A Tightly-coupled Unsupervised Clustering and Bilingual Alignment Model for Transliteration

Tingting Li[1], **Tiejun Zhao**[1], **Andrew Finch**[2], **Chunyue Zhang**[1]
[1]Harbin Institute of Technology, Harbin, China
[2]NICT, Japan
[1]{ttli, tjzhao, cyzhang}@mtlab.hit.edu.cn
[2]andrew.finch@nict.go.jp

Abstract

Machine Transliteration is an essential task for many NLP applications. However, names and loan words typically originate from various languages, obey different transliteration rules, and therefore may benefit from being modeled independently. Recently, transliteration models based on Bayesian learning have overcome issues with over-fitting allowing for many-to-many alignment in the training of transliteration models. We propose a novel coupled Dirichlet process mixture model (cDPMM) that simultaneously clusters and bilingually aligns transliteration data within a single unified model. The unified model decomposes into two classes of non-parametric Bayesian component models: a Dirichlet process mixture model for clustering, and a set of multinomial Dirichlet process models that perform bilingual alignment independently for each cluster. The experimental results show that our method considerably outperforms conventional alignment models.

1 Introduction

Machine transliteration methods can be categorized into phonetic-based models (Knight et al., 1998), spelling-based models (Brill et al., 2000), and hybrid models which utilize both phonetic and spelling information (Oh et al., 2005; Oh et al., 2006). Among them, statistical spelling-based models which directly align characters in the training corpus have become popular because they are language-independent, do not require phonetic knowledge, and are capable of achieving state-of-the-art performance (Zhang et al., 2012b). A major problem with real-word transliteration corpora is that they are usually not clean, may contain name pairs with various linguistic origins and this can hinder the performance of spelling-based models because names from different origins obey different pronunciation rules, for example:

"Kim Jong-il/金正恩" (Korea),
"Kana Gaski/金崎" (Japan),
"Haw King/霍金" (England),
"Jin yong/金庸' (China).

The same Chinese character "金" should be aligned to different romanized character sequences: "Kim", "Kana", "King", "Jin". To address this issue, many name classification methods have been proposed, such as the supervised language model-based approach of (Li et al., 2007), and the unsupervised approach of (Huang et al., 2005) that used a bottom-up clustering algorithm. (Li et al., 2007) proposed a supervised transliteration model which classifies names based on their origins and genders using a language model; it switches between transliteration models based on the input. (Hagiwara et al., 2011) tackled the issue by using an unsupervised method based on the EM algorithm to perform a soft classification.

Recently, non-parametric Bayesian models (Finch et al., 2010; Huang et al., 2011; Hagiwara et al., 2012) have attracted much attention in the transliteration field. In comparison to many of the previous alignment models (Li et al., 2004; Jiampojamarn et al., 2007; Berg-Kirkpatrick et al., 2011), the non-parametric Bayesian models allow unconstrained monotonic many-to-many alignment and are able to overcome the inherent over-fitting problem.

Until now most of the previous work (Li et al., 2007; Hagiwara et al., 2011) is either affected by the multi-origins factor, or has issues with over-fitting. (Hagiwara et al., 2012) took these two factors into consideration, but their approach still operates within an EM framework and model order selection by hand is necessary prior to training.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 393–398,
Sofia, Bulgaria, August 4-9 2013.

We propose a simple, elegant, fully-unsupervised solution based on a single generative model able to both cluster and align simultaneously. The coupled Dirichlet Process Mixture Model (cDPMM) integrates a Dirichlet process mixture model (DPMM) (Antoniak, 1974) and a Bayesian Bilingual Alignment Model (BBAM) (Finch et al., 2010). The two component models work synergistically to support one another: the clustering model sorts the data into classes so that self-consistent alignment models can be built using data of the same type, and at the same time the alignment probabilities from the alignment models drive the clustering process.

In summary, the key advantages of our model are as follows:

- it is based on a single, unified generative model;
- it is fully unsupervised;
- it is an infinite mixture model, and does not require model order selection – it is effectively capable of discovering an appropriate number of clusters from the data;
- it is able to handle data from multiple origins;
- it can perform many-to-many alignment without over-fitting.

2 Model Description

In this section we describe the methodology and realization of the proposed cDPMM in detail.

2.1 Terminology

In this paper, we concentrate on the alignment process for transliteration. The proposed cDPMM segments a bilingual corpus of transliteration pairs into bilingual character sequence-pairs. We will call these sequence-pairs Transliteration Units (TUs). We denote the source and target of a TU as $s_1^m = \langle s_1, ..., s_m \rangle$ and $t_1^n = \langle t_1, ..., t_n \rangle$ respectively, where s_i (t_i) is a single character in source (target) language. We use the same notation $(\mathbf{s}, \mathbf{t}) = (\langle s_1, ..., s_m \rangle, \langle t_1, ..., t_n \rangle)$ to denote a transliteration pair, which we can write as $x = (s_1^m, t_1^n)$ for simplicity. Finally, we express the training set itself as a set of sequence pairs: $D = \{x_i\}_{i=1}^I$. Our aim is to obtain a bilingual alignment $\langle (\mathbf{s}_1, \mathbf{t}_1), ..., (\mathbf{s}_l, \mathbf{t}_l) \rangle$ for each transliteration pair x_i, where each $(\mathbf{s}_j, \mathbf{t}_j)$ is a segment of the whole pair (a TU) and l is the number of segments used to segment x_i.

2.2 Methodology

Our cDPMM integrates two Dirichlet process models: the DPMM clustering model, and the BBAM alignment model which is a multinomial Dirichlet process.

A *Dirichlet process mixture model*, models the data as a mixture of distributions – one for each cluster. It is an infinite mixture model, and the number of components is not fixed prior to training. Equation 1 expresses the DPMM hierarchically.

$$\begin{aligned} G_c|\alpha_c, G_{0c} &\sim DP(\alpha_c, G_{0c}) \\ \theta_k|G_c &\sim G_c \\ x_i|\theta_k &\sim f(x_i|\theta_k) \end{aligned} \quad (1)$$

where G_{0c} is the base measure and $\alpha_c > 0$ is the concentration parameter for the distribution G_c. x_i is a name pair in training data, and θ_k represents the parameters of a candidate cluster k for x_i. Specifically θ_k contains the probabilities of all the TUs in cluster k. $f(x_i|\theta_k)$ (defined in Equation 7) is the probability that mixture component k parameterized by θ_k will generate x_i.

The alignment component of our cDPMM is a *multinomial Dirichlet process* and is defined as follows:

$$\begin{aligned} G_a|\alpha_a, G_{0a} &\sim DP(\alpha_a, G_{0a}) \\ (\mathbf{s}_j, \mathbf{t}_j)|G_a &\sim G_a \end{aligned} \quad (2)$$

The subscripts 'c' and 'a' in Equations 1 and 2 indicate whether the terms belong to the clustering or alignment model respectively.

The generative story for the cDPMM is simple: first generate an infinite number of clusters, choose one, then generate a transliteration pair using the parameters that describe the cluster. The basic sampling unit of the cDPMM for the clustering process is a transliteration pair, but the basic sampling unit for BBAM is a TU. In order to integrate the two processes in a single model we treat a transliteration pair as a sequence of TUs generated by a BBAM model. The BBAM generates a sequence (a transliteration pair) based on the joint source-channel model (Li et al., 2004). We use a blocked version of a Gibbs sampler to train each BBAM (see (Mochihashi et al., 2009) for details of this process).

2.3 The Alignment Model

This model is a multinomial DP model. Under the Chinese restaurant process (CRP) (Aldous, 1985)

interpretation, each unique TU corresponds to a dish served at a table, and the number of customers in each table represents the count of a particular TU in the model.

The probability of generating the j^{th} TU $(\mathbf{s}_j, \mathbf{t}_j)$ is,

$$P\Big((\mathbf{s}_j, \mathbf{t}_j)|(\mathbf{s}_{-j}, \mathbf{t}_{-j})\Big) = \frac{N\Big((\mathbf{s}_j, \mathbf{t}_j)\Big) + \alpha_a G_{0a}\Big((\mathbf{s}_j, \mathbf{t}_j)\Big)}{N + \alpha_a} \tag{3}$$

where N is the total number of TUs generated so far, and $N\Big((\mathbf{s}_j, \mathbf{t}_j)\Big)$ is the count of $(\mathbf{s}_j, \mathbf{t}_j)$. $(\mathbf{s}_{-j}, \mathbf{t}_{-j})$ are all the TUs generated so far except $(\mathbf{s}_j, \mathbf{t}_j)$. The base measure G_{0a} is a joint spelling model:

$$\begin{aligned} G_{0a}\Big((\mathbf{s}, \mathbf{t})\Big) &= P(|\mathbf{s}|)P(\mathbf{s}||\mathbf{s}|)P(|\mathbf{t}|)P(\mathbf{t}||\mathbf{t}|) \\ &= \frac{\lambda_s^{|\mathbf{s}|}}{|\mathbf{s}|!} e^{-\lambda_s} \mathbf{v}_s^{-|\mathbf{s}|} \times \frac{\lambda_t^{|\mathbf{t}|}}{|\mathbf{t}|!} e^{-\lambda_t} \mathbf{v}_t^{-|\mathbf{t}|} \end{aligned} \tag{4}$$

where $|\mathbf{s}|$ ($|\mathbf{t}|$) is the length of the source (target) sequence, $\mathbf{v}_s$ ($\mathbf{v}_t$) is the vocabulary (alphabet) size of the source (target) language, and λ_s (λ_t) is the expected length of source (target) side.

2.4 The Clustering Model

This model is a DPMM. Under the CRP interpretation, a transliteration pair corresponds to a customer, the dish served on each table corresponds to an origin of names.

We use $z = (z_1, ..., z_I), z_i \in \{1, ..., K\}$ to indicate the cluster of each transliteration pair x_i in the training set and $\theta = (\theta_1, ..., \theta_K)$ to represent the parameters of the component associated with each cluster.

In our model, each mixture component is a multinomial DP model, and since θ_k contains the probabilities of all the TUs in cluster k, the number of parameters in each θ_k is uncertain and changes with the transliteration pairs that belong to the cluster. For a new cluster (the $K + 1^{\text{th}}$ cluster), we use Equation 4 to calculate the probability of each TU. The cluster membership probability of a transliteration pair x_i is calculated as follows,

$$P(z_i = k|D, \theta, z_{-i}) \propto \frac{n_k}{n - 1 + \alpha_c} P(x_i|z, \theta_k) \tag{5}$$

$$P(z_i = K + 1|D, \theta, z_{-i}) \propto \frac{\alpha_c}{n - 1 + \alpha_c} P(x_i|z, \theta_{K+1}) \tag{6}$$

where n_k is the number of transliteration pairs in the existing cluster $k \in \{1, ..., K\}$ (cluster $K + 1$ is a newly created cluster), z_i is the cluster indicator for x_i, and z_{-i} is the sequence of observed clusters up to x_i. As mentioned earlier, basic sampling units are inconsistent for the clustering and alignment model, therefore to couple the models the BBAM generates transliteration pairs as a sequence of TUs, these pairs are then used directly in the DPMM.

Let $\gamma = \langle(\mathbf{s}_1, \mathbf{t}_1), ..., (\mathbf{s}_l, \mathbf{t}_l)\rangle$ be a derivation of a transliteration pair x_i. To make the model integration process explicit, we use function f to calculate the probability $P(x_i|z, \theta_k)$, where f is defined as follows,

$$f(x_i|\theta_k) = \begin{cases} \sum_{\gamma \in R} \prod_{(\mathbf{s},\mathbf{t}) \in \gamma} P(\mathbf{s}, \mathbf{t}|\theta_k) & k \in \{1, ..., K\} \\ \sum_{\gamma \in R} \prod_{(\mathbf{s},\mathbf{t}) \in \gamma} G_{0c}(\mathbf{s}, \mathbf{t}) & k = K + 1 \end{cases} \tag{7}$$

where R denotes the set of all derivations of x_i, G_{0c} is the same as Equation 4.

The cluster membership z_i is sampled together with the derivation γ in a single step according to $P(z_i = k|D, \theta, z_{-i})$ and $f(x_i|\theta_k)$. Following the method of (Mochihashi et al., 2009), first $f(x_i|\theta_k)$ is calculated by forward filtering, and then a sample γ is taken by backward sampling.

3 Experiments

3.1 Corpora

To empirically validate our approach, we investigate the effectiveness of our model by conducting English-Chinese name transliteration generation on three corpora containing name pairs of varying degrees of mixed origin. The first two corpora were drawn from the "Names of The World's Peoples" dictionary published by Xin Hua Publishing House. The first corpus was constructed with names only originating from English language (EO), and the second with names originating from English, Chinese, Japanese evenly (ECJ-O). The third corpus was created by extracting name pairs from LDC (Linguistic Data Consortium) Named Entity List, which contains names from all over the world (Multi-O). We divided the datasets into training, development and test sets for each corpus with a ratio of 10:1:1. The details of the division are displayed in Table 2.

cDPMM Alignment	BBAM Alignment
mun|蒙 din|丁 ger|格(0, English) ding|丁 guo|果(2, Chinese) tei|丁 be|部(3, Japanese)	mun|蒙 din|丁 ger|格 din|丁 g|_ guo|果 t|_ _|丁 e|_ ibe|部
fan|范 chun|纯 yi|一(2, Chinese) hong|洪 il|一 sik|植(5, Korea) sei|静 ichi|一 ro|郎(4, Japanese)	fan|范 chun|纯 y|_ i|一 hong|洪 i|一 l|_ si|植 k|_ seii|静 ch|_ i|一 ro|郎
dom|东 b|布 ro|罗 w|夫 s|斯 ki|基(0, Russian) he|何 dong|东 chang|昌(2, Chinese) b|布 ran|兰 don|东(0, English)	do|东 mb|布 ro|罗 w|夫 s|斯 ki|基 he|何 don|东 gchang|昌 b|布 ran|兰 don|东

Table 1: Typical alignments from the BBAM and cDPMM.

3.2 Baselines

We compare our alignment model with GIZA++ (Och et al., 2003) and the Bayesian bilingual alignment model (BBAM). We employ two decoding models: a phrase-based machine translation decoder (specifically Moses (Koehn et al., 2007)), and the DirecTL decoder (Jiampojamarn et al., 2009). They are based on different decoding strategies and optimization targets, and therefore make the comparison more comprehensive. For the Moses decoder, we applied the grow-diag-final-and heuristic algorithm to extract the phrase table, and tuned the parameters using the BLEU metric.

Corpora	Corpus Scale		
	Training	Development	Testing
EO	32,681	3,267	3,267
ECJ-O	32,500	3,250	3,250
Multi-O	33,291	3,328	3,328

Table 2: Statistics of the experimental corpora.

To evaluate the experimental results, we utilized 3 metrics from the Named Entities Workshop (NEWS) (Zhang et al., 2012a): word accuracy in top-1 (ACC), fuzziness in top-1 (Mean F-score) and mean reciprocal rank (MRR).

3.3 Parameter Setting

In our model, there are several important parameters: 1) max_s, the maximum length of the source sequences of the alignment tokens; 2) max_t, the maximum length of the target sequences of the alignment tokens; and 3) nc, the initial number of classes for the training data. We set $max\_s = 6$, $max\_t = 1$ and $nc = 5$ empirically based on a small pilot experiment. The Moses decoder was used with default settings except for the distortion-limit which was set to 0 to ensure monotonic decoding. For the DirecTL decoder the following settings were used: $cs = 4, ng = 9$ and $nBest = 5$. cs denotes the size of context window for features, ng indicates the size of n-gram features and $nBest$ is the size of transliteration candidate list for updating the model in each iteration. The concentration parameter α_c, α_a of the clustering model and the BBAM was learned by sampling its value. Following (Blunsom et al., 2009) we used a vague gamma prior $\Gamma(10^{-4}, 10^4)$, and sampled new values from a log-normal distribution whose mean was the value of the parameter, and variance was 0.3. We used the Metropolis-Hastings algorithm to determine whether this new sample would be accepted. The parameters λ_s and λ_t in Equation 4 were set to $\lambda_s = 4$ and $\lambda_t = 1$.

	Model	EO	ECJ-O	Multi-O
#(Clusters)	cDPMM	5.8	9.5	14.3
#(Targets)	GIZA++	14.43	5.35	6.62
	BBAM	6.06	2.45	2.91
	cDPMM	9.32	3.45	4.28

Table 3: Alignment statistics.

3.4 Experimental Results

Table 3 shows some details of the alignment results. The #(Clusters) represents the average number of clusters from the cDPMM. It is averaged over the final 50 iterations, and the classes which contain less than 10 name pairs are excluded. The #(Targets) represents the average number of English character sequences that are aligned to each Chinese sequence. From the results we can see that in terms of the number of alignment targets: GIZA++ > cDPMM > BBAM. GIZA++ has considerably more targets than the other approaches, and this is likely to be a symptom of it overfitting the data. cDPMM can alleviate the overfitting through its BBAM component, and at the same time effectively model the diversity in Chinese character sequences caused by multi-origin. Table 1 shows some typical TUs from the alignments produced by BBAM and cDPMM on corpus Multi-O. The information in brackets in Table 1, represents the ID of the class and origin of

Corpora	Model	Evaluation		
		ACC	M-Fscore	MRR
EO	GIZA	0.7241	0.8881	0.8061
	BBAM	0.7286	0.8920	0.8043
	cDPMM	**0.7398**	**0.8983**	**0.8126**
ECJ-O	GIZA	0.5471	0.7278	0.6268
	BBAM	0.5522	0.7370	0.6344
	cDPMM	**0.5643**	**0.7420**	**0.6446**
Multi-O	GIZA	0.4993	0.7587	0.5986
	BBAM	0.5163	0.7769	0.6123
	cDPMM	**0.5237**	**0.7796**	**0.6188**

Table 4: Comparison of different methods using the Moses phrase-based decoder.

Corpora	Model	Evaluation		
		ACC	M-Fscore	MRR
EO	GIZA	0.6950	0.8812	0.7632
	BBAM	0.7152	0.8899	0.7839
	cDPMM	**0.7231**	**0.8933**	**0.7941**
ECJ-O	GIZA	0.3325	0.6208	0.4064
	BBAM	0.3427	0.6259	0.4192
	cDPMM	**0.3521**	**0.6302**	**0.4316**
Multi-O	GIZA	0.3815	0.7053	0.4592
	BBAM	0.3934	0.7146	0.4799
	cDPMM	**0.3970**	**0.7179**	**0.4833**

Table 5: Comparison of different methods using the DirecTL decoder.

the name pair; the symbol '_' indicates a "NULL" alignment. We can see the Chinese characters "丁(ding) 一(yi) 东(dong)" have different alignments in different origins, and that the cDPMM has provided the correct alignments for them.

We used the sampled alignment from running the BBAM and cDPMM models for 100 iterations, and combined the alignment tables of each class together. The experiments are therefore investigating whether the alignment has been meaningfully improved by the clustering process. We would expect further gains from exploiting the class information in the decoding process (as in (Li et al., 2007)), but this remains future research. The top-10 transliteration candidates were used for testing. The detailed experimental results are shown in Tables 4 and 5.

Our proposed model obtained the highest performance on all three datasets for all evaluation metrics by a considerable margin. Surprisingly, for dataset EO although there is no multi-origin factor, we still observed a respectable improvement in every metric. This shows that although names may have monolingual origin, there are hidden factors which can allow our model to succeed, possibly related to gender or convention. Other models based on supervised classification or clustering with fixed classes may fail to capture these characteristics.

To guarantee the reliability of the comparative results, we performed significance testing based on paired bootstrap resampling (Efron et al., 1993). We found all differences to be significant ($p < 0.05$).

4 Conclusion

In this paper we propose an elegant unsupervised technique for monotonic sequence alignment based on a single generative model. The key benefits of our model are that it can handle data from multiple origins, and model using many-to-many alignment without over-fitting. The model operates by clustering the data into classes while simultaneously aligning it, and is able to discover an appropriate number of classes from the data. Our results show that our alignment model can improve the performance of a transliteration generation system relative to two other state-of-the-art aligners. Furthermore, the system produced gains even on data of monolingual origin, where no obvious clusters in the data were expected.

Acknowledgments

We thank the anonymous reviewers for their valuable comments and helpful suggestions.We also thank Chonghui Zhu, Mo Yu, and Wenwen Zhang for insightful discussions. This work was supported by National Natural Science Foundation of China (61173073), and the Key Project of the National High Technology Research and Development Program of China (2011AA01A207).

References

D.J. Aldous. 1985. Exchangeability and Related Topics. *École d'Été St Flour 1983*. Springer, 1985, 1117:1–198.

C.E. Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*. 2:1152, 174.

Taylor Berg-Kirkpatrick and Dan Klein. 2011. Simple effective decipherment via combinatorial optimization. In *Proc. of EMNLP*, pages 313–321.

P. Blunsom, T. Cohn, C. Dyer, and Osborne, M. 2009. A Gibbs sampler for phrasal synchronous grammar induction. In *Proc. of ACL*, pages 782–790.

Eric Brill and Robert C. Moore. 2000. An Improved Error Model for Noisy Channel Spelling Correction. In *Proc. of ACL*, pages 286–293.

B. Efron and R. J. Tibshirani 1993. An Introduction to the Bootstrap. Chapman & Hall, New York, NY.

Andrew Finch and Eiichiro Sumita. 2010. A Bayesian Model of Bilingual Segmentation for Transliteration. In *Proc. of the 7th International Workshop on Spoken Language Translation*, pages 259–266.

Masato Hagiwara and Satoshi Sekine. 2011. Latent Class Transliteration based on Source Language Origin. In *Proc. of ACL (Short Papers)*, pages 53-57.

Masato Hagiwara and Satoshi Sekine. 2012. Latent semantic transliteration using dirichlet mixture. In *Proc. of the 4th Named Entity Workshop*, pages 30–37.

Fei Huang, Stephan Vogel, and Alex Waibel. 2005. Clustering and Classifying Person Names by Origin. In *Proc. of AAAI*, pages 1056–1061.

Yun Huang, Min Zhang and Chew Lim Tan. 2011. Nonparametric Bayesian Machine Transliteration with Synchronous Adaptor Grammars. In *Proc. of ACL*, pages 534–539.

Sittichai Jiampojamarn, Grzegorz Kondrak and Tarek Sherif. 2007. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. In *Proc. of NAACL*, pages 372–379.

Sittichai Jiampojamarn, Aditya Bhargava, Qing Dou, Kenneth Dwyer and Grzegorz Kondrak. 2009. DirecTL: a Language Independent Approach to Transliteration. In *Proc. of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 1056–1061.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Journal of Computational Linguistics*, pages 28–31.

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

Haizou Li, Min Zhang, and Jian Su 2004. A joint source-channel model for machine transliteration. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 159.

Haizhou Li, Khe Chai Sim, Jin-Shea Kuo, and Minghui Dong. 2007. Semantic Transliteration of Personal Names. In *Proc. of ACL*, pages 120–127.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling. In *Proc. of ACL/IJCNLP*, pages 100–108.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Journal of Comput. Linguist.*, 29(1):19-51.

Jong-Hoon Oh, and Key-Sun Choi. 2005. Machine Learning Based English-to-Korean Transliteration Using Grapheme and Phoneme Information. *Journal of IEICE Transactions*, 88-D(7):1737-1748.

Jong-Hoon Oh, Key-Sun Choi, and Hitoshi Isahara. 2006. A machine transliteration model based on correspondence between graphemes and phonemes. *Journal of ACM Trans. Asian Lang. Inf. Process.*, 5(3):185-208.

Min Zhang, Haizhou Li, Ming Liu and A Kumaran. 2012a. Whitepaper of NEWS 2012 shared task on machine transliteration. In *Proc. of the 4th Named Entity Workshop (NEWS 2012)*, pages 1–9.

Min Zhang, Haizhou Li, A Kumaran and Ming Liu. 2012b. Report of NEWS 2012 Machine Transliteration Shared Task. In *Proc. of the 4th Named Entity Workshop (NEWS 2012)*, pages 10–20.