intransitive verb, vt=transitive verb, n=noun).

3.2 Ainu Place Names Dictionary

Momouchi and Kobayashi (2010) [8], in their research on developing a system for translation of Ainu topological names, created a dictionary of Ainu place names in a form of a database. Originally the dictionary consists of 1,282 entries, each containing three types of information: tokenized transcription in roman alphabet, part-of-speech, and Japanese translation of the place name. However, the dictionary was not applicable in our research in its original form, since the dictionary entries often consisted of multiple tokens. Therefore we modified the dictionary to make it applicable in POST-AL. We tokenized all dictionary entries and used each token as a separate entry in our modified version of the dictionary. The original entries that consisted of more than one token were used as usage examples, which POST-AL uses for POS disambiguation. The modified dictionary (later abbreviated to `M-PL`) contained 873 unique entries with place names as usage examples. An example of modification is presented in Figure 1.

3.3 Yukar 10-13 Bootstrapped Dictionary

Apart from developing the place name dictionary, Momouchi, with Azumi and Kadoya (2008) [9] began a process of annotating Ainu "yukar" stories for the need of developing a machine translation system for Ainu language. One of their annotated stories, namely *Pon Okikirmuy yayeyukar "kutnisa kutunkutun"* (The "Kutnisa kutunkutun" story told by Small Okikirmuy himself) was used by Ptaszynski et al. [6] in their evaluation experiment of POST-AL.

At present there exist four additional annotated *yukar* stories from the collection by Chiri (1978) [12]. The annotation was performed using a bootstrapping technique. At first, *yukar 10* was annotated fully manually. The dictionary generated from the annotations was used to annotate *yukar 11*. Then errors were corrected and missing annotations were added by hand. Again, a dictionary was generated from stories 10 and 11, and used to annotate story 12. The process was repeated until story 13 was fully annotated.

Dictionary (later abbreviated to `M-BT`) generated from *yukar* stories 10-13 contains 422 entries, each of the entry containing such information as word (token), POS, and Japanese translation. Additionally Momouchi et al. [9] added POS information for Japanese meaning for further comparative analysis of the two languages.

3.4 Ainu Conversational Dictionary

Ainugo kaiwa jiten (Ainu conversational dictionary) [14] is one of the first dictionaries for Ainu language collected by a Japanese researcher. Shōzaburo Kanazawa, with help of Kotora Jinbo, collected it firstly around 1895 and 1897, right after Piłsudski's first collection [16], and a few years before Batchelor published his first Ainu-English-Japanese dictionary [10]. The dictionary was reprinted several times, with the most recent reprint dating on 1986.

In its present form, the dictionary contains 3,839 entries. For the need of the present research we have developed two versions of the dictionary. First one, using the original contents of the dictionary, second one, modified in a similar way to Momouchi's Ainu Place Names Dictionary [8]. Namely, for entries containing more than one word, such as phrases and short sentences, we have divided the entries and created separate one-word entries. Thus in the following sections we will refer to the Jinbo-Kanazawa dictionary in two contexts, its original form (abbreviated to `JK-or`), and further modified with one-word reference entries (later `JK-1w`). Additionally, we have also added alphabet transcriptions and English translations provided by Bugaeva and Endo [11].

4 Evaluation Experiment

4.1 Dataset Description

As the dataset for evaluation we used a collection of 13 Ainu stories (yukar) included in *Ainu shin-yōshū* (Ainu Songs of Gods) gathered by Chiri (1978) [12]. At present five yukar have been annotated with POS by expert annotators (Momouchi et al. [9]), in particular, stories from 9 to 13 form the collection gathered by Chiri.

In the evaluation we used yukar 9 and yukar 10, the latter also applied previously in evaluation of POST-AL by Ptaszynski et al. [6]. However, for the present experiment we did not use stories 11-13, although they were available. It was done due to the fact that one of the dictionaries (`M-BT`) was bootstrap-generated on the basis of those stories. Therefore it would most probably achieve the highest results. We used Yukar 10 only for confirmation of this fact and to keep similar evaluation settings as in previous research.

Table 1: Comparison of all dictionaries applied in the research.

Dictionary	# of entries	Included information					
		word	morpheme	POS	translation	examples	additional information
JK-1w [14]	12,855	◯	*◯	◯	JP/EN	N/A	N/A
JK-or [14]	3,839	◯	*◯	◯	JP/EN	N/A	N/A
KK [7]	2,019	◯	◯	◯	JP	◯	referring *yukar* story
M-PL [8]	873* (1,282)	◯	◯	◯	JP	**◯	N/A
M-BT [9]	422	◯	*◯	◯	JP	N/A	POS for Japanese

*equal to "word"; **after modification;

4.2 Experiment Setup

Although POST-AL is equipped with several functions (tokenization, token translation, etc.), in the present evaluation experiment we focused only on POS tagging and the effect of using different dictionaries on its performance. Contents of the dictionaries other than related to POS and its disambiguation, such as meaning (Japanese translation of the dictionary entry), or references to the story a record appears in were not taken into account. This means that, even if the dictionary contained incorrect translation of the entry, but the POS information was correct, the output was considered positive.

All results were calculated with the means of Precision (P), Recall (R) and balanced F-score (F), standard score calculation methods used in tasks such as POS tagging. Precision is the percentage showing how many annotations made by the system were correct. It is calculated as in equation 1. Recall is the percentage showing how many correct annotations the system made comparing to a gold standard. It is calculated as in equation 2. The balanced F-score is a harmonic mean of the two values. It is calculated as in equation 3. All results are represented in Table 2.

$$P = \frac{correct\ \ annotations}{all\ \ system's\ \ annotations} \tag{1}$$

$$R = \frac{correct\ \ annotations}{all\ \ gold\ \ standard\ \ annotations} \tag{2}$$

$$F_1 = 2\frac{P * R}{P + R} \tag{3}$$

4.3 Results and Discussion

The results of the comparison of POS tagging performance was as follows. The highest and the most balanced results were achieved with the use of Kirikae's dictionary (KK). This dictionary, used also originally by Ptaszynski et al. [6], was based on the Yukar stories. Therefore it was predictable that its results would score as one of the highest. Equally good performance was achieved by the Bootstrapped Dictionary (M-BT). It was also based on Yukar stories, however, the process of its creation differed from Kirikae's dictionary. In particular, Kirikae developed his dictionary fully manually, while the Bootstrapped Dictionary was created half-automatically. The fact that for the story used in bootstrapping the system achieved higher F-score than Kirikae's means that this method of dictionary creation is valid and promising. We also checked, why for yukar 10 the system did not achieve 100% of F-score. POST-AL uses higher order Hidden Markov Model trained on usage examples to disambiguate the parts of speech. However, for some words the examples are not available, and thus the disambiguation is performed based on statistics. Therefore, parts of speech, which represent some words less often are prone to cause errors. A solution to this would be to add more usage examples in the dictionary or continue the bootstrapping method to improve the quality of the dictionary. In the future it would also be useful to check the system performance on Ainu texts other than yukar stories.

An interesting result was presented by the Jinbo-Kanazawa Dictionary (JK). Although it was the largest of the applied dictionaries it did not achieve the highest scores. Even modifying the dictionary and expanding its coverage of the dictionary over four times (3 thousands from JK-or expanded to nearly 13 thousands in JK-1w), although improving the Recall (29% on average improved to 61%), did not noticeably improve the Precision. This could suggest that the dictionary is of good quality in general, but its coverage does not fully overlap with yukar stories.

The lowest results were achieved by the Ainu Place Names Dictionary (M-PL). It is reasonable, as it is the smallest dictionary. However, the fact that, despite the small size, the dictionary allows achieving close to 50% of Recall is interesting in its own. This confirms previous observation by Ptaszynski that the majority of topological names in Ainu language is directly derived from everyday vocabulary. For example, the name Sapporo (city name), derived from Ainu name *sat poro pet* (see Table 1), means "dry, great river". Ptaszynski et al. [6] consider this an interesting discovery, since it shows a striking resemblance to how Native Americans created topological names. For example, the city name Ohio in USA is derived from Iroquoian, where it means "great river".

5 Conclusions and Future Work

In this paper we presented our study in comparing multiple dictionaries for the support of part-of-speech tagging in Ainu language. We collected five different dictionaries and checked the POS tagging performance of the state-of-the-art POS tagger for Ainu language, POST-AL, when each of the dictionary was applied. We found out that its is not the size of the dictionary that makes the difference, but the way the dictionary was created (e.g., on data similar to the analyzed contents). We also confirmed that bootstrapping method is

Table 2: Results of POS tagging for different dictionaries.

Dictionary	Yukar 09			Yukar 10			Average		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
JK-1w	87%	61%	72%	88%	61%	72%	88%	61%	72%
JK-or	85%	29%	44%	88%	28%	42%	87%	29%	43%
KK	95%	97%	96%	92%	98%	95%	94%	98%	96%
M-PL	15%	42%	22%	24%	46%	31%	20%	44%	27%
M-BT	83%	73%	78%	93%	100%	96%	88%	87%	87%

useful in semiautomatic construction of such dictionaries.

In the future we plan to further enlarge the database by adding other dictionaries, such as the one by Nakagawa (1995) [15], or Tamura (1998) [17]. We also plan to add English translations, e.g., from Batchelor (1905) [10] to make the tool usable also for non-Japanese speaking researchers. Some of the additional translations have already been made available to the public by Bugaeva and Endo [11] for the dictionary by Jinbo and Kanazawa [14].

As the next step in this research we also plan to combine various dictionaries to check how joining various language resources influences the performance of POS tagging, and other functions, such as tokenization and word-to-word translation.

References

[1] Margaret Sleeboom. 2004. *Academic Nations in China and Japan*. Routledge: UK.

[2] Masayoshi Shibatani. 1990. *The Languages of Japan*. Cambridge university Press, London.

[3] Colin Ronald Baker, Sylvia Prys Jones. 1998. *Encyclopedia/Bilingualism/Bili*. Multilingual Matters. p. 163.

[4] Anna Bugaeva. 2010. Internet Applications for Endangered Languages: A Talking Dictionary of Ainu. *Waseda Institute for Advanced Study Research Bulletin*, No.3, pp. 73-81.

[5] Skye Hohmann. 2008. The Ainu's modern struggle. In *World Watch*, Vol 21., No. 6, pp. 20E4.

[6] Michal Ptaszynski and Yoshio Momouchi. 2012. Part-of-Speech Tagger for Ainu Language Based on Higher Order Hidden Markov Model, *Expert Systems With Applications*, Vol. 39, Issue 14 (2012), pp. 11576-11582, 2012.

[7] Hideo Kirikae. 2003. *Ainu shin-yōshū jiten: tekisuto bumpō kaisetsu tsuki* (Lexicon to Yukie Chiri's Ainu Shin-yōsyū (Ainu Songs of Gods) with Text and Grammatital Notes) [In Japanese]. Published by Daigaku Shorin.

[8] Yoshio Momouchi and Ryosuke Kobayashi. 2010. Dictionaries and Analysis Tools for the Componential Analysis of Ainu Place Name [In Japanese]. *Engineering Research: The Bulletin of Graduate School of Engineering at Hokkai-Gakuen University*, No.10, pp.39-49.

[9] Yoshio Momouchi, Yasunori Azumi and Yukio Kadoya. 2008. Research Note: Construction and Utilization of Electronic Data for "Ainu Shin-yōsyū" [In Japanese]. Bulletin of the Faculty of Engineering at Hokkai-Gakuen University, No. 35, pp. 159-171.

[10] John Batchelor. 1905. *An Ainu-English-Japanese dictionary (including a grammar of the Ainu language)*. Tokyo Methodist Pub. House.

[11] Anna Bugaeva and Shiho Endo (eds.), speaker: Setsu Kurokawa, multimedia developer: David Nathan. 2010. A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational dictionary. SOAS, University of London http://lah.soas.ac.uk/projects/ainu/

[12] Yukie Chiri. 1978. *Ainu shin-yōshū*. Tokyo, Iwanami Shoten.

[13] Katsunobu Izutsu. 2006. Ainu Parts of Speech Revisited: with Special Reference to So-called Personal Pronouns [in Japanese]. *Journal of Hokkaido University of Education (Humanities and Social Sciences)*, Vol. 56, No. 2, pp. 13-27.

[14] Kotora Jinbo and Shouzaburo Kanazawa. 1986. *Ainugo kaiwa jiten* (Ainu conversational dictionary) [In Japanese]. Sapporo: *Hokkaido publication project center*. First edition: 1898, Tokyo: Kinkōdo press.

[15] Hiroshi Nakagawa. 1995. *Ainugo Chitose Hōgen Jiten: The Ainu-Japanese Dictionary: Chitose Dialect* [In Japanese]. Sōfūkan.

[16] Bronisław Piłsudski (Author), Alfred F. Majewicz (Editor). 2004. *The Collected Works of Bronislaw Pilsudski: Materials for the Study of the Ainu Language and Folklore*, v.3, Pt. 2: Materials for the Study of the Ainu, (Trends in Linguistics: Documentation). Mouton de Gruyter (Oct 2004)

[17] Suzuko Tamura. 1998. *Ainugo Chitose Hōgen Jiten: The Ainu-Japanese Dictionary: Saru Dialect* [In Japanese]. Sōfūkan.