POST-AL: Part-of-Speech Tagger for Ainu Language

Michal Ptaszynski †　　Yoshio Momouchi ‡

† JSPS Research Fellow / High-Tech Research Center, Hokkai-Gakuen University
ptaszynski@hgu.jp

‡ Department of Electronics and Information Engineering,
Faculty of Engineering, Hokkai-Gakuen University
momouchi@eli.hokkai-s-u.ac.jp

Abstract

This paper presents POST-AL, a part-of-speech tagger for Ainu language. The 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 tasks related to the research on Ainu language, such as content analysis or translation, which till now have been done mostly manually.

1 Introduction

It is estimated that there are about 6000 to 7000 languages spoken in the world today. Half of them are endangered, with a probability of extinction by the year 2050. Ainu language is recognized among them as one of the most critically endangered, one step prior to extinction [1]. Ainu language is a language of Ainu[1] people, mostly living in northern parts of Japan. The population of Ainu is estimated on about 23 thousand people [2]. However, the latest estimate of the number of native speakers of Ainu (people who can fluently use Ainu language in conversation) is strikingly less than a hundred [3]. For several years there has been a noticeable movement to preserve and revive the Ainu language. By this research we wish to contribute to the reviving of Ainu language. There have been numerous research on the language done from the point of view of linguistics and anthropological linguistics, and only a few attempts to process the language computationally. In this research we propose the first part of speech tagger for the Ainu language, a tool which could become useful in any kind of language-related research.

The paper outline is as follows. In section 2 we describe some of the previous research on Ainu language from the points of view of linguistics and anthropological linguistics, and the few research approaching the Ainu language from the view of computational linguistics. Section 3 presents the dictionary we used as the base for the POS tagger and how it was transformed from a written form into a database. In section 4 we describe all components of the system, including tokenization, POS tagging and token translation. Section 5 presents the evaluation of the system. Finally, in section 6, we conclude the paper, propose some ideas to improve the system as well as possible applications.

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

2 Previous Research

Some of the first research on Ainu language are dated on the end of 19th century. It was performed by Bronisław Piłsudski, a Polish cultural anthropologist. Piłsudski studied Ainu culture and language, and prepared some of the first glossaries [4]. A few years later Batchelor [5] published his *Ainu-English-Japanese Dictionary*. Among linguistic research done in modern times, most consist of collections of Ainu epic stories and myths [4, 6], dictionaries and lexicons [7, 8, 9], and grammar descriptions [10, 11, 13, 12]. As for the research in NLP, there have been mostly only two. In the first one, Bugaeva [2] describes an attempt to transform Ainu language dictionary into an online database. The second research is an attempt done by Momouchi and colleagues to create a machine translation system for Ainu (to Japanese). Azumi and Momouchi [14, 15] prepared ground for analysis and retrieval of hierarchical Ainu-Japanese translations. Momouchi et al. [16] began a process of annotating Ainu "yukar" stories for the need of machine translation system. At present they performed annotations of one story, namely *Pon Okikirmuy yayeyukar "kutnisa kutunkutun"* (The "Kutnisa kutunkutun" story told by Small Okikirmuy himself). Lastly, Momouchi and Kobayashi [17] began creation of a system for translation of Ainu place names.

3 Dictionary Construction

Most of what remained of the language till present are transcribed narratives, such as *yukar* (epic stories), or *uwepeker* (old stories). The exact number of *yukar* and other poetry has not been estimated, although it is counted in thousands. Most of Ainu language studies have been based on analysis of this kind of narratives.

Therefore as the base dictionary for POST-AL we used

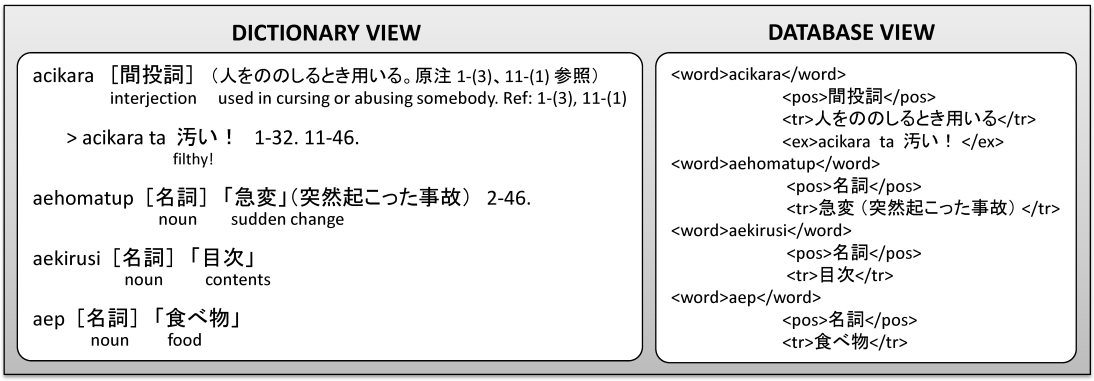

Figure 1: An example showing part of the dictionary and how it appears in the database.

the one solely based on analysis of yukar, namely *Ainu shin-yōshū jiten* (Lexicon to Yukie Chiri's Ainu Shin-yōsyū (Ainu Songs of Gods)) by Kirikae [9]. It is one of the newest Ainu 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*, "nominal particles", such as *i, kur* or *p*, or "count verbs". The last one represents a feature called pluractionality, recognized in less than ten other languages in the world, which expresses plural form of action or object of action (e.g., *kor* "to have [something]" as opposed to *kor-pa* "to have plenty of [something]").

The dictionary was transformed into an XML database using dictionary source files provided by the author of the dictionary. The original text of the dictionary contains five types of information: token (word, morpheme, etc.), part of speech, meaning (in Japanese), reference to the story it appears in and usage examples (the latest two not for all cases) In POST-AL we used all the above information, except the reference which is irrelevant to part-of-speech tagging. An example showing part of the dictionary and how it appears in the database is represented in figure 1.

4 POST-AL: System Description

4.1 Tokenization

The problem preceding the part-of-speech tagging task is tokenizatin. Tokenization is a process in which the text is separated into tokens. In general tokens consist of words and punctuation marks. In languages such as English, where the writing system assumes separating words by spaces, the tokenization process is of less difficulty. On the other hand, languages such as Chinese or Japanese are not spaced, which makes the tokenization process crucial to and inseparable from POS tagging. For Ainu language the situation is even more complex. Ainu language was only a spoken language before it was studied by researchers. It did not have a writing system. The written transcripts appeared only after the studies on the language began. Texts in the Ainu language, which usually include stories and narratives, most often appear in their printed form either undivided, or with chunks of text separated with a caesura (pause in recitation within one line of a poem). Therefore we needed to apply a tokenization method to be able to perform POS tagging. The method we applied is based on a standard approach to tokenization, namely dictionary lookup (DL). In the DL method one performs tokenization by matching all words in the lexicon to the untokenized string of text. In the method, called **DL-LSM:** (*Dictionary Lookup with Longest String Matching*), the input text is firstly glued together disregarding any potential separations and caesura. Then the dictionary lookup is performed according to the Longest Match Principe, which assumes that the matching is done beginning with the longest words in the lexicon, and ending on the shortest ones.

4.2 Part-of-Speech Tagging

We developed and compared two methods for part-of-speech tagging. The first one is based on statistics of parts of speech in the lexicon. The second one, based on a higher order HMM, is using n-grams as contextual information for the processed word.

S-POST: (*Statistical Part of Speech Tagging*) In this method all words of the same lexical form in the dictionary are treated as a separate list. The parts of speech describing the words are counted and the part of speech with the highest occurrence is set as the most probable tag for the given word.

CON-POST: (*Contextual Part of Speech Tagging*) This method uses a standard approach to POS tagging based on a higher order Hidden-Markov Model (HMM). HMM is a model in which a given word is analyzed with respect to the word preceding or succeeding it (bigrams). A higher order HMM is taking into account not one, but two or more succeeding words (trigrams and longer). We

trained the HMM model on the examples that appear in the original dictionary on which the system is based.

4.3 Token Translation

An additional task we included in the system is translating the tokens annotated with POS. Although this is not a standard feature of POS taggers, we assumed in this case it is more than necessary. There is only a very small number of people who understand the Ainu language without using a dictionary. Moreover, a great number of the preserved Ainu narratives is only transcribed, and the process of translating is important. Therefore we added the option of translating the annotated tokens. However, it must be noticed that we do not claim to propose a machine translation system for Ainu. What we aim to offer by this option is a support for translators, who will be given an automatic glossary lookup for the translated text. The token translation option could also be useful in training the actual machine translation system for the Ainu language [14, 15, 16, 17].

The translations of tokens are selected from the lexicon. We compared two methods for selecting the translations: random and contextual. They work similarly when there is only one token available for translation, but differ in dealing with cases of ambiguity.

RAN-ToT: (*Random Token Translation*) This method is used as an extension of S-POST. The translation is selected randomly from the list of words having the same annotated POS feature but different meaning (for example, in English the word "table" can have only one verb meaning but at least two noun meanings (related to either "furniture" or "information categorization") according to the Cambridge Advanced Learner's Dictionary & Thesaurus by Cambridge University Press[2]).

CON-ToT: (*Contextual Token Translation*) This method is the extension of CON-POST. The translation is selected specifically for the word selected in the contextual part-of-speech tagging, based on Hidden Markov Model trained on the dictionary examples.

As an option we also added two possible versions of output. The first one, vertical, typical for POS taggers, and the second one, horizontal, more readable and familiar to language anthropologists studying Ainu language. The two types of output are represented in figure 2.

5 Evaluation

5.1 Dataset Description

As the dataset for evaluation of the system we used 13 Ainu stories (yukar) included in *Ainu shin-yōshū* (Ainu Songs of Gods) gathered by Chiri [6]. The stories have been partially processed by Kirikae [9]. Kirikae added the tokenization of the stories according to linguistic, not

[2]http://dictionary.cambridge.org/

Sentence:
Ci nukar wa ci eramesinne pet esoro hosippa as .
Translation:
When I saw this I was relieved and came back with the river current.

Vertical output

ci	人称接辞, 意味:私（たち） [I/we]
nukar	他動詞, 意味:見る [see]
wa	接続助詞, 意味:て [conj.]
ci	人称接辞, 意味:私（たち） [I/we]
eramesinne	他動詞, 意味:安心する [be relieved]
pet	名詞, 意味:川 [river]
esoro	他動詞, 意味:沿うて下る [swim down with the current]
hosippa	自動詞, 意味:戻る [come back]
as	人称接辞, 意味:私 [I]

Horizontal output
ci nukar wa ci eramesinne pet esoro hosippa as .
人称接辞 他動詞 接続助詞 人称接辞 他動詞 名詞 他動詞 自動詞 人称接辞 ピリオド
私（たち） 見る て 私（たち） 安心する 川 沿うて下る 戻る 私 .

Figure 2: The two types of output in POST-AL.

poetic rules as it was originally in Chiri. Therefore this set is ideal for evaluating firstly the tokenization performance in POST-AL. Unfortunately, Kirikae did not annotate parts of speech on the stories. At present there exists only one yukar annotated with POS, namely Yukar 10: *Pon Okikirmuy yayeyukar "kutnisa kutunkutun"* (The "Kutnisa kutunkutun" story told by Small Okikirmuy himself). It has been annotated with parts of speech and Japanese translations of tokens by expert annotators [16]. We used this annotated yukar in evaluation of POS tagging and Token Translation.

5.2 Evaluation Experiments

We performed evaluation experiments for all components of POST-AL: tokenization (DL-LSM), POS tagging (S-POST and CON-POST) and token translation (RAN-ToT and CON-ToT). All results were calculated with the means of Precision (P), Recall (R) and balanced F-score (F), standard score calculation methods used in Information Extraction. 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.

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

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

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

Tokenization: At first we performed tokenization of all stories. The DL-LSM method achieved high results with Precision over 99% and Recall over 97%. There were 69

Table 1: Evaluation results for all parts of POST-AL.

	Precision	Recall	F-score
Tokenization			
DL-LSM	99.29%	97.64%	98.46%
POS Tagging			
S-POST	83.64%	97.66%	90.11%
CON-POST	96.26%	97.66%	96.96%
Token Translation			
RAN-ToT	83.64%	97.66%	90.11%
CON-ToT	99.07%	97.66%	98.36%

errors, which were caused by inconsistencies of the dictionary to the original tokenization.

Part-of-Speech Tagging: In part-of-speech tagging there were large differences between the two methods. Statistical method (S-POST) achieved 83% of Precision, while contextual method (CON-POST) achieved over 96%, which is an improvement of about 13 percentage points. However, there were still some errors, all of them for words that were not equipped with examples in the original dictionary. This proves that the HMM based disambiguation of parts of speech is applicable also for language isolate like Ainu.

Token Translation: The task of translating the tokens showed similar result tendencies as POS tagging. The contextual method achieved much better results (Precision = 99.07%) than the random method (Precision = 83.64%). This shows an improvement of over 15 percentage points.

6 Conclusions and Future Work

In this paper we presented POST-AL, the first part-of-speech tagger for Ainu language. The Ainu language is close to extinction and it is estimated there could be no native speaker of this language in one or two generations. Our obligation is to keep the language alive. This could be done by performing analytic research on the language, or by developing a virtual character-based story teller. In both cases a long list of intermediary tools is needed. The first tool on the list, without which no language-related task could be performed automatically, is a part-of-speech tagger. POST-AL performs three main tasks: tokenization, part-of-speech tagging and token translation. At present POST-AL uses a database created on one dictionary [9]. In the future we will enlarge the database by adding other dictionaries [7, 8] and add English translations [5] to make the tool usable also for non-Japanese speaking researchers. Having annotated a larger number of Ainu stories we plan to perform a robust evaluation of the annotations with the help of several experts and Ainu native speakers. After the annotations are evaluated we will be able to bootstrap the system for even better performance. We also plan to apply POST-AL to machine translation and develop a dependency parser for the Ainu language. In the future we also wish to contribute to the development of an artificial story teller for Ainu "yukar" stories.

Acknowledgments

This research was supported by (JSPS) KAKENHI Grant-in-Aid for JSPS Fellows (Project Number: 22-00358).

The authors express their gratitude to Associate Professor Hideo Kirikae of Hokkai-Gakuen University for his contribution of the dictionary source files and numerous insightful comments on the desirable structure and usability of the system.

References

[1] Christopher Moseley (ed.). 2010. *Atlas of the World? Languages in Danger*, 3rd ed. Paris, UNESCO Publishing. Online version: http://www.unesco.org/culture/languages-atlas/

[2] 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.

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

[4] 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)

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

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

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

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

[9] 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]. Sapporo: Hokkaidō Daigaku Bungakubu Gengōgaku.

[10] Mashiho Chiri. 1974. *Ainu gohō gaisetu* (An outline of Ainu grammar). In *Chiri Mashiho chosakushuu* (Collection of works by Machiho Chiri) [In Japanese], vol. 4, 3-197. Tokyo, Heibonsha. Reprint from 1936.

[11] Kyōko Murasaki. 1979. *Karafuto ainugo. Bunpō-hen* (Sakhalin Ainu. Grammar volume) [In Japanese]. Tokyo, Kokushokan-kōkai.

[12] Suzuko Tamura. 2000. *The Ainu Language*. Tokyo. Sanseido.

[13] Tomomi Satō. 2008. *Ainugo bunpō no kiso* (The basics of Ainu grammar) [In Japanese]. Tokyo, Daigakushorin.

[14] Yasunori Azumi and Yoshio Momouchi. 2009a. Development of Analysis Tool for Hierarchical Ainu-Japanese Translation Data [In Japanese]. *Bulletin of the Faculty of Engineering at Hokkai-Gakuen University*, No.36, pp.175-193.

[15] Yasunori Azumi and Yoshio Momouchi. 2009b. Development of Tools for retrieving and analyzing Ainu-Japanese translation data and their applications to Ainu-Japanese machine translation system [In Japanese]. *Engineering Research: The Bulletin of Graduate School of Engineering at Hokkai-Gakuen University*, No.9, pp.37-58.

[16] 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.

[17] 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.