Robust Transliteration Mining from Comparable Corpora with Bilingual Topic Models

John Richardson[†] **Toshiaki Nakazawa**[‡] **Sadao Kurohashi**[†]
[†]Graduate School of Informatics, Kyoto University, Kyoto 606-8501
[‡]Japan Science and Technology Agency, Kawaguchi-shi, Saitama 332-0012
john@nlp.ist.i.kyoto-u.ac.jp, nakazawa@pa.jst.jp, kuro@i.kyoto-u.ac.jp

Abstract

We present a high-precision, language-independent transliteration framework applicable to bilingual lexicon extraction. Our approach is to employ a bilingual topic model to enhance the output of a state-of-the-art grapheme-based transliteration baseline. We demonstrate that this method is able to extract a high-quality bilingual lexicon from a comparable corpus, and we extend the topic model to propose a solution to the out-of-domain problem.

1 Introduction

A large, high-quality bilingual lexicon is of great utility to any dictionary-based system that processes bilingual data. The ability to automatically generate such a lexicon without relying on expensive training data or pre-existing lexical resources allows us to find translations for rare and unknown words with high efficiency.

Transliteration[1] is particularly important as new words are often created by importing words from other languages, especially English. It would be an almost impossible task to create and maintain a dictionary of such words by hand, as new words appear rapidly, especially in online texts, and word usage can vary over time.

In this paper we construct a language-independent transliteration framework. Our model builds on previous transliteration work, improving extraction and generation precision by including semantic as well as purely lexical features. The proposed model can be trained on comparable corpora, thereby not relying on expensive or often unavailable parallel data.

The motivation behind the approach of combining lexical and semantic features is that these two components are largely independent, greatly improving the effectiveness of their combination. This is particularly important for word-sense disambiguation. For example, a purely lexical approach is not sufficient to transliterate the Japanese ソース (*soosu*), as it can mean either 'sauce' or 'source' depending on the context.

2 Previous Work

Previous work has considered various methods for transliteration, ranging from simple edit distance and noisy-channel models (Brill et al., 2001) to conditional random fields (Ganesh et al., 2008) and finite state automata (Noeman and Madkour, 2010). We construct a baseline by modelling transliteration as a Phrase-Based Statistical Machine Translation (PB-SMT) task, a popular and well-studied approach (Matthews, 2007; Hong et al., 2009; Antony et al., 2010).

The vast majority of previous work on transliteration has considered only lexical features, for example spelling similarity and transliteration symbol mapping, however we build on the inspiration of Li et al. (2007) and later Hagiwara and Sekine (2012), who introduced semantic features to a transliteration model.

Li et al. (2007) proposed the concept of 'semantic transliteration', which is the consideration of inherent semantic information in transliterations. Their example is the influence of the source language and gender of foreign names on their transliterations into Chinese. Hagiwara and Sekine (2012) expanded upon this idea by considering a 'latent class'

[1]This paper considers both 'transliteration' (EN–XX) and 'back-transliteration' (XX–EN). For simplicity we refer to both tasks as 'transliteration'.

International Joint Conference on Natural Language Processing, pages 261–269,
Nagoya, Japan, 14-18 October 2013.

transliteration model considering transliterations to be grouped into categories, such as language of origin, which can give additional information about their formation. For example, if we know that a transliteration is of Italian origin, we are more likely to recover the letter sequence 'gli' than if it were originally French.

While these methods consider limited semantic features, they do not make use of the rich contextual information available from comparable corpora. We show such contextual information, in the form of bilingual topic distributions, to be highly effective in generating transliterations.

Bilingual lexicon mining from non-parallel data has been tackled in recent research such as Tamura et al. (2012) and Haghighi et al. (2008), and we build upon the techniques of multilingual topic extraction from Wikipedia pioneered by Ni et al. (2009). Previous research in lexicon mining has tended to focus on semantic features, such as context similarity vectors and topic models, but these have yet to be applied to the task of transliteration mining. We use the word-topic distribution similarities explored in Vulić et al. (2011) as baseline word similarity measures.

In some cases it is possible to use monolingual corpora for transliteration mining, as English is often written alongside transliterations (Kaji et al., 2011), however we consider the more general setting where such information is unavailable.

3 Baseline Transliteration Model

We begin by constructing a baseline transliteration system trained only on lexical features. This baseline system will allow us to compare directly the effectiveness of the addition of a semantic model to a traditional transliteration framework.

Our baseline model is a grapheme-based machine transliteration system. We model transliteration as a machine translation task on a character rather than word level, treating character groups as phrases. The model is trained by learning phrase alignments such as that shown in Figure 1. The field of phrase-based SMT has been well studied and there exists a standard toolset enabling the construction of an easily reproducable baseline system.

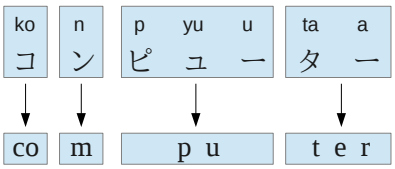

Figure 1: Example of Japanese–English transliteration phrase alignment.

We use the default configuration of Moses (Koehn et al., 2007) to train our baseline system, with the distortion limit set to 1 (as transliteration requires monotonic alignment). Character alignment is performed by GIZA++ (Och and Ney, 2003) with the 'grow-diag-final' heuristic for training. We apply standard tuning with MERT (Och, 2003) on the BLEU (Papineni et al., 2001) score. The language model is built with SRILM (Stolcke, 2002) using Kneser-Ney smoothing (Kneser and Ney, 1995).

The system described above has been implemented as specified in previous work such as Matthews (2007) (Chinese and Arabic), Hong et al. (2009) (Korean), and Antony et al. (2010) (Kannada). We demonstrate that this standard, highly-regarded baseline can be greatly improved with our proposed method.

4 Semantic Model

Having set up the baseline system, we turn to the task of combining a semantic model with our transliteration engine. We employ the method of bilingual LDA (Mimno et al., 2009), an extension of monolingual Latent Dirichlet Allocation (LDA) (Blei et al., 2003) as the semantic model.

Monolingual LDA takes as its input a set of monolingual documents and generates a word-topic distribution ϕ classifying words appearing in these documents into semantically similar topics. Bilingual LDA extends this by considering pairs of comparable documents in each of two languages, and outputs a pair of word-topic distributions ϕ and ψ, one for each input language. The graphical model for bilingual LDA is illustrated in Figure 2.

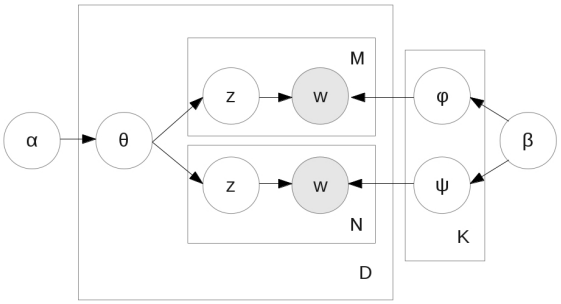

Figure 2: Graphical model for Bilingual LDA with K topics, D document pairs and hyperparameters α and β. Topics for each document are sampled from the common distribution θ, and the two languages have word-topic distributions ϕ and ψ.

4.1 Motivation for Bilingual LDA

We choose to employ a bilingual topic model to measure semantic similarity (i.e. topic similarity) of word pairs rather than the more intuitive method of comparing monolingual context similarity vectors (Rapp, 1995) for reasons of robustness and scalability.

Measuring context similarity on a word level requires a bilingual lexicon to match cross-language word pairs and such bilingual data is often expensive or unavailable. There are also problems with directly comparing collocations and word concurrence of distant language pairs as they do not always correspond predictably. Therefore our proposed method provides a more robust approach using coarser semantic features.

The use of topic models as a semantic similarity measure is a scalable method because document-aligned bilingual training data is growing ever more widely available. Examples of such sources are Wikipedia, multilingual newspaper articles and mined Web data.

4.2 Semantic Similarity Measures

In order to apply bilingual topic models to a transliteration task, we must construct an effective word similarity measure for source and target transliteration candidates. We improve upon three natural similarity measures, Cos, Cue and KL, based on those considered in Vulić et al. (2011), by proposing two methods of feature combination: reordering and SVM combination.

The reranking method considers hybrid scores Base+Cos, Base+Cue and Base+KL. These are generated by reranking the top-10 baseline (Base) transliteration candidates by their respective semantic scores (Cos, Cue or KL). We used 10 candidates for filtering as we found this gave the best balance between volume and accuracy in preliminary experiments. Approximately 75–85% of correct transliterations (depending on language pair) were within the top-10 candidates and this is therefore an upper bound for the hybrid model accuracy. As a comparison, the top-100 candidates contained roughly 80–85% of correct transliterations, the remainder failing to be identified by the baseline.

We additionally consider the combination of all three semantic features with the baseline (Moses) transliteration scores using a Support Vector Machine (SVM) (Vapnik, 1995). The SVM is used to classify candidate pairs as 'transliteration' (positive) or 'not transliteration' (negative), and we rerank the candidates by SVM predicted values. The features used for SVM training are baseline, Cos, Cue and KL scores.

The similarity measures Cos, Cue and KL are defined below.

4.2.1 Cos Similarity

The **Cos** method calculates the cosine similarity of the topic distribution vectors ψ_{k,w_e} and ϕ_{k,w_f} for transliteration pair candidates w_e and w_f.

$$Cos(w_e, w_f) = \frac{\sum_{k=1}^{K} \psi_{k,w_e} \phi_{k,w_f}}{\sqrt{\sum_{k=1}^{K} \psi_{k,w_e}^2} \sqrt{\sum_{k=1}^{K} \phi_{k,w_f}^2}} \quad (1)$$

4.2.2 Cue Similarity

The **Cue** method expresses the mean of the two probabilities $P(w_e \mid w_f)$ of a transliteration w_e given some source language string w_f and $P(w_f \mid w_e)$ of the reverse. We define:

$$P(w_e \mid w_f) = \sum_{k=1}^{K} \psi_{k,w_e} \frac{\phi_{k,w_f}}{Norm_\phi}$$

and likewise for $P(w_e \mid w_f)$, with the

normalization factors given by $Norm_\phi = \sum_{k=1}^{K} \phi_{k,w_f}$ and $Norm_\psi = \sum_{k=1}^{K} \psi_{k,w_e}$.

Finally, we consider:

$$Cue(w_e, w_f) = \frac{1}{2}(P(w_e \mid w_f) + P(w_f \mid w_e)) \quad (2)$$

4.2.3 KL Similarity

The **KL** method considers the averaged Kullback-Leibler divergence:

$$KL(w_e, w_f) = \frac{1}{2}(KL_{e,f} + KL_{f,e}) \quad (3)$$

$$KL_{e,f} = \sum_{k=1}^{K} \frac{\phi_{k,w_e}}{Norm_\phi} \log \frac{\phi_{k,w_e}/Norm_\phi}{\psi_{k,w_f}/Norm_\psi}$$

$$KL_{f,e} = \sum_{k=1}^{K} \frac{\psi_{k,w_f}}{Norm_\psi} \log \frac{\psi_{k,w_f}/Norm_\psi}{\phi_{k,w_e}/Norm_\phi}$$

using the same normalization factors as for Cue similarity.

5 Experiments

In order to demonstrate the effectiveness of our proposed model, we constructed an evaluation framework for a transliteration extraction task. The language pairs English–Japanese (EN–JA), Japanese–English (JA–EN), English–Korean (EN–KO) and Korean–English (KO–EN) were chosen to verify that this method is effective for a variety of languages and in both transliteration directions. Indeed, the methods introduced in this paper could also be applied directly to other languages with many transliterations, such as Chinese, Arabic and Hindi.

While it is possible to make language-specific optimizations, we decided only to preprocess the data minimally (such as removing punctuation) in order to demonstrate that our model works effectively in a language-independent setting. Examples of language-specific preprocessing techniques that we did not perform include segmentation of Japanese compound nouns (Nakazawa et al., 2005) and splitting of Korean syllabic blocks (*eumjeols*) into smaller components (*jamo*) (Hong et al., 2009).

Language Pairs	Train	Tune	Test
EN–JA/JA–EN	59K	1K	1K
KO–EN/EN–KO	21K	1K	1K

Table 1: Number of aligned word pairs in each fold of data.

5.1 Data Set

We chose to build our data set from Wikipedia articles, as they provide document-aligned comparable data across a variety of languages. Figure 3 shows how the Wikipedia data was split.

5.1.1 Baseline Training Data

We trained our baseline system on aligned Wikipedia page titles. This data consisted of pairs of English and Japanese/Korean words extracted from the freely available Wikipedia XML dumps. The aligned titles were filtered with hand-written rules[2] to extract only transliteration pairs, and the test data was verified for correctness by hand. This data will be made available to encourage comparison for future transliteration research[3].

The composition of this data is shown in Table 1. Aligned word pairs were shuffled randomly before splitting into the three folds to ensure an even topic distribution across each of 'Train', 'Tune' and 'Test'.

5.1.2 Bilingual Topic Model

The bilingual topic model was trained on the body text of Wikipedia articles aligned with Wikipedia inter-language links. These correspond to articles covering the same content, however they are rarely of similar length and not necessarily close transliterations.

We first pre-processed the most recent Wikipedia XML dumps to remove all tags and data other than plain text sentences, then aligned articles with language links to generate comparable document pairs. Words occurring fewer than 10 or more than 100K times were also removed to reduce noise and computation time.

[2]Heuristic rules included extraction of Japanese katakana, a script used primarily for transliterations, and words aligned with proper nouns as defined in a name dictionary.

[3]http://orchid.kuee.kyoto-u.ac.jp/~john

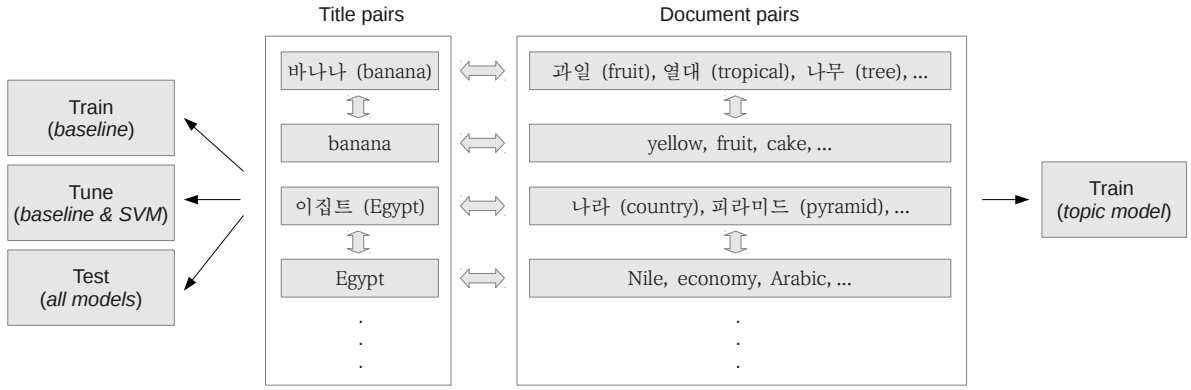

Figure 3: We extracted aligned title pairs (only transliterations) and aligned document pairs from Wikipedia using inter-language links. The baseline was trained and tuned on title pairs ('Train' and 'Tune'), the topic model was trained on document pairs and the SVM was trained on the title pairs 'Tune' fold.

5.1.3 SVM Hybrid Model

The training data for the proposed SVM hybrid model was built from the same data used for the baseline (tuning fold). We first generated the top-10 distinct transliteration candidates for the tuning data using the 'n-best-list' option in Moses. These candidates were then labeled as 'transliteration' or 'not-transliteration' and feature scores (Base, Cos, Cue, KL) were generated for each candidate. The SVM model was trained using these labels and feature scores.

5.2 LDA Implementation Details

PolyLDA++, our implementation of multilingual LDA, was based on GibbsLDA++ (Phan et al., 2007), a toolkit for monolingual LDA. This software is available for free[4].

Each topic model was trained over 1000 iterations, and the standard Dirichlet prior hyperparameters for the LDA model were set as $\alpha = 50/K$ for K topics and $\beta = 0.1$.

The choice of number of topics is important, as demonstrated in Figure 4, which shows the top-1 accuracy of the SVM hybrid model using various numbers of topics K. The optimal value of K seems to be between around 100 for this data.

The model accuracy gradually decreases with adding more than 100 topics. We believe that this is because the granularity of the topics becomes too fine to accommodate for the wide differences in semantic usage of English and Japanese/Korean transliteration pairs. A higher number of topics could be more suitable for more closely related language pairs, such as Italian and English (Vulić et al., 2011), because the higher similarity of word usage would allow for topics of more limited semantic scope. Such experiments are to be considered in future work. The results below are for $K = 100$.

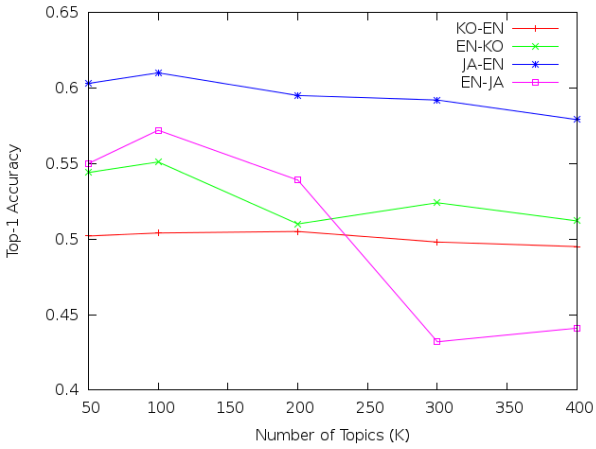

Figure 4: Top-1 accuracy of SVM for various K.

5.3 Results

Table 2 compares the top-1 accuracy of our proposed hybrid models to the baseline perfor-

[4]http://orchid.kuee.kyoto-u.ac.jp/~john

	JA–EN	EN–JA	KO–EN	EN–KO
Base	0.334	0.363	0.296	0.421
Base+Cos	0.608	0.559	0.494	0.516
Base+Cue	0.608	0.551	**0.507**	0.504
Base+KL	0.365	0.398	0.261	0.373
SVM	**0.610**	**0.572**	0.504	**0.551**

Table 2: Top-1 accuracy of proposed model for each hybrid scoring method.

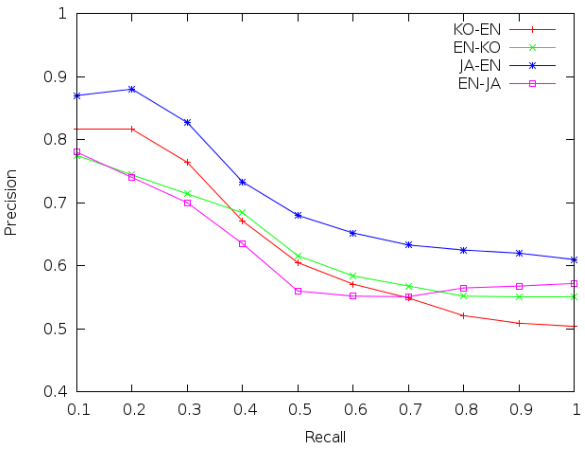

Figure 5: Precision-recall curve for SVM hybrid model.

mance. The SVM hybrid model outperformed the baseline for every language pair, by as much as 0.276 for JA–EN. This suggests that the addition of a bilingual topic model significantly improves transliteration accuracy.

In general the SVM was the most effective hybrid score, outperforming Base+Cos, Base+Cue and Base+KL in all but KO–EN, where it performed very slightly worse than Base+Cue.

Figure 5 shows the precision-recall curve for the SVM hybrid model over the test set. We vary recall by ranking the hybrid model scores for all test pairs and selecting only the highest scoring fraction to evaluate. This simulates a lexicon extraction task where we wish to sacrifice recall for precision. The results demonstrate that it is possible to improve significantly the precision of a set of extracted transliterations by reducing the recall. This large improvement is made possible because the topic similarity scores are particularly effective at measuring confidence in each transliteration candidate, allowing effective selection of the correct transliterations.

5.4 Comparison with Previous Work

The results compare favorably to the top-1 accuracy of similar existing systems, such as DirecTL+ (Jiampojamarn et al., 2010), which also used Wikipedia titles (EN–JA 0.398), and Hagiwara and Sekine (2012) (EN–JA 0.349).

Our baseline transliteration system can be measured against previous work using Moses and GIZA++ alignment, such as Matthews (2007) (EN–AR 0.43, AR–EN 0.39, EN–ZH 0.38, ZH–EN 0.35) and Hong et al. (2009) (EN–KO 0.45). These scores are consistent with our baseline results.

While it is difficult to compare directly the accuracy of transliteration systems across different languages and data sets, especially since we use additional data to train the semantic model, the results above show that our model has made a considerable improvement over the state-of-the-art baseline.

6 Extension to Out-of-Domain Words

The model described in this paper revolves around the use of a bilingual topic model to improve transliteration quality. What happens then when a source word is not covered by the topic model? This is a very important problem in a practical setting, and we show that even in such cases our model can improve considerably upon the baseline system. We define 'out-of-domain' words as source language words that did not appear in the topic model training data and hence do not have a known topic distribution.

6.1 Model Details

Our proposed approach is to consider not the word-topic distribution of the source word w_e itself, but rather that of the words in the surrounding context. We consider two methods for calculating the modified topic similarity

scores over the set of words W_e in the same context as the source word.

Let $S(w_e, w_f)$ be a basic topic similarity score Cos, Cue or KL, then we define the extended scores $ExtMean(W_e, w_f)$ and $ExtWeight(W_e, w_f)$ as follows:

$$ExtMean(W_e, w_f) = \frac{\sum_{w_e \in W_e} S(w_e, w_f)}{|W_e|} \quad (4)$$

$$ExtWeight(W_e, w_f) = \frac{\sum_{w_e \in W_e} c'_{w_e} S(w_e, w_f)}{\sum_{w_e \in W_e} c'_{w_e}} \quad (5)$$

where $c'_{w_e} = (\log c_{w_e})^{-1}$ for the frequency c_{w_e} of w_e appearing in the semantic model training data.

ExtMean corresponds to the mean topic similarity for each word in the context W_e. ExtWeight is weighted by the inverse log frequency of each word, allowing consideration of their semantic importance. These extended scores are used to train the SVM in place of the original scores.

6.2 Out-of-Domain Experiment

We performed an additional experiment where we transliterated a set of 25 Japanese words unknown to the topic model into English. These words appeared in Wikipedia fewer than 10 times and therefore were not included in our training data. We extracted the sentences and documents in which these words occurred, and back-transliterated the Japanese words into English by hand. We considered both sentence-level and document-level contexts for W_e, and evaluated each extended metric ExtMean and ExtWeight.

The results of the out-of-domain experiment are shown in Table 3, which gives the top-1 accuracy of the SVM hybrid model trained on the ExtMean and ExtWeight counterparts of Cos, Cue and KL similarities. Base is the top-1 accuracy using only the Moses baseline.

The most effective settings were to use ExtWeight on a sentence level context. There is a balance between size and relevance of context, with document-level context containing too many misleading words. The improvement of ExtWeight over ExtMean shows the importance of weighting contextual words based on their importance (i.e. inverse log frequency).

	Base	ExtMean	ExtWeight
Document	0.27	0.44	0.48
Sentence		0.48	0.52

Table 3: Top-1 accuracy for out-of-domain model extension (JA–EN).

The results show a large improvement (+0.25) over the baseline scores that is comparable to that of the in-domain model (+0.28, see Table 2). This suggests that the proposed model is an effective solution to the out-of-domain problem.

7 Discussion and Error Analysis

An example of the top candidates for a successful and an incorrect transliteration are given in Tables 4 and 5 respectively. We can see that the topic model has succeeded in finding the correct transliteration of 'batik', a traditional Javanese fabric, however a low score was given to the Korean transliteration of the name 'Bernard' appearing in the training data.

The benefits of the addition of a topic model is made clear with the example of 'batik' in Table 4. The semantic similarity measures give a higher score to 'batik' than 'Batic', a Slavic surname, despite 'Batic' being the more likely transliteration according to the baseline.

The improvement over the baseline for back-transliteration (XX–EN), on average +0.24, was considerably greater than that for transliteration (EN–XX), on average +0.17. We believe that this is due to the vast range of transliteration spelling variations in the non-English target languages. Since there is only one correct spelling variation defined in our test data and the topic distributions for each spelling variation are very similar, it is not possible to guess correctly. For an example of this problem, see Table 5.

7.1 Topic Alignment Difficulties

The majority of transliteration errors were caused by unsuccessful topic alignment between the source and target words. This was partly caused by the differences in usage of the original English words and the transliterated Japanese or Korean. For example, the

Candidate	Baseline	Cos	Cue	KL	SVM
batik	-1.29	**0.989**	**2.54e-04**	**-0.327**	**1.10**
baetic	-1.32	0.0764	1.67e-06	-1.65	-1.39
batic	**-0.708**	0.00	0.0	0.0	-1.48
batick	-0.788	0.00	0.0	0.0	-1.53
butic	-1.09	0.00	0.0	0.0	-1.68

Table 4: A good transliteration – バティック (*batikku* / 'batik') → batik.

Candidate	Baseline	Cos	Cue	KL	SVM
베르나르 *bereunareu*	-2.96	**0.642**	**4.78e-04**	**-1.72**	**0.112**
베르나르드 *bereunareudeu*	-3.65	0.243	3.84e-05	-2.41	-0.909
베른하르트 *bereunhareuteu*	-3.58	0.188	7.64e-05	-1.81	-0.969
베르나르트 *bereunareuteu*	-4.24	0.217	8.24e-05	-2.69	-1.02
버나드 *beonadeu*	**-2.78**	0.123	4.33e-05	-3.01	-1.23

Table 5: An incorrect transliteration – bernard → 베르나르트 (*bereunareuteu*).

Japanese バイキング (*baikingu*) is a transliteration of 'Viking', however it is almost always used to mean 'buffet', deriving from the Scandinavian smorgasbord. In this case, we can expect the Japanese to be associated with food-related topics, quite different from 'Viking'.

There are also many cases where words that do not clearly fit into one topic have unclear distributions across many groups. For example, the word 로마 (*roma* / 'Rome') could be more strongly categorized with 'cities' and 'sightseeing' in English but 'history' and 'classical civilization' in Korean, giving a low overall topic correlation.

7.2 Effect of Word Length and Frequency

We found that our model was more successful at finding the correct transliteration of longer words, as smaller words tend to have more spelling variations and are orthographically more similar to other words. By removing words of length 5 characters or less from the test data, we were able to improve the top-1 accuracy (SVM) to 0.593 (KO–EN, +0.089) and 0.721 (JA–EN, +0.111). In a practical lexicon extraction task over the entirety of Wikipedia this would cover roughly 35–45% of words (depending on language).

There was almost no variation in transliteration accuracy based on word frequency. The baseline is relatively unaffected by word frequency, with the exception of finding very rare character phrases not in the training data, and the topic model proved to be robust across words of both high and low frequency.

8 Conclusion and Future Work

In this paper we demonstrated that the addition of semantic features can significantly improve transliteration accuracy. Specifically, it is possible to outperform the top-1 accuracy of a state-of-the-art phrase-based SMT transliteration baseline through the addition of a bilingual topic model.

Furthermore, our extended model is able to produce a considerable improvement in accuracy even for out-of-domain source words that have an unknown topic distribution. The experimental data set was constructed to simulate the task of extracting unknown word pairs from a comparable corpus, however our extension model results suggest that it will be possible to extract high-quality transliterations from larger and less comparable corpora than ever before.

In the future we would like to explore in depth the improvements to machine translation made possible by this approach.

Acknowledgements

We would like to thank the anonymous reviewers for their feedback. The first author is supported by a Japanese Government (MEXT) research scholarship.

References

P.J. Antony, V.P. Ajith, and K.P. Soman 2010. Statistical Method for English to Kannada Transliteration. *BAIP 2010, CCIS 70, pp. 356–362.*

David Blei, Andrew Ng and Michael Jordan. 2003. Latent Dirichlet Allocation. In *The Journal of Machine Learning Research*, Volume 3.

Eric Brill, Gary Kacmarcik and Chris Brockett. 2001. Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, 2001.

Surya Ganesh, Sree Harsha, Prasad Pingali, Vasudeva Varma. 2008. Statistical Transliteration for Cross Language Information Retrieval using HMM alignment model and CRF. In *2nd International Workshop on Cross Language Information Access, IJCNLP 2008.*

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick and Dan Klein. 2008. Learning Bilingual Lexicons from Monolingual Corpora. In *ACL 2008.*

Masato Hagiwara and Satoshi Sekine. 2011. Latent Class Transliteration based on Source Language Origin. In *ACL 2011.*

Masato Hagiwara and Satoshi Sekine. 2012. Latent Semantic Transliteration using Dirichlet Mixture. In *ACL 2012.*

Gumwon Hong, Min-Jeong Kim, Do-Gil Lee and Hae-Chang Rim. 2009. A Hybrid Approach to English-Korean Name Transliteration. In *Proceedings of 2009 Named Entities Workshop, ACL-IJCNLP.*

Sittichai Jiampojamarn, 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, ACL 2010.*

Nobuhiro Kaji and Masaru Kitsuregawa. 2011. Splitting Noun Compounds via Monolingual and Bilingual Paraphrasing: A Study on Japanese Katakana Words. In *EMNLP 2011.*

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1.*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007.*

Haizhou Li, Khe Chai Sim, Jin-Shea Kuo, Minghui Dong. 2007. Semantic Transliteration of Personal Names. In *ACL 2007.*

David Matthews. 2007. Machine Transliteration of Proper Names. *Masters Thesis, School of Informatics, University of Edinburgh.*

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP 2009.*

Toshiaki Nakazawa, Daisuke Kawahara and Sadao Kurohashi. 2005. Automatic Acquisition of Basic Katakana Lexicon from a Given Corpus. In *IJCNLP 2005.*

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, Zheng Chen. 2009. Mining Multilingual Topics from Wikipedia. In *WWW 2009.*

Sara Noeman and Amgad Madkour. 2010. Language independent transliteration mining system using finite state automata framework. In *Proceedings of the 2010 Named Entities Workshop.*

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *ACL 2003.*

Franz Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics 2003.*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of Machine Translation. *Technical Report RC22176, IBM.*

Xuan-Hieu Phan and Cam-Tu Nguyen. 2007. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA).

Reinhard Rapp. 1995. Identifying Word Translations in Non-Parallel Texts. In *ACL 1995.*

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of ICSLP, Volume 2.*

Akihiro Tamura, Taro Watanabe and Eiichiro Sumita. 2012. Bilingual Lexicon Extraction from Comparable Corpora Using Label Propagation. In *EMNLP-CoNLL 2012.*

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.

Ivan Vulić, Wim De Smet and Marie-Francine Moens. 2011. Identifying Word Translations from Comparable Corpora Using Latent Topic Models. In *ACL 2011.*