Inducing Romanization Systems

Keiko Taguchi
Doshisha University / 1-3 Miyamaki
Tatara, Kotanabe-shi, Kyoto
610-0321, Japan
dun0153@mail4.doshisha.ac.jp

Andrew Finch
NICT / 2-2-2 Hikaridai
Seika-cho, Soraku-gun, Kyoto
619-0288, Japan
andrew.finch@nict.go.jp

Seiichi Yamamoto
Doshisha University / 1-3 Miyamaki
Tatara, Kotanabe-shi, Kyoto
610-0321, Japan
seyamamo@mail.doshisha.ac.jp

Eiichiro Sumita
NICT / 2-2-2 Hikaridai
Seika-cho, Soraku-gun, Kyoto
619-0288, Japan
eiichiro.sumita@nict.go.jp

Abstract

We propose a method for inducing romanization systems directly from a bilingual alignment at the grapheme level. First, transliteration word pairs are aligned using a non-parametric Bayesian approach, and then for each grapheme sequence to be romanized, a particular romanization is selected according to a user-specified criterium. We apply our approach to the task of transliteration mining, and used Levenshtein distance as the selection criterium. We performed experiments on three languages with differing characteristics: Japanese, Russian and Chinese. Our experiments show that the mining system built from the induced romanization system is able to outperform existing baseline romanization systems. By extending our approach to induce romanization systems based on other criteria we expect our technique may find more general application in the future.

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.

This paper investigates the possibility of discovering systems of romanization automatically from data. The process is based on two steps: first bilingual alignment of transliteration pairs is made, yielding a set of possible romanization candidates; second a candidate is chosen according to a specific selection criterium. The contribution of our work is twofold: first we propose the first system capable of learning to fully romanize from a corpus, and second we show this system, when used with an appropriate romanization selection criterium, is able to improve the discrimination capability of a state-of-the-art transliteration mining approach.

We now move on to motivate the development our our approach, and detail the existing related research in the area. Section 4 describes the methodology we used. In Section 5.1 we present the experiments we performed on inducing a romanization system for Japanese, and compare our methodology to other plausible automatic strategies as well as the two principle romanization systems in general use. We also analyze the characteristics of the induced romanization, and expose the mechanism by which it is able to improve mining performance. Sections 5.2.1 and 5.2.2 gives details of our experiments in Russian and Chinese, and presents a study of the effect of training data size on the quality of the induced romanization. Finally, in Section 6 we

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 19–26.

conclude and suggest avenues for future research.

2 Motivation

At first glance it would seem strange to attempt to induce romanization systems for languages that already have established systems in general use. However, recently romanization systems have taken on new roles for which they were not originally designed, and it is possible that more optimal systems may be waiting to be discovered for these purposes.

One example of a new use for a romanization system is the transliteration mining task we study in this paper. Romanization is being used as a means of performing cross language cross-lingual word similarity between languages whose scripts are not directly comparable. Converting the scripts into a common representation (for example Roman characters, or a phonetic representation like Soundex), allows comparisons to be made between languages with different scripts.

Another example of a new use of a romanization system is for the input of text into a machine. In many languages the native character set is too large to represent directly on a user interface. A simple, commonly-adopted solution to this problem is to to use a Roman keyboard and input text as a sequence of Roman characters in accordance with an existing romanization scheme. An example of such a system Pinyin for entry of Chinese. There are problems, however. First and foremost, existing romanization systems were not originally conceived as user input methods, and many are cumbersome and long-winded to enter; they may be very explicit about precisely how a grapheme ought to be pronounced, and even make clear the subtle differences in pronunciation between one character and another, but at the cost of the lengths of the character sequences required to express this information.

Second, there are often multiple competing romanization systems, and users may use one or the other or a mixture them for input. Both of these problems are illustrated in the input system for Japanese in which input is possible in a mixture of Hepburn and Nihon shiki romanization. For example the Japanese character 'ち' can be input as 'chi' (Hepburn) or 'ti' (Nihon shiki), with 'ti' usually being preferred because it is shorter even though 'chi' reflects the pronunciation of the character more accurately. On the other hand, 'じゃ' can be input as 'ja' (Hepburn) or 'zya' (Nihon shiki) however in this case the Hepburn form is almost always used because it is both shorter and represents the phonetics of the syllable adequately. Therefore, for Japanese *neither* of the existing romanization systems used for user input is ideal. Users typically differ in the manner in which they input text, but it is clear that a better system than either existing system must exist, if only it can be discovered.

For the task of discovering a romanization system suitable for user input three factors need to be taken into account: how well the romanization represents the phonetics of the characters it is romanizing; how efficient the system is for input; and whether or not words can be input unambiguously using the system. In this paper we chose to work with the problem of romanizing for cross lingual word similarity because the criteria for choosing among candidate romanizations can be simple and well-defined, and also the performance of the resulting system is straightforward to evaluate and analyze. We believe however that our technique is more generally applicable and 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 the model, and that it can be used to either induce romanization systems for languages that have none, or lack a standard system of romanization (for example Myanmar (Oo and Thein, 2011)), 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 well-established existing schemes.

3 Related Work

In many transliteration mining approaches (Aransa et al., 2012; Htun et al., 2012), 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. A simple system of automatic romanization was used by (Jiampojamarn et al., 2010) to great effect in the shared mining task of the NEWS2010 workshop. Their system allowed cross language comparison between word pairs in different scripts by aligning single characters in one script to either single Roman characters or to NULL. Their romanization rules roman-

ized by substitution with the most representative single character, or by deletion. Our work differs from theirs in that we are aiming to induce a full romanization involving multiple characters on the target side[1] without the deletion of the characters to be romanized. The fact that romanization was only performed with a single character in their approach may lead to problems for languages such as Japanese and Chinese where single graphemes align naturally to multiple Roman characters; we investigate these issues in Section 5. Nonetheless, the system of (Jiampojamarn et al., 2010) is capable of state-of-the-art performance; the system achieved the top rank in the shared evaluation for most of the tracks in which the automatic romanization strategy was used, motivating the research presented here.

As far as the authors are aware this work and the work in this paper are the only romanization induction techniques reported in the literature to date. The advantage of these methods is that they can be applied to many different languages without the need for an existing romanization system, and can be optimized to fit a specific purpose. Furthermore, as we will show later, the strategy seems quite robust to noise in the data, and we were able to build effective systems from data that contained non-transliteration pairs.

4 Methodology

Our method induces a romanization system directly from a non-parametric Bayesian bilingual alignment (Finch and Sumita, 2010) 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 both 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 corpora of source and target words respectively. Each s_i and t_i are represented as 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 Π might be syllables, and for English the set Ω could be the alphabet. The romanization rules $\mathcal{R}$ are defined to be a set of tuples (o_j, r_j), where o_j and r_j are source and target grapheme sequences: $\forall j\ o_j \in \Pi$ and $r_j \in \Omega$.

$$\mathcal{R} = \{(o_1, r_1), (o_2, r_2), \ldots, (o_J, r_J)\} \quad (1)$$

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 o_j: $\mathcal{C}_j = \{c_1, c_2, \ldots, c_K\}$. The romanization r_j of o_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 $\mathcal{R}$:

$$\phi(o_j) = \arg\min_{c_k \in \mathcal{C}_j} E[D(c_k)] \quad (2)$$

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

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

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

5 Experiments

5.1 Inducing Japanese Romanization

5.1.1 Data

For training and evaluation in our experiments we used the Japanese-English translation mining corpus of (Fukunishi et al., 2011). 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 noise.

5.1.2 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

[1] Although in principle it is possible to learn romanizations with multiple characters on the source side, in the experiments in this paper we do not attempt to learn the source language segmenation as this could lead to issues of ambiguity when applying romanization rules. For Japanese, we used a universally accepted set compound kana groupings.

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.

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 (Unigram 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.

5.1.3 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 (Aransa et al., 2012; Htun et al., 2012; Jiampojamarn et al., 2010). 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 (Jiampojamarn et al., 2010) (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, but it was only slightly better than the frequency-based variant of the approach. 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 nature 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.

5.1.4 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 (Hanley and McNeil, 1982). We found that all the AUCs of adjacent lines in the graph are significantly different ($\alpha<0.05$) with the exception of

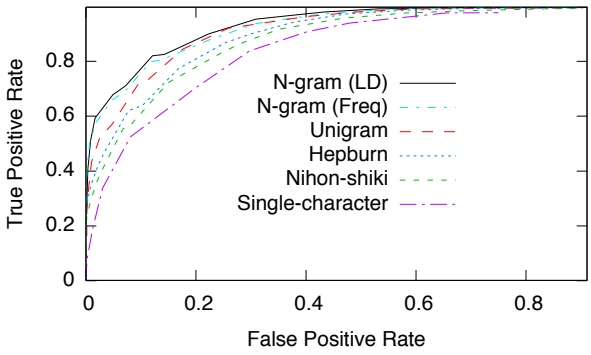

Figure 1: ROC curves for various mining approaches on Japanese data.

the two best approaches based on the N-gram technique (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.

5.1.5 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 (PDFs) 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 Levenshtein 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 5.1.4). 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.

5.1.6 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.

5.2 Application to Other Languages

We applied our approach to Russian and Chinese investigate the behavior of our approach on both simpler and more challenging languages to romanize. These experiments were carried out on the task of transliteration mining, using our N-gram approach to induce the romanization system. For these experiments we decided to only induce the romanization from clean data because the proportion of non-transliteration pairs in the corpora was far higher than in the Japanese-English data. Although it is possible our approach may work from such noisy data, it remains future research.

5.2.1 Inducing Chinese Romanization

We induced a romanization for each Chinese grapheme from a 1000-pair corpus of clean transliteration data of Chinese named entities that was the seed data set used in the NEWS2010 Shared Mining Task. The test data consisted of the 621-pair reference data set for this task. Unfortunately, as might be expected our approach was not able to succeed on this task due to the much larger grapheme set.

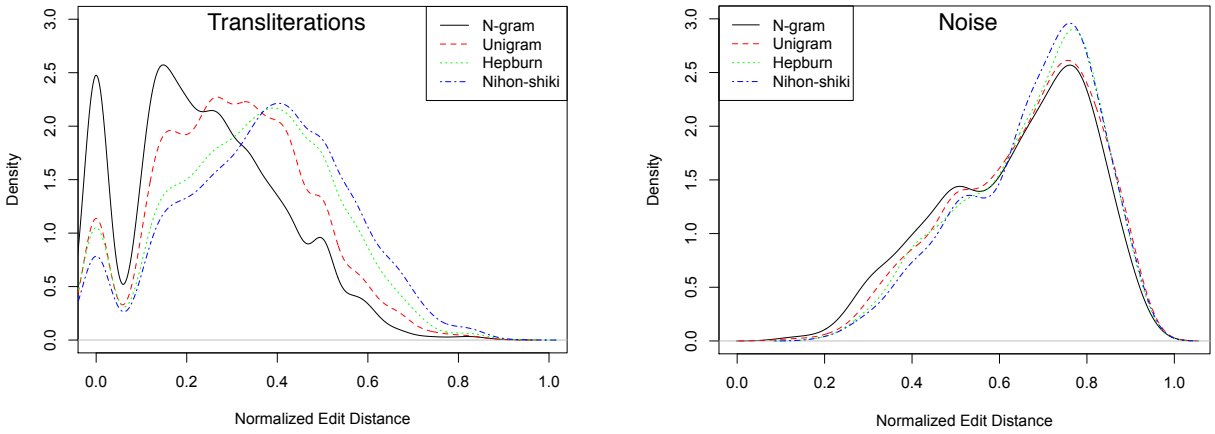

Figure 2: Kernel density plots of NED for transliteration pairs and noise.

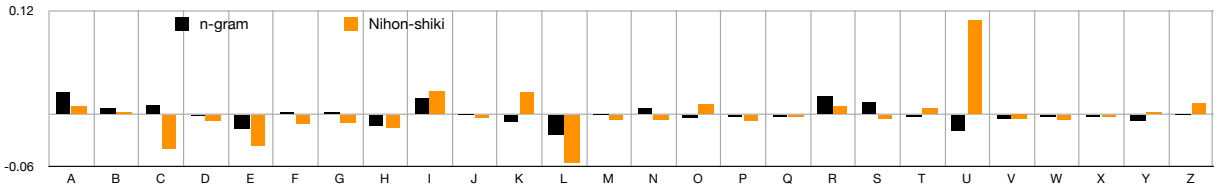

Figure 3: Character occurrence frequencies relative to English.

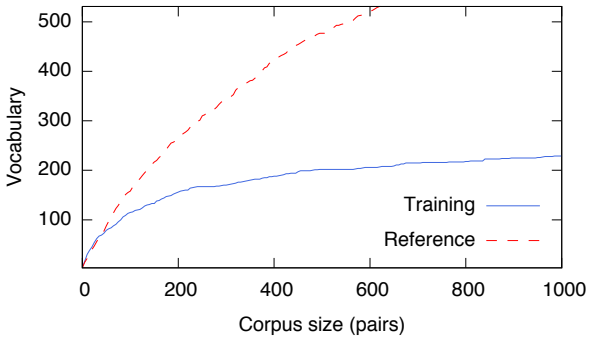

Figure 4: The relationship between corpus size and vocabulary size for the Chinese corpora.

The problem encountered with grapheme set size, and also the differences in characteristics between the training and reference sets is illustrated in Figure 4 which shows the relationship between corpus size and vocabulary size (the size of the set of single Kanji) for each corpus. It can be seen that the grapheme set size for the reference corpus exceeds 500 for the reference set, and is growing rapidly. In Chinese the full set of Kanji is more than 50,000 characters (Shu and Anderson, 1999). However, not all Kanji are used for the same purpose; for example, some are reserved for fortune-telling, and only a subset of them are typically used in transliteration. Nonetheless, as can be seen from Figure 4, the grapheme set size used for transliteration consists of several hundred characters. The grapheme set size for the training data is more limited, as it contains mainly foreign personal names, and country names. We will look in more detail at the effect of corpus size on the quality of the induced transliteration rules in Section 5.3.

5.2.2 Inducing Russian Romanization

The Russian language uses the Cyrillic alphabet in its writing system, and like Japanese there are several existing systems for romanization. In fact, recently a new romanization system was adopted as the standard for Russian international passports, we will call this system "Passport2010". For Russian, we used a bilingual corpus of NEWS2010 (Kumaran et al., 2010) data for inducing a Russian romanization system. We used the 1000 transliteration word pair corpus of seed data for training, and the test data consisted of the 885-pair reference data set. We compared our proposed system to the Passport2010 system, and also to a system that romanizes each Cyrillic character to the single English character that it most frequently aligns to.

Figure 5 presents ROC curves for each approach. All the different systems achieved approximately the same high level of performance on this task. This result reveals that there was little ambiguity in

Russian	А	Б	В	Г	Д	Е	Ё	Ж	З	И	Й	К	Л	М	Н	О
Passport2010	A	B	V	G	D	E	E	ZH	Z	I	I	K	L	M	N	O
Induced	A	B	V	G	D	E	O	Z	S	I	EI	K	L	M	N	O
Russian	П	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Ъ	Ы	Ь	Э	Ю	Я
Passport2010	P	R	S	T	U	F	KH	TC	CH	SH	–	Y	–	E	IU	IA
Induced	P	R	S	T	U	F	CH	C	CH	SH	Y	Y	L	E	U	A

Table 3: The Russian romanization rules for the official passport system, and a set of rules induced by our method.

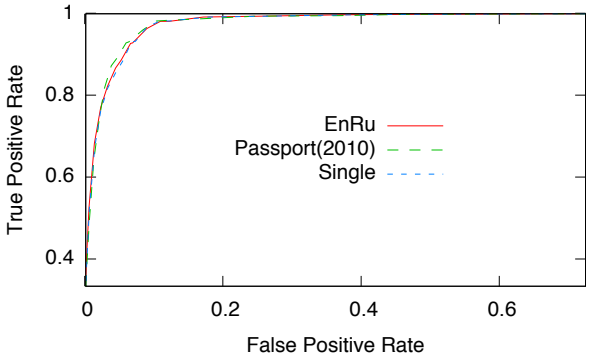

Figure 5: ROC curves for Russian.

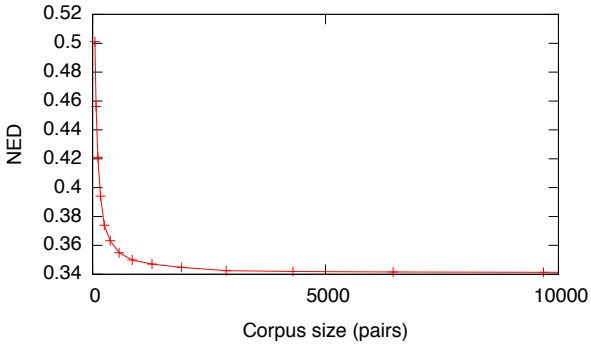

Figure 6: The relationship between corpus size and romanization quality for Japanese.

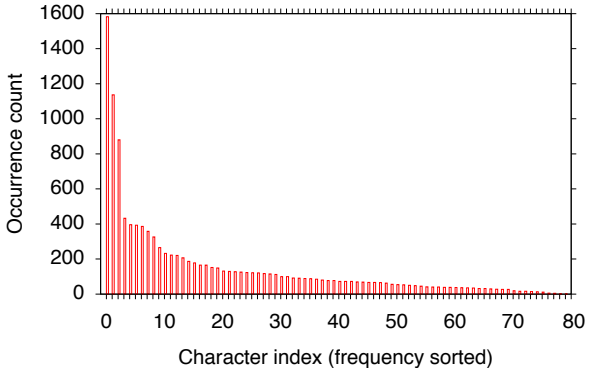

Figure 7: The distribution of occurrence counts for Japanese kana.

the bilingual alignment between Russian and English. This is in line with intuition because these language both belong to the same Indo-European language family. The full set of induced romanization rules is shown in Table 3 alongside the Passport2010 set. The 'Ъ' and 'Ь' characters have no direct phonetic value, and are left unromanized in Passport2010. Our approach has romanized them, however it would be more appropriate to leave them unromanized. This could be accomplished by extending our approach to allow NULL alignments.

Most of the rules are identical, and are simple conversions of the Cyrillic characters into their Roman counterparts. In most of the cases where the two systems differ, the Passport2010 system uses a longer form, making differences in pronunciation more explicit. There is major and interesting difference however: the Passport2010 system romanizes 'Ё' as 'E', whereas the induced system chose 'O'. Phonetically 'Ё' is a stressed /o/, as in 'York', and given the primary purpose of a romanization system for passports is to aid foreigners with the pronunciation of personal names, it may be that our induced system is indicating a more appropriate romanization for this character.

5.3 Required Corpus Size

The difficulties encountered in Section 5.2.1 suggest that for languages with large grapheme set sizes, the size of the training corpus will become an important factor. To study this effect, we performed an experiment using subsets of varying sizes sampled from a larger set of Japanese transliteration pairs than those used in Section 5.1 taken from Wikipedia inter-language links. The results are shown in Figure 6. Each point on the graph is the mean of 100 experiments each based on a random sample of the training data. The NED of the romanized corpus stops decreasing at

around 3,000-4,000 word pairs. Figure 7 shows the grapheme occurrence count distribution for an experiment that used 3200 training pairs. This distribution has a typical long-tailed form, with counts for the most frequent graphemes on the left-side of the graph, and the least frequent on the right-side. Most graphemes have occurred several hundred times, and therefore the amount of training data required depends on the amount of data available to train the rarer graphemes. The most frequent grapheme occurred 1583 times, the rarest only once, and the average occurrence count per grapheme was 146. There are approximately 42 transliteration pairs per grapheme.

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, Russian and Chinese for the purposes of transliteration mining. In these experiments we used corpora derived from of Wikipedia interlanguage link titles, and a criterium based on Levenshtein distance. For Japanese 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. For Chinese, our approach required more data than was available due to the large grapheme set size, and this motivated us to provide an analysis of the effect of corpus size on romanization quality. On Russian data our method was able to induce a system that was very close to the official system used for Russian passports. 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 extension 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

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

Finch, Andrew and Eiichiro Sumita. 2010. A Bayesian Model of Bilingual Segmentation for Transliteration. In Federico, Marcello, Ian Lane, Michael Paul, and François Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 259–266.

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

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

Htun, Ohnmar, Andrew Finch, Eiichiro Sumita, and Yoshiki Mikami. 2012. Improving transliteration mining by integrating expert knowledge with statistical approaches. *International Journal of Computer Applications*, 57, November.

Jiampojamarn, Sittichai, Kenneth Dwyer, Shane Bergsma, Aditya Bhargava, Qing Dou, Mi-Young Kim, and Grzegorz Kondrak. 2010. Transliteration generation and mining with limited training resources. In *Proceedings of the 2010 Named Entities Workshop*, pages 39–47. Association for Computational Linguistics.

Kumaran, A, Mitesh M Khapra, and Haizhou Li. 2010. Report of news 2010 transliteration mining shared task. In *Proceedings of the 2010 Named Entities Workshop*, pages 21–28. Association for Computational Linguistics.

Oo, Nandar Pwint and Ni Lar Thein. 2011. itextmm: Intelligent text input system for myanmar language on android smartphone. In *IT Convergence and Services*, pages 661–670. Springer.

Shu, Hua and Richard C Anderson. 1999. Learning to read chinese: The development of metalinguistic awareness. *Reading Chinese script: A cognitive analysis*, pages 1–18.