E-005

A Study on Automatic Thai Nickname Recommendation Considering the Similarities to Japanese Names

Shiho Hoshi Nobesawa[*1] Pannathorn Naksung[*2] Sirichai Khomleart[*2]
Yusuke Sakai[*1] Hideo Furugori[*1] Kodai Inada[*1] HuanYuan Zhao[*3]
Kazumasa Fujita[*4] Aliaksei Khadanovich[*4]

1.Introduction

Thai nicknames are more than just for convenience or terms of endearment among family members and close friends; they can be bound up in the very nature of Thai-ness and an important part of Thai culture. Thus we here propose our ideas on automatic Thai nickname recommendation methods. Our aim is a communication support system, not an automatic nickname proposal system, as there may not be any significance in proposing a nickname which won't be used.

In this paper we show our recommendation methods of suitable Thai nicknames for each Japanese kanji input. We consider three types of similarities between Thai nicknames and Japanese full names; possible meanings included in the original name, Soundex index, and string edit distance in romaji.

2.Recommendation Methods

Fig.1 shows the basic idea of our recommendation methods. We take a Japanese full name in kanji as an

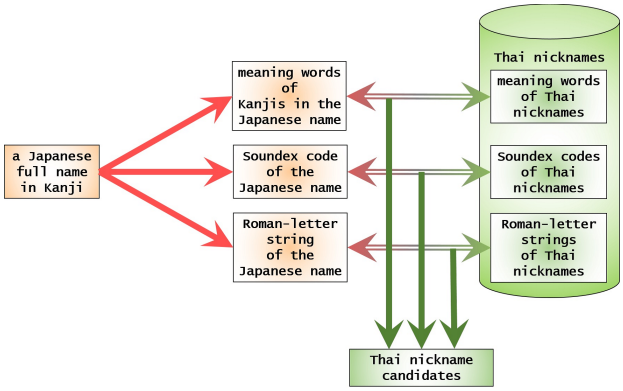

Fig.1: Basic Idea on Thai Nickname Recommendation for Japanese Names

input, and output Thai nickname candidates (Fig.1). Our system also considers the gender.

Our goal here is not on the selection of one best Thai nickname candidate for a Japanese full name. We rather intend to propose some basic ideas for the similarity matching between foreign names. Thus in this paper we propose three types of similarities for names; the similarity in the meaning (sim_m), the similarity in the pronunciation (sim_p) and the similarity in the spelling (sim_s).

For Thai nickname dataset, we collected Thai nicknames from an online source with basic information like gender and real names.

2.1.Similarity in the Meaning: sim_m

The basic idea on our meaning-based similarity is to compare the meanings of kanji letters contained in the Japanese full name and the meanings of Thai nicknames.

Our system chunks the input Japanese full name into a bag of words. Since Thai nicknames mostly contain only one meaning chunk, we didn't have to do the same segmentation for Thai nicknames. We even tried bag-of-word expansion to increase matching possibility.

Both these Japanese words and Thai nicknames get translated to English words for the similarity matching (Fig.2). We used Wu-Palmer Similarity[1, 2] based

坂, 井, 優, 介	⇒	
well, wellcrib, ···,	⇔	*bank, deposit, savings*
superiority, gentleness,		*bank, trust, camber,*
slope, incline, hill		*cant, ···*
	⇒	แบงค์ baeng (*bank*)

Fig.2: An Example of Similarity in Meaning

on the WordNet corpus to measure the meaning similarity between a Japanese full name and Thai nicknames.

2.2.Similarity in the Pronunciation: sim_p

Our pronunciation similarity matching is based on Soundex (Fig.3). Both Japanese names and Thai nicknames are romanized into pronunciation form to obtain Soundex representations. Then we use Levenshtein distance to calculate the distance between two pronunciation forms.

坂井優介	⇒	
sakai yū kai yasa yuu	⇒	
S222	⇔	K200
	⇒	เค้ก khek (*cake*)

Fig.3: An Example of Similarity in Pronunciation

For the better use of Soundex, we should consider the pronunciation of target languages. The Japanese name in Fig.3 includes [y] sound, which is counted as a consonant both in Japanese and Thai. But common Soundex takes [y] as a vowel and omit it, so we should consider adding it to the local Soundex rule. This also happens to the sound [h], which should be added to the sound group for [b, f, p, v]. As the pronunciation of the sound [l] is often mixed up with the sound [r] in Japanese, we may be able to gain better accuracy by considering these two as in a same sound group.

Automatic romanization of Japanese names is not always successful. A name can be chunked into parts (Fig.3), or mis-romanized such as `kazunari`, a common pronunciation for 和成, where `kazumasa` was rare but correct for the person. This mistake can not be avoided because of the creativity of Japanese names, thus we should include both kanji and romaji for Japanese name inputs.

[*1]Tokyo City University
[*2]Sirindhorn International Institute of Technology, Thammasat University
[*3]Tokyo City University/Dalian Jiaotong University
[*4]Graduate School of Engineering, Tokyo City University

2.3.Similarity in the Spelling: sim_s

We also tried an edit distance matching between a Japanese full name and Thai nicknames. Levenshtein distance was used to quantizing the difference between these romanized strings (Fig.4).

坂井優介 ⇒
sakai yū kai yasa yuu ⇔ ไข่ไก่ khaikai (*egg*)

Fig.4: An Example of Similarity in Spelling

We expected this sim_s to have better output than sim_p, as sim_s has less string replacement comparing to sim_p. However, as this method uses a full name as an input, its edit distance between Thai nicknames are mostly not small enough. We should segment input names into smaller pieces for the better use of this spelling similarity.

3.Discussion

We prepared 40 names in Japanese letters for our input. Two of them were of foreign names, one was a Chinese name written in kanji, and the other was a Belarusian name written in katakana. As for the Belarusian name the system could not output any valid candidate, because of its non-kanji representation. For the Chinese name, the system succeeded in output valid candidates just as Japanese names, for its pronunciation and its romanized spelling were both obtained in a Japanese fashion. Thus 39 out of 40 full names had valid output (examples in Tab.1).

Tab.1: Thai Nickname Recommendation Results

kanji	sim_m	sim_p	sim_s
稲田弘大	bik	ainamtatuen	ingnoinanak
Inada Kōdai	*big*	*teary*	*guess*
坂井優介	baeng	khek	khaikai
Sakai Yūsuke	*bank*	*cake*	*egg*
櫻井克憲	ket	usrakon	sueanoi
Sakurai Katsunori	*sneaky*	*Urasakorn*	*little tiger*
田中耕平	ket	tang	namkhing
Tanaka Kōhei	*sneaky*	*gluten*	*ginger juice*
藤田和成	thinni	phunithat	kutnai
Fujita Kazumasa	*thin*	*Purdue*	*good night*
古郡英朗	sai	rak	smrueti
Furugōri Hideo	*clear*	*love*	*consciousness*
丸野裕貴	khan	marina	mattuni
Maruno Yūki	*gift*	*marina*	*Moszini*
趙煥元	peni	chao	ch
Chō Kangen	*agile*	*brave*	*the*
延澤志保	hang	poetia	baihmon
Nobesawa Shiho	*hope*	*Persia*	*mulberry*

Due to the small Thai nickname dataset, there were some Thai nicknames frequently recommended. All the three Japanese names which contained a kanji 一 (*one*) had the same Thai nickname hnueng (*one*) as the meaning-based candidate. Both two Japanese names with 大 (*big*) output bik (*big*), both two Japanese names with 山 (*mountain*) had phupa (*mountain*). This problem may be avoidable by uniting several types of similarity matching methods.

The outputs for sim_p and sim_s may be improved by using only given names, or using given names and family names separately, as they seem to be longer than the outputs of sim_m. Our method requires a Japanese full name as its input. Since Thai nicknames are often very short, and Japanese names mostly consist of two to five moras, we may be able to use only Japanese given names to obtain a better match based on the pronunciation and the spelling. We can give priority to given names, and use family names as additional inputs.

If our goal is to select one best-fit Thai nickname based on these similarity features, we need to combine their similarity scores effectively. Fig.5 shows the ranges of similarity scores of the three similarity methods. sim_{max} in Fig.5 shows the distribution of

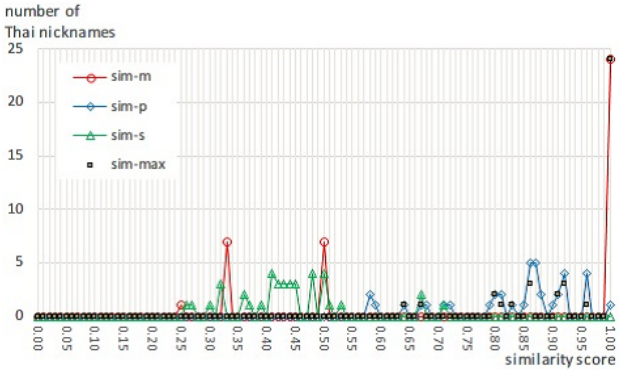

Fig.5: Distribution of Similarity Scores

$max(sim_m, sim_p, sim_s)$ for each input name. Fig.5 indicates that we need to examine better similarity-score estimation methods. According to Fig.5, sim_m takes only four values as its similarity scores (1, $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{4}$). Score sim_s is comparatively low (.44 on average) to score sim_p (.83 on average). To merge these similarities to rank Thai nicknames considering several features, we first need to study on the appropriate estimation of similarities. Then we need to examine effective weighting and effective combination of the similarity scores, considering which feature should be more significant in nickname recommendation.

4.Conclusion

Here in this paper we proposed a similarity-based nickname recommendation between foreign names. Our approach considers the meaning, the pronunciation and the spelling to select Thai nickname candidates for input Japanese names. As both Japanese and Thai have their own letter sets, we romanized the names to match them based on the pronunciation and the spelling. And we also tried to make the best use of Japanese kanji. With proper machine learning models and proper dataset, we may be able to expect a better result.

We believe our system helps better communication between Japanese people and Thai people.

References

[1] Yuanyuan Cai, Qingchuan Zhang, Wei Lu, and Xiaoping Che. A hybrid approach for measuring semantic similarity based on IC-weighted path distance in WordNet. In *Journal of Intelligent Information Systems*, volume 51, pages 23–47, August 2018.

[2] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL '94)*, pages 133–138, June 1994.