Inducing Romanization Systems from Bilingual Alignment

Keiko Taguchi Doshisha University
Andrew Finch NICT
Seiichi Yamamoto Doshisha University
Eiichiro Sumita NICT

{ buj1077@mail4.doshisha.ac.jp, andrew.finch@nict.go.jp, seyamamo@mail.doshisha.ac.jp, eiichiro.sumita@nict.go.jp }

Abstract

In this paper we present a method for inducing romanization systems for languages directly from a bilingual alignment at the grapheme level. Our approach learns how to romanize by first aligning transliteration word pairs using a non-parametric Bayesian approach, and then for each grapheme sequence to be romanized, selecting the optimal romanization according to a user-specified criteria. We apply our approach to the task of transliteration mining where the edit distance to romanized text is often used as a measure of cross-language word similarity. For these experiments we therefore used edit distance as our criteria for the selection of the romanization. Our experiments on Japanese-English showed that mining performance is strongly dependent on the romanization system used: for example Hepburn romanization was significantly more effective than Nihon-shiki romanzation on this task. Furthermore we show that the mining system built from the romanization system induced using our technique was able to outperform both of the existing baseline romanization systems, and we provide a detailed analysis of the underlying process in order to explain this result.

1 Introduction

Romanization is the process of producing a string in roman script from a string in another language with a different writing system. In Japan there are two prominent systems for romanization: the Hepburn system (ヘボン式ローマ字) and the Nihon-shiki system (日本式ローマ字). The former follows the principle of phonemic transcription and attempts to render the significant sounds (phonemes) of English as faithfully as possible. The latter attempts to transliterate the original script (kana syllables) with less emphasis on how the result sounds when pronounced according to the English, and more emphasis on how the kana syllables are pronounced.

Pure transcriptions are generally not possible, as the one language usually contains sounds and distinctions not found in the other language; these are often made explicit in the romanization by inserting characters that to represent them. In general, building a usable romanization system involves trade-offs between the two extremes of transliteration and transcription.

Recently romanization systems have taken on new roles for which they were not originally designed. Examples being the cross-lingual word similarity task we study in this paper, and as methods of textual input for languages where the native character set is too large to represent directly on a user interface. In this paper we demonstrate the effectiveness of our method on the well-defined task of cross-lingual word similarity but in principle our method could be extended to encompass more complex and realistic criteria necessary for romanizing for other purposes.

The main merits of our approach are that it can be applied to any language where data are available to train our model, and that it can be used to either induce romanization systems for languages that have none, or produce alternative romanization systems for languages that have existing systems. We will show later in this paper that in our chosen application, it is possible to induce a romanization system that is more effective than simply choosing from existing schemes.

2 Related Work

In many transliteration mining approaches [1, 5], romanization is required to compare words across languages, typically using normalized edit distance metrics. Statistical transliteration systems can be used, but these need large amounts of training data which may not be available. As far as the authors are aware the only other reported automatic romanization induction system was reported by [6]. The advantage of their method is that it can be applied to many different languages without the need for an existing romanization system. However, their approach romanized every foreign language grapheme with only a single Roman character, potentially causing problems for languages such as Japanese and Chinese where single graphemes align naturally to multiple Roman characters; we investigate these issues in Section 4.2.

3 Romanization Induction

Our method induces a romanization system directly from a non-paramteric Bayesian bilingual alignment [2] between source and target grapheme sequences. This model has been shown to align consistently, without a tendency to overfit the data, and is therefore suitable for one-to-many and many-to-many alignment. We use Levenshtein distance (LD) to select an appropriate romanization from a set of candidates derived from the alignment.

More formally, let $\mathcal{S} = (s_1, s_2, \ldots, s_I)$ and $\mathcal{T} = (t_1, t_2, \ldots, t_I)$ be a corpus of source and target words respectively, each s_i and t_i are sequences of graphemes in their respective writing systems.

Let Π and Ω be sets of grapheme sequences in the source and target writing systems respectively. For example, for Japanese the set may be syllables, and for English the set could be the alphabet. The romanization rules $\mathcal{R}$ are defined to be a set of tuples (s_j, r_j), where s_j and r_j are source and target grapheme sequences: $\forall j\ s_j \in \Pi$ and $r_j \in \Omega$.

$$\mathcal{R} = \{(s_1, r_1), (s_2, r_2), \ldots, (s_J, r_J)\} \tag{1}$$

The r_j are selected by choosing from the set $\mathcal{C}_j$ of all target grapheme sequences aligned in the corpus to the source grapheme sequence s_j: $\mathcal{C}_j = \{c_1, c_2, \ldots, c_K\}$. The romanization r_j of s_j is chosen from this set in order to minimize the expected cost in terms of Levenshtein distance to the English in the manner described below.

Let $\phi : \Pi \mapsto \Omega$ be the romanization function defined by $\mathcal{R}$:

$$\phi(s_j) = \underset{c_k \in \mathcal{C}_j}{\arg\min}\, E[D(c_k)] \tag{2}$$

Where $D(c_k)$ is the cost in terms of Levenshtein distance from using romanization rule (s_j, c_k). For a single occurrence of s_j in the corpus, this cost is $LD(c_k, \psi(s_j))$, the Levenshtein distance between romanization candidate sequence c_k and $\psi(s_j)$, the target grapheme sequence aligned to s_j.

The expected value of this cost over the corpus is calculated according to:

$$E[D(c_k)] = \sum_{l=1..K} p(c_l) LD(c_k, c_l) \tag{3}$$

4 Experimental Methodology

4.1 Data

For training and evaluation in our experiments we used the Japanese-English translation mining corpus of [3]. This corpus consists of 4339 Japanese-English word pairs extracted from Wikipedia interlanguage link titles, all of which are annotated as correct/incorrect transliteration pairs. 3800 of the word pairs were correct transliterations and 539 word pairs were incorrect.

4.2 Induced Systems

We induced two different romanization systems from the data. The simplest method (Unigram) discovered romanizations for each individual kana character. A more sophisticated method learned romanizations for multiple sequences of kana (N-gram). Table 1 shows example romanization rules for a selection of characters that differed in romanization from the Hepburn/Nihon-shiki systems. It is interesting to note that our two induced systems (Ungram and N-gram) learned the same romanization rules as the standard systems for most Japanese graphemes (grapheme sequences in the case of the N-gram system); the N-gram approach shares 69% of its romanization rules with the Hepburn system. The romanization of the character ル exemplifies two of the main differences between the human and machine produced systems. Both of the automatic methods prefer romanizing with an ‘l’ rather than an ‘r’ because ‘l’ is more frequently used in English with this syllable. Furthermore, the automatic methods have dropped the ‘u’ which is used in the Japanese pronunciation of the syllable, but rarely occurs in the English spellings.

Kana	Hepburn (Nihon-shiki)	N-gram	Unigram
カ	KA	CA	CA
ク	KU	C	K
グ	GU	G	G
ケ	KE	CE	KE
コ	KO	CO	CO
シ	SHI (SI)	SI	S
ジ	JI (ZI)	GI	G
ス	SU	S	S
ズ	ZU	S	S
ゼ	ZE	SE	SE
ツ	TSU (TU)	TS	TS
ト	TO	T	T
ド	DO	D	D
フ	FU (HU)	F	F
ブ	BU	B	B
プ	PU	P	P
ム	MU	M	M
ユ	YU	U	U
ヨ	YO	JO	JO
ル	RU	L	L
キャ	KIYA(KYA)	CA	-
クィー	KUII	QUEE	-

Table 1: The romanization rules from two standard systems, and two systems automatically induced from data.

5 Results and Analysis

5.1 Mining Performance

In order to classify the data into correct/incorrect transliteration pairs we used normalized edit distance (NED). A similar approach was taken by [1, 5, 6]. We calculated the NED between English words and corresponding romanized forms produced by each system. LD determines the similarity of two strings: the minimum number of insertions, deletions, and substitutions required to transform one string into the other. In our experiments, NED was calculated by dividing the LD between the two sequences by the length of the edit path, and yields a value between 0 and 1 that is robust to differences in sequence length.

We applied a range of thresholds to the NED to produce the receiver operating characteristic (ROC) curves for the classifiers shown in Figure 1. The ROC is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. The ROC is shown for our proposed systems (N-gram(LD), Unigram), as well as well known Japanese systems (Hepburn, Nihon-shiki), and the approach taken by [6] (Single-character) that romanizes each kana to the single English character that it most frequently aligns to. Also on the plot is a curve (N-gram(Freq)) for a system which used the same Bayesian alignment as our N-gram(LD) system, but selected the romanizations according to frequency rather than minimizing the Levenshtein distance. The results show that our proposed N-gram romanization system achieves the best performance. It is also interesting to note that the Hepburn system outperforms the Nihon-shiki system. One explanation for this is that the Hepburn system was designed as a way for foreigners to read Japanese and is therefore more likely to be similar to English in character than the Nihon-shiki system which is focused on expressing pronunciation characteristics. The performance of (Single-character) was quite poor indicating this approach is not suitable for some language pairs, even though it performed well on the Russian-English task in the NEWS2010 workshop.

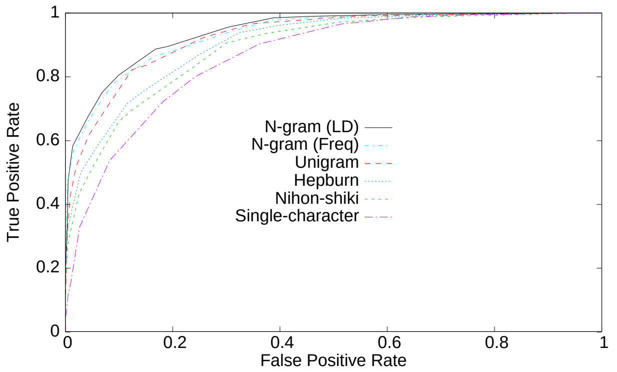

Figure 1: ROC curves for various mining approaches on Japanese-English data.

5.1.1 Statistical significance

The AUC statistics for each approach are shown in Table 2. The AUC represents the probability that a classifier will rank a randomly chosen transliteration pair instance higher than a randomly chosen noise pair. We ran significance tests on the area under curve (AUC) statistics using the method set out [4]. We found that all the AUCs of adjacent lines in the graph are significantly different (α<0.05) with the exception of the two best approaches based on the N-gram techique (α=0.13).

Approach	AUC	Length	Mean LD
N-gram (LD)	0.942	6	2.6
N-gram (Freq)	0.936	6	2.7
Unigram	0.927	6	3.1
Hepburn	0.907	7	3.7
Nihon-shiki	0.892	7	4.0
Single-character	0.867	3	4.6

Table 2: Statistics from the romanization approaches.

5.2 Effect on the Distributions of NED

In order to gain some insight into the mechanism by which our approach improves the mining performance, we show kernel density plots of the probability density functions (PDF) of NED for correct/incorrect transliteration pairs for various romanization systems in Figure 2. From visual inspection of the incorrect pair plots, it appears that the choice of romanization system has little effect on the NED PDFs for the incorrect pairs. We performed a Kolmogorov-Smirnov test (a non-parametric test for the equality of distributions) on the incorrect pair distributions. All pairs of distributions were equal at α=0.05 according to this test, with the exception of the N-gram to Hepburn/Nihon-shiki comparisons.

Moreover, from the correct pair plots it appears that the better the romanization system performed in our experiments, the further the NED PDFs are shifted to the left. This gives a visually intuitive explanation of how our approach operates: by reducing the edit distance to the English, the correct pair PDF is shifted to the left while the incorrect pair PDF remains fixed in position, resulting in a separation of the two distributions (see Section 5.1.1). We performed a Wilcoxon signed-rank test on samples from the correct pair distributions and found that all distributions were significantly different (α=0.05).

Finally, it is interesting to observe the densities where the NED is zero. This is the case where the English spelling is generated exactly from the Japanese. The N-gram system generated the correct spelling approximately twice as often as the best of the other systems.

5.3 Qualitative difference

We calculated the probability of occurrence of each roman character in the N-gram romanization, Nihon-shiki romanization, and the reference English. Figure 3 shows the relative difference in probability with respect to the reference English. The major differences are that the Nihon-shiki system tends to over-generate the vowel ‘u’ due to the fact that consonants are always romanized as consonant vowel pairs. It under-generates the consonants ‘c’ and ‘l’ since the system never uses them, instead using ‘k’ and ‘r’ respectively. For example, the word スクール is romanized as ‘SUKUURU’ with the Nihon-shiki system and as ‘SCOOL’ using the induced N-gram system.

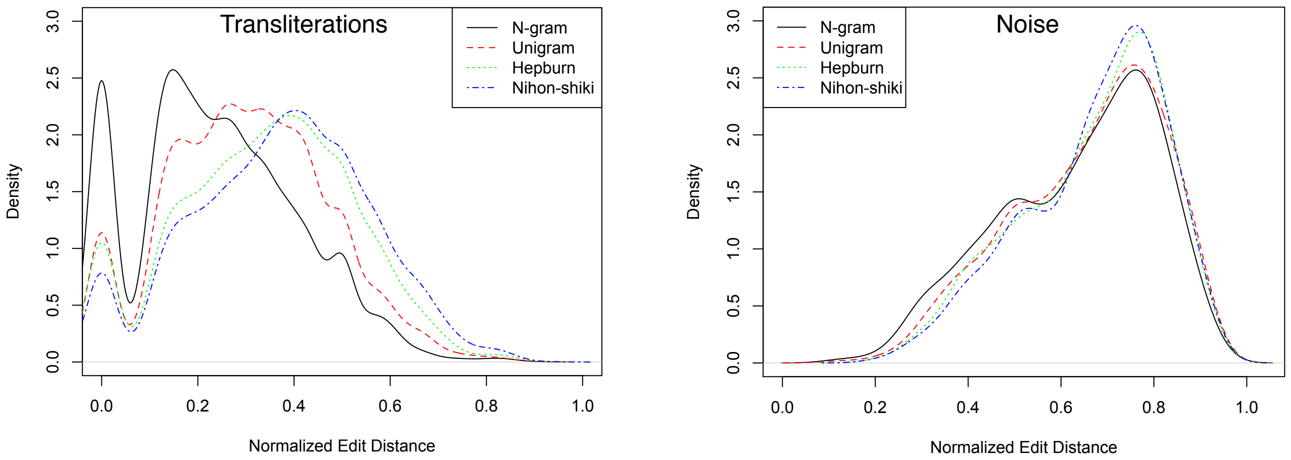

Figure 2: Kernel density plots of NED for transliteration pairs and noise.

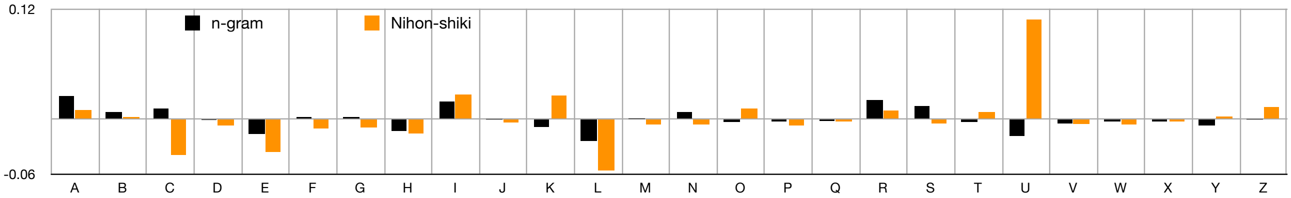

Figure 3: Character occurrence frequencies relative to English.

6 Conclusion

In this paper we introduced a novel unsupervised romanization technique for the induction of a complete system of romanization automatically from a bilingual corpus. First, a bilingual corpus of words is aligned using a many-to-many non-parametric Bayesian sequence alignment method, and then for each sequence of characters to be romanized, a set of possible candidate romanization rules is extracted with reference to the alignment. Finally, the best romanization rules are chosen from this set according to an appropriate criterium. We applied our technique to the task of producing a romanized script similar to English from Japanese, for the purposes of transliteration mining. In these experiments we used a corpus of Wikipedia interlanguage link titles, and a criterium based on Levenshtein distance. We found that mining performance depends heavily on the choice of romanization system used. Furthermore, we show that using our approach gives rise to a romanization system that significantly outperformed two existing romanization schemes on the mining task. Our approach, was trained on noisy data and performed well enough that a bootstrapping approach was not attempted. It is theoretically language independent and requires only a training corpus and a well-defined criterium for selecting among possible romanization candidates from the alignment. In the future we would like to investigate the performance of our approach on other language pairs using different criteria for romanization. In particular it would interesting to build a system capable of finding a more-humanlike romanization scheme that captures the tradeoffs between transliteration and transcription. Such an approach could be used as an aid to creating romanization systems for languages that do not yet have a standard system. We believe another important future extention of our technique could be in the automatic discovery of systems for textual input in romanized form that are both efficient and also sufficiently capture the phonetics of the underlying graphemes.

References

[1] W. Aransa, H. Schwenk, L. Barrault, and F. Le Mans. Semi-supervised transliteration mining from parallel and comparable corpora. *Proceedings IWSLT 2012*, 2012.

[2] A. Finch and E. Sumita. A Bayesian Model of Bilingual Segmentation for Transliteration. In M. Federico, I. Lane, M. Paul, and F. Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 259–266, 2010.

[3] T. Fukunishi, A. Finch, S. Yamamoto, and E. Sumita. Using features from a bilingual alignment model in transliteration mining. In *2011 Named Entities Workshop*, page 49, 2011.

[4] J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.

[5] O. Htun, A. Finch, E. Sumita, and Y. Mikami. Improving transliteration mining by integrating expert knowledge with statistical approaches. *International Journal of Computer Applications*, 57, November 2012.

[6] S. Jiampojamarn and G. Kondrak. Letter-phoneme alignment: An exploration. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 780–788. Association for Computational Linguistics, 2010.