# Noise-aware Character Alignment for Bootstrapping Statistical Machine Transliteration from Bilingual Corpora

**Katsuhito Sudoh**[*†] **Shinsuke Mori**[‡] **Masaaki Nagata**[*]
[*]NTT Communication Science Laboratories
[†]Graduate School of Informatics, Kyoto University
[‡]Academic Center for Computing and Media Studies, Kyoto University
sudoh.katsuhito@lab.ntt.co.jp

## Abstract

This paper proposes a novel noise-aware character alignment method for bootstrapping statistical machine transliteration from automatically extracted phrase pairs. The model is an extension of a Bayesian many-to-many alignment method for distinguishing non-transliteration (noise) parts in phrase pairs. It worked effectively in the experiments of bootstrapping Japanese-to-English statistical machine transliteration in patent domain using patent bilingual corpora.

## 1 Introduction

Transliteration is used for providing translations for source language words that have no appropriate counterparts in target language, such as some technical terms and named entities. Statistical machine transliteration (Knight and Graehl, 1998) is a technology to solve it in a statistical manner. Bilingual dictionaries can be used to train its model, but many of their entries are actually *translation* but not *transliteration*. Such non-transliteration pairs hurt the transliteration model and should be eliminated beforehand.

Sajjad et al. (2012) proposed a method to identify such non-transliteration pairs, and applied it successfully to *noisy* word pairs obtained from automatic word alignment on bilingual corpora. It enables the statistical machine transliteration to be bootstrapped from bilingual corpora. This approach is beneficial because it does not require carefully-developed bilingual transliteration dictionaries and it can learn domain-specific transliteration patterns from bilingual corpora in the target domain. However, their transliteration mining approach is sample-wise; that is, it makes a decision whether a bilingual phrase pair is transliteration or not. Suppose that a compound word in a language A is transliterated into two words in another language B. Their correspondence may not be fully identified by automatic word alignment and a wrong alignment between the compound word in A and only one component word in B is found. The sample-wise mining cannot make a correct decision of *partial transliteration* on the aligned candidate, and may introduces noise to the statistical transliteration model.

This paper proposes a novel transliteration mining method for such partial transliterations. The method uses a noise-aware character alignment model that distinguish non-transliteration (noise) parts from transliteration (signal) parts. The model is an extension of a Bayesian alignment model (Finch and Sumita, 2010) and can be trained by a sampling algorithm extended for a constraint on noise. Our experiments of Japanese-to-English transliteration achieved 16% relative error reduction in transliteration accuracy from the sample-wise method. The main contribution of this paper is two-fold:

- we formulate alignment over string pairs with partial noise and present a solution with a noise-aware alignment model;
- we proved its effectiveness by experiments with frequent unknown words in actual Japanese-to-English patent translation data.

*Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 204–209,
Seattle, Washington, USA, 18-21 October 2013.

## 2 Bayesian many-to-many alignment

We briefly review a Bayesian many-to-many character alignment proposed by Finch and Sumita (2010) on which our model is based. The model is based on a generative process of bilingual substring pairs $\langle \bar{\boldsymbol{s}}, \bar{\boldsymbol{t}} \rangle$ by the following Dirichlet process (DP):

$$\begin{aligned} G|_{\alpha,G_0} &\sim \mathrm{DP}(\alpha, G_0) \\ \langle \bar{\boldsymbol{s}}, \bar{\boldsymbol{t}} \rangle | G &\sim G, \end{aligned}$$

where G is a probability distribution over substring pairs according to a DP prior with base measure $G_0$ and hyperparameter $\alpha$. $G_0$ is modeled as a joint spelling model as follows:

$$G_0 \left( \langle \bar{\boldsymbol{s}}, \bar{\boldsymbol{t}} \rangle \right) = \frac{\lambda_s^{|\bar{\boldsymbol{s}}|}}{|\bar{\boldsymbol{s}}|!} e^{-\lambda_s} v_s^{-|\bar{\boldsymbol{s}}|} \times \frac{\lambda_t^{|\bar{\boldsymbol{t}}|}}{|\bar{\boldsymbol{t}}|!} e^{-\lambda_t} v_t^{-|\bar{\boldsymbol{t}}|}. \quad (1)$$

This is a simple joint probability of the spelling models, in which each alphabet appears based on a uniform distribution over the vocabulary (of size $v_s$ and $v_t$) and each string length follows a Poisson distribution (with the average length $\lambda_s$ and $\lambda_t$).

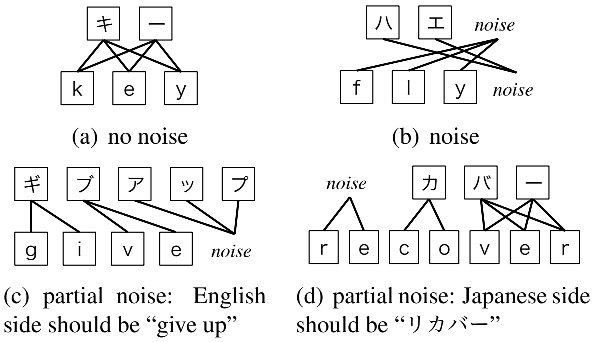
The model handles infinite number of substring pairs according to the Chinese Restaurant Process (CRP). The probability of a substring pair $\langle \bar{s_k}, \bar{\boldsymbol{t}_k} \rangle$ is based on the counts of all other substring pairs as follows:

$$\begin{aligned} &p \left( \langle \bar{\boldsymbol{s}}_k, \bar{\boldsymbol{t}}_k \rangle | \{ \langle \bar{\boldsymbol{s}}, \bar{\boldsymbol{t}} \rangle \}_{-k} \right) \\ &\qquad = \frac{N \left( \langle \bar{\boldsymbol{s}}_k, \bar{\boldsymbol{t}}_k \rangle \right) + \alpha G_0 \left( \langle \bar{\boldsymbol{s}}_k, \bar{\boldsymbol{t}}_k \rangle \right)}{\sum_i N \left( \langle \bar{\boldsymbol{s}}_i, \bar{\boldsymbol{t}}_i \rangle \right) + \alpha}. \quad (2) \end{aligned}$$

Here $\{ \langle \bar{\boldsymbol{s}}, \bar{\boldsymbol{t}} \rangle \}_{-k}$ means a set of substring pairs excluding $\langle \bar{s_k}, \bar{\boldsymbol{t}_k} \rangle$, and $N \left( \langle \bar{\boldsymbol{s}}_k, \bar{\boldsymbol{t}}_k \rangle \right)$ is the number of $\langle \bar{s_k}, \bar{\boldsymbol{t}_k} \rangle$ in the current sample space. This alignment model is suitable for representing very sparse distribution over arbitrary substring pairs, thanks to reasonable CRP-based smoothing for unseen pairs based on the spelling model.

## 3 Proposed method

We propose an extended many-to-many alignment model that can handle partial noise. We extend the model in the previous section by introducing a noise symbol and state-based probability calculation.



(a) no noise

(b) noise

(c) partial noise: English side should be "give up"

(d) partial noise: Japanese side should be "リカバー"

Figure 1: Three types of noise in transliteration data. Solid lines are correct many-to-many alignment links.

### 3.1 Partial noise in transliteration data

Figure 1 shows transliteration examples with "no noise," "noise," and "partial noise." Solid lines in the figure show correct many-to-many alignment links. The examples (a) and (b) can be distinguished effectively by Sajjad et al. (2012). We aim to do alignment as in the examples (c) and (d) by distinguishing its non-transliteration (noise) part, which cannot be handled by the existing methods.

### 3.2 Noise-aware alignment model

We introduce a *noise symbol* to handle partial noise in the many-to-many alignment model. Htun et al. (2012) extended the many-to-many alignment for the sample-wise transliteration mining, but its noise model only handles the sample-wise noise and cannot distinguish partial noise. We model partial noise in the CRP-based joint substring model.

Partial noise in transliteration data typically appears in compound words as mentioned earlier, because their counterparts consisting of two or more words may not be fully covered in automatically extracted words and phrases as shown in Figure 1(c). Another type of partial noise is derived from morphological differences due to inflection, which usually appear in the sub-word level as prefixes and suffixes as shown in Figure 1(d). According to this intuition, we assume that partial noise appears in the beginning and/or end of transliteration data (in case of sample-wise noise, we assume the noise is in the beginning). This assumption derives a constraint between signal and noise parts that helps to avoid a welter of transliteration and non-transliteration parts. It also has a shortcoming that it is generally

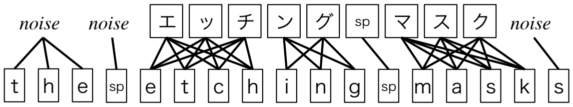


Figure 2: Example of many-to-many alignment with partial noise in the beginning and end. "*noise*" stands for the noise symbol and "sp" stands for a white space.

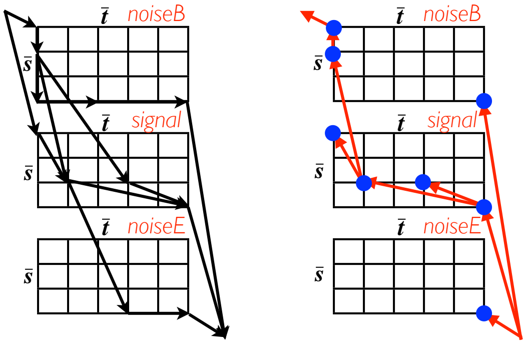
not appropriate for noise in the middle, but handling arbitrary number of noise parts increases computational complexity and sparseness. We rely on this simple assumption in this paper and consider a more complex mid-noise problem as future work.

Figure 2 shows a partial noise example in both the beginning and end. This example is actually correct translation but includes noise in a sense of transliteration; an article "the" is wrongly included in the phrase pair (no articles are used in Japanese) and a plural noun "masks" is transliterated into "マスク"(mask). These non-transliteration parts are aligned to noise symbols in the proposed model. The noise symbols are treated as zero-length substrings in the model, same as other substrings.

### 3.3 Constrained Gibbs sampling

Finch and Sumita (2010) used a blocked Gibbs sampling algorithm with forward-filtering backward-sampling (FFBS) (Mochihashi et al., 2009). We extend their algorithm for our noise-aware model using a state-based calculation over the three states: non-transliteration part in the beginning (noiseB), transliteration part (signal), non-transliteration part in the end (noiseE).

Figure 3 illustrates our FFBS steps. At first in the forward filtering, we begin with transition to noiseB and signal. The calculation of forward probabilities itself is almost the same as Finch and Sumita (2010) except for state transition constraints: from noiseB to signal, from signal to noiseE. The backward-sampling traverses a path by probability-based sampling with true posteriors, starting from the choice of the ending state among noiseB (means full noise), signal, and noiseE. This algorithm increases the computational cost by three times to consider three different states, compared to that of Finch and Sumita (2010).



(a) Forward filtering (b) Backward sampling

Figure 3: State-based FFBS for the proposed model.

## 4 Experiments

We conducted experiments comparing the proposed method with the conventional sample-wise method for the use in bootstrapping statistical machine transliteration using Japanese-to-English patent translation dataset (Goto et al., 2013).

### 4.1 Training data setup

First, we trained a phrase table on the 3.2M parallel sentences by a standard training procedure using Moses, with Japanese tokenization using MeCab[1]. We obtained 591,840 phrase table entries whose Japanese side was written in *katakana* (Japanese phonogram) only[2]. Then, we iteratively ran the method of Sajjad et al. (2012) on these entries and eliminate non-transliteration pairs, until the number of pairs converged. Finally we obtain 104,563 *katakana*-English pairs after 10 iterations; they were our *baseline training set* mined by sample-wise method. We used Sajjad et al.'s method as preprocessing for filtering sample-wise noise while the proposed method could also do that, because the proposed method took much more training time for all phrase table entries.

### 4.2 Transliteration experiments

The transliteration experiment used a translation-based implementation with Moses, using a

[1] http://code.google.com/p/mecab/

[2] This *katakana*-based filtering is a language dependent heuristic for choosing potential transliteration candidate, because transliterations in Japanese are usually written in *katakana*.

character-based 7-gram language model trained on 300M English patent sentences. We compared three transliteration models below.

The test set was top-1000 unknown (in the Japanese-to-English translation model) *katakana* words appeared in 400M Japanese patent sentences. They covered 15.5% of all unknown *katakana* words and 8.8% of all unknown words (excluding numbers); that is, more than a half of unknown words were *katakana* words.

#### 4.2.1 Sample-wise method (BASELINE)

We used the baseline training set to train statistical machine transliteration model for our baseline. The training procedure was based on Moses: MGIZA++ word alignment, grow-diag-final-and alignment symmetrization and phrase extraction with the maximum phrase length of 7.

#### 4.2.2 Proposed method (PROPOSED)

We applied the proposed method to the baseline training set with 30 sampling iterations and eliminated partial noise. The transliteration model was trained in the same manner as BASELINE after eliminating noise.

The hyperparameters, $\alpha$, $\lambda_s$, and $\lambda_t$, were optimized using a held-out set of 2,000 *katakana*-English pairs that were randomly chosen from a general-domain bilingual dictionary. The hyperparameter optimization was based on F-score values on the held-out set with varying $\alpha$ among 0.01, 0.02, 0.05, 0.1, 1.0, and $\lambda$s among 1, 2, 3, 5.

Table 1 compares the statistics on the training sets of BASELINE and PROPOSED. Note that we applied the proposed method to BASELINE data (the sample-wise method was already applied until convergence). The proposed method eliminated only two transliteration candidates in sample-wise but also eliminated 5,714 (0.64%) *katakana* and 55,737 (4.1%) English characters[3].

#### 4.2.3 Proposed method using aligned joint substrings as phrases (PROPOSED-JOINT)

The many-to-many character alignment actually induces substring pairs, which can be used as phrases in statistical machine transliteration and improved transliteration performance (Finch and Sumita, 2010). We extracted them by: 1) generate many-to-many word alignment, in which all possible word alignment links in many-to-many correspondences (e.g., 0-0 0-1 0-2 1-0 1-1 1-2 for ⟨コ ン, c o m⟩), 2) run phrase extraction and scoring same as a standard Moses training. This procedure extracts longer phrases satisfying the many-to-many alignment constraints than the simple use of extracted joint substring pairs as phrases.

Table 1: Statistics of the training sets.

| Method | #pairs | #Ja chars. | #En chars. |
|---|---|---|---|
| BASELINE | 104,563 | 899,080 | 1,372,993 |
| PROPOSED | 104,561 | 893,366 | 1,317,256 |

### 4.3 Results

Table 2 shows the results. We used three evaluation metrics: ACC, F-score, and BLEU$_\mathrm{c}$. ACC is a sample-wise accuracy and F-score is a character-wise F-measure-like score (Li et al., 2010). BLEU$_\mathrm{c}$ is BLEU (Papineni et al., 2002) in the character level with $n$=4.

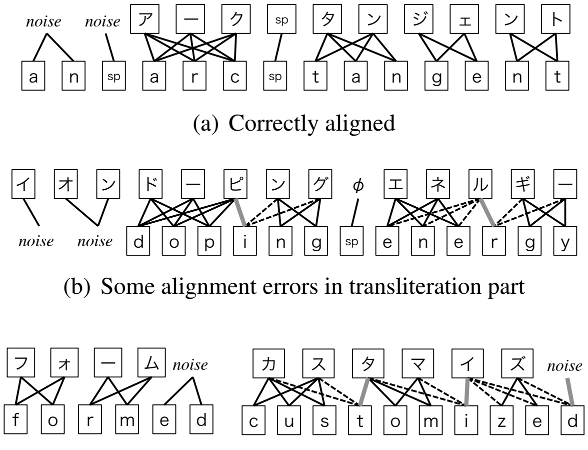
PROPOSED achieved 63% in ACC (16% relative error reduction from BASELINE), and 94.6% in F-score (25% relative error reduction from BASELINE). These improvements clearly showed an advantage of the proposed method over the sample-wise mining. BLEU$_\mathrm{c}$ showed a similar improvements. Recall that BASELINE and PROPOSED had a small difference in their training data, actually 0.64% (*katakana*) and 4.1% (English) in the number of characters. The results suggest that the partial noise can hurt transliteration models.

PROPOSED-JOINT showed similar performance as PROPOSED with a slight drop in BLEU$_\mathrm{c}$, although many-to-many substring alignment was expected to improve transliteration as reported by Finch and Sumita (2010). The difference may be due to the difference in coverage of the phrase tables; PROPOSED-JOINT retained relatively long substrings by the many-to-many alignment constraints in contrast to the less-constrained grow-diag-final-and alignments in PROPOSED. Since the training data in our bootstrapping experiments con-

[3] The reason of larger number of partial noise in English side would be a syntactic difference as shown in Figure 2 and the *katakana*-based filtering heuristics.

Table 2: Japanese-to-English transliteration results for top-1000 unknown *katakana* words. ACC and F-score stand for the ones used in NEWS workshop, $\text{BLEU}_\text{c}$ is character-wise BLEU.

| Method | ACC | F-score | $\text{BLEU}_\text{c}$ |
|---|---|---|---|
| BASELINE | 0.56 | 0.929 | 0.864 |
| PROPOSED | 0.63 | 0.946 | 0.897 |
| PROPOSED-JOINT | 0.63 | 0.943 | 0.888 |


noise noise ア ー ク sp タ ン ジ ェ ン ト
a n sp a r c sp t a n g e n t
イ オ ン ド ー ピ ン グ φ エ ネ ル ギ ー
noise noise d o p i n g sp e n e r g y
フ ォ ー ム noise
f o r m e d
カ ス タ マ イ ズ noise
c u s t o m i z e d

(a) Correctly aligned

(b) Some alignment errors in transliteration part

(c) Correctly aligned

(d) Errors in partial noise

Figure 4: Examples of noise-aware many-to-many alignment in the training data. $\phi$ stands for a zero-length substring. Dashed lines show incorrect alignments, and bold grey lines mean their corrections.

tained many similar phrases unlike dictionary-based data in Finch and Sumita (2010), the phrase table of PROPOSED-JOINT may have a small coverage due to long and sparse substring pairs with large probabilities even if the many-to-many alignment was good. This sparseness problem is beyond the scope of this paper and worth further study.

### 4.4 Alignment Examples

Figure 4 shows examples of the alignment results in the training data. As expected, partial noise both in Japanese and English was identified correctly in (a), (b), and (c). There were some alignment errors in the signal part in (b), in which characters in boundary positions were aligned incorrectly to adjacent substrings. These alignment errors did not directly degrade the partial noise identification but may cause a negative effect on overall alignment performance in the sampling-based optimization. (d) is a negative example in which partial noise was incorrectly aligned. (c) and (d) have similar partial noise in their English word endings, but it could not be identified in (d). One possible reason for that is the sparseness problem mentioned above, as shown in erroneous long character alignments in (d).

## 5 Conclusion

This paper proposed a noise-aware many-to-many alignment model that can distinguish partial noise in transliteration pairs for bootstrapping statistical machine transliteration model from automatically extracted phrase pairs. The model and training algorithm are straightforward extension of those by Finch and Sumita (2010). The proposed method was proved to be effective in Japanese-to-English transliteration experiments in patent domain.

Future work will investigate the proposed method in other domains and language pairs. The partial noise would appear in other language pairs, typically between agglutinative and non-agglutinative languages. It is also worth extending the approach into word alignment in statistical machine translation.

## Acknowledgments

We would like to thank anonymous reviewers for their valuable comments and suggestions.

## References

Andrew Finch and Eiichiro Sumita. 2010. A Bayesian Model of Bilingual Segmentation for Transliteration. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 259–266.

Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. 2013. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In *The 10th NTCIR Conference*, June.

Ohnmar Htun, Andrew Finch, Eiichiro Sumita, and Yoshiki Mikami. 2012. Improving Transliteration Mining by Integrating Expert Knowledge with Statistical Approaches. *International Journal of Computer Applications*, 58(17):12–22, November.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Haizhou Li, A Kumaran, Min Zhang, and Vladimir Pervouchine. 2010. Whitepaper of NEWS 2010 Shared Task on Transliteration Generation. In *Proceedings of the 2010 Named Entities Workshop*, pages 12–20, Uppsala, Sweden, July. Association for Computational Linguistics.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 100–108, Suntec, Singapore, August. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2012. A statistical model for unsupervised and semi-supervised transliteration mining. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 469–477, Jeju Island, Korea, July. Association for Computational Linguistics.