Comparative Evaluation of Spanish Segmentation Strategies for Spanish-Chinese Transliteration

Rafael E. Banchs
Human Language Technology Department, Institute for Infocomm Research
1 Fusionopolis Way, #21-01 Connexis South, Singapore 138632
rembanchs@i2r.a-star.edu.sg

Abstract

This work presents a comparative evaluation among three different Spanish segmentation strategies for Spanish-Chinese transliteration. The transliteration task is implemented by means of Statistical Machine Translation, using Chinese characters and Spanish sub-word segments as the textual units to be translated. Three different Spanish segmentation strategies are evaluated: character-based, syllabic-based and a proposed sub-syllabic segmentation scheme. Experimental results show that syllabic-based segmentation is the most effective strategy for Spanish-to-Chinese transliteration, while the proposed sub-syllabic segmentation is the most effective scheme in the case of Chinese-to-Spanish transliteration.

1 Introduction

Transliteration can be defined as the process of transcribing a word from one language to another by using the characters of the latter's alphabet. This actually constitutes a "phonetic translation of names across languages" (Zhang *et al.*, 2011). Transliteration is typically used to construct appropriate translations for words that either do not have specific equivalents or are inexistent in the target language, such as, for instance, names of people, institutions or geographical locations.

Although they are conceptually similar tasks, technically speaking, translation and transliteration exhibit some important differences. For instance, while translation mainly operates at the word level, transliteration does it at the sub-word level. Perhaps, the most important difference is the fact that in the transliteration task, reordering of units is not required. As in the case of translation, transliteration results are not necessarily unique, i.e. one word might have different valid transliterations.

The transliteration task can be approached from either a rule-based or a statistical perspective, but in any case, the problem can be theoretically grounded on Finite-state Automata Theory (Knight, 2009). Several different approaches to transliteration have been proposed in the literature (Arbabi *et al.*, 1994; Divay and Vitale, 1997; Knight and Graehl, 1998; Al-Onaizan and Knight, 2002; Li *et al.*, 2004; Tao *et al.*, 2006; Yoon *et al.*, 2007; Jansche and Sproat, 2009) covering specific transliteration tasks between English and a large variety of languages such as Japanese (Knight and Graehl, 1998), French (Divay and Vitale, 1997), Arabic (Arbabi *et al.*, 1994; Al-Onaizan and Knight, 2002), Chinese (Ren et al., 2009; Kwong, 2009), Hindi (Chinnakotla and Damani, 2009; Das *et al.*, 2009; Haque *et al.*, 2009), Tamil (Vijayanand, 2009) and Korean (Hong *et al.*, 2009), among others.

Nevertheless, despite of the large body of research on automatic transliteration, and as far as we are concerned, there have not been research efforts reported on this area for the specific case of Spanish and Chinese. According to this, the main objective of this work is twofold: first, to create an experimental dataset for transliteration between Chinese and Spanish; and, second, to report some research results on transliteration tasks between these two languages.

The remaining of the paper is structured as follows. First, in section 2, the main technical issue evaluated in this work, which is the segmentation of Spanish words into sub-word units, is introduced and motivated. Then, in section 3, the selected SMT-based approach for Chinese-Spanish transliteration, is described. In section 4, the creation of an experimental dataset for Chinese-Spanish transliteration is described in detail. In section 5, experimental results are presented and discussed. Finally, in section 6, main conclusions and future research ideas are provided.

Proceedings of the 2011 Named Entities Workshop, IJCNLP 2011, pages 41–48,
Chiang Mai, Thailand, November 12, 2011.

2 Spanish Word Segmentation

The concept of isochronism in language was first introduced by Pike (1945). Three types of rhythmic patterns can be distinguished: stress-timed, syllable-timed and mora-timed. Although this theory has not been fully accepted, there is some accepted empirical evidence that both Spanish (Pamies Bertran, 1999) and Chinese (Lin and Wang, 2007) belong to the syllable-timed rhythmic group.

In the case of Chinese, syllabic segmentation is naturally induced by the basic association between the characters and their corresponding sounds. On the contrary, in the case of Spanish, as well as many other western languages, syllabic segmentation is a phonetic property that does not exhibit a direct or explicit association with orthographic properties of the language.

According to this, syllabic segmentation or syllabification constitutes a problem of interest in some natural language processing applications. This problem can be addressed by means of either rule-based or data-driven approaches (Adsett *et al.*, 2009). Syllabification algorithms based on finite-state transducers have been proposed for languages such as English and German (Kiraz and Mobius, 1998). For the effects of the present work, we implemented our own rule-based syllabic segmentation algorithm for Spanish by following the work of Cuayahuitl (2004).

Three different strategies for Spanish word segmentation are studied in this work with the objective of determining the most appropriate segmentation scheme for Chinese-Spanish transliteration. These three strategies are: character segmentation (the simple division of a word in characters), syllabic segmentation (the division of a word according to Spanish syllabic phonetic units) and an intermediate segmentation to be referred to as sub-syllabic segmentation. The rest of this section is devoted to motivate and explain this latter segmentation scheme.

The main motivation for the proposed sub-syllabic segmentation of Spanish words is the observed fact that, although they agree in most of the cases, syllabifications can often differ between Spanish and Chinese transliterated names. Consider, for instance, the examples presented in Figure 1. The first two examples illustrate cases in which the Chinese name contains less syllables than the corresponding Spanish name. On the other hand, the last three examples illustrate cases in which the Chinese name contains more syllables than the corresponding Spanish name.

Chinese – Pinyin	Spanish
马 其 顿 – mǎ qí dùn	ma ce do nia
亚 略 巴 古 – yà è ba gǔ	a re ó pa go
亚 历 山 大 – yà lì shān dà	a le jan dro
塞 缪 尔 – sāi móu ěr	sa muel
亚 伯 拉 罕 – yà bó lā hǎn	a bra ham
埃 利 亚 斯 – āi lì yà sī	e lí as

Figure 1. Some examples of Chinese-Spanish name transliterations

A detailed analysis on the syllabic length ratios between Chinese and Spanish names on our experimental dataset (more details on the dataset are provided in section 4) reveals that the most common situation is that both Chinese and Spanish names have the same number of syllables. This occurs in about 75% of the cases. From the remaining 25% of cases, about 15% (and 10%) correspond to cases in which the Chinese versions of the names contain more (and less) syllables than their corresponding Spanish versions.

Further analysis show that some clear patterns for sub-syllabic segmentation can be observed in those cases of Chinese transliterations containing more syllables than their corresponding Spanish versions, which is not the case for the opposite situation. Some of these patterns include the segmentation of Spanish diphthongs such as *ue* into *u-e*, which will generate the more appropriate segmentation *sa-mu-el* for the fourth example in Figure 1; the separation of some multiple consonant constructions such as *br* into *b-r*, which will provide the more appropriate segmentation *a-b-ra-ham*; and the separation of some ending consonants such as *as* into *a-s*, which will generate *e-li-a-s*. This sub-syllabic segmentation strategy is expected to improve the performance of the transliteration task as it both reduces the vocabulary size of Spanish syllabic units and improves syllable correspondences between Chinese and Spanish. The complete set and sequence of rules implemented for sub-syllabic segmentation is presented in Figure 2.

Notice that the proposed sub-syllabic segmentation strategy is only addressing those cases in which the Chinese versions of the names contain more syllables than their corresponding Spanish

versions. Addressing the opposite case, would require instead the definition of rules for merging consecutive Spanish syllables. We have not considered this case because of two reasons: first, according to our exploratory analysis of the data, it does not seem to be clear patterns for syllabic merging; and, second, a merging strategy would lead to an increment of the vocabulary of Spanish Syllabic units, which is not desirable in terms of the resulting transliteration model sparseness.

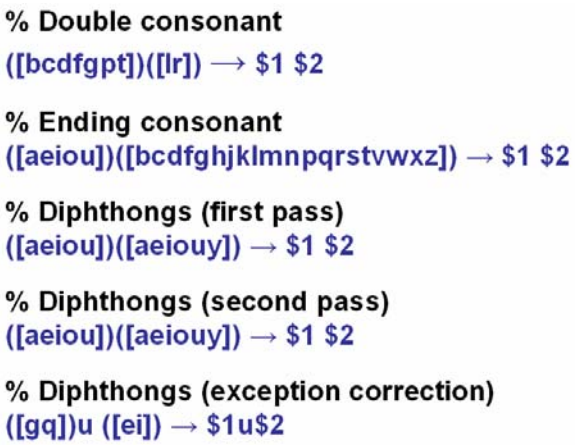

Figure 2. Rules and their sequence of application for sub-syllabic segmentation

Notice that, those cases in which the Chinese versions of the names contain less syllables than their corresponding Spanish versions are basically unaddressed by our proposed segmentation strategy. This, however, should not constitute a problem in the case of Spanish-to-Chinese transliteration as the transliteration model just should be required to learn how to throw away some Spanish syllables. On the other hand, this certainly posses a problem for the case of Chinese-to-Spanish transliteration as the transliteration model must be able to generate Spanish syllables from no Chinese correspondents. However, we still expect an overall gain as the former case is more common that the latter one.

3 Transliteration Approach

For implementing the transliteration system, we have used the Phrase-Based Statistical Machine Translation approach, which has been proven to be a good strategy for transliteration (Noeman, 2009; Jia *et al.*, 2009). Within this approach, transliteration is performed as a machine translation task over substring units of both the source and the target languages. More specifically, we use the MOSES toolkit (Koehn *et al.*, 2007).

Although several parameters can be varied in order to study their effect over the overall transliteration performance, we will focus our study in three specific parameters, which we consider could have the largest incidence, as well as make an important difference, on quality for both transliterations directions under consideration: Spanish-to-Chinese and Chinese-to-Spanish.

The first parameter of interest is substring segmentation. Although we only consider Chinese characters as substring units for Chinese; in the case of Spanish, we consider three different types of substring units according to the three segmentation schemes described in the previous section. More specifically, characters, syllables and the proposed sub-syllabic units are considered for Spanish.

The other two parameters to be considered for evaluation purposes are the order of the target language model and the alignment strategy used for phrase extraction. In the case of the target language model, four different orders are compared, namely: *1*-gram, *2*-gram, *3*-gram and *4*-gram; and in the case of the alignment strategy, three different methods are compared, namely: source-to-target, target-to-source and grow-diag-final-and (Koehn *et al.*, 2007).

According to this, our experimental work involves the construction of 72 different transliteration systems, by considering 2 transliteration directions, 3 Spanish segmentation schemes, 4 target language model orders, and 3 alignment strategies. In each of these transliteration systems, the standard set of phrase-based features, which include the forward and backward relative frequencies and lexical models, as well as the target language and phrase-length penalty models, are used.

As evaluation metric for assessing transliteration quality we use the BLEU score (Papineni *et al.*, 2001). In the case of Spanish-to-Chinese transliterations, BLEU is computed at the Chinese character level. Similarly, and in order to make results among all three different Spanish segmentation schemes comparable, in the case of Chinese-to-Spanish transliterations, BLEU is computed at the character level too.

Finally, each of the implemented systems is tuned by means of the minimum error rate training procedure (Och, 2003), in which the BLEU score is minimized over a development dataset. Final system scores are computed over a test dataset, which is transliterated by using the tuned parameters. More details on the datasets are provided in the following section.

4 Dataset Construction

As no named entity dataset is available for transliteration purposes between Spanish and Chinese, the first objective of this work was the creation of such a dataset. Despite the fact that Chinese and Spanish are the most spoken native languages in the word, the amount of bilingual resources for this specific language pair happens to be very scarce (Costa-jussa *et al.* 2011).

According to this, we used one of the few bilingual resources that are available, the Holy Bible (Table 1 presents the basic statistics for this dataset), for constructing an experimental dataset for transliteration research purposes.

Language	Sentences	Words	Vocab.
Chinese	29,887	781,113	28,178
Spanish	29,887	848,776	13,126

Table 1. Basic statistics of the Bible dataset

In his section we present a description of the procedure followed for creating the dataset, as well as the basic statistics and characteristics of the constructed dataset.

The construction of the experimental dataset for transliteration can be summarized according to the following steps:

- A list of named entities was extracted from the Spanish side of the dataset. This extraction was conducted by using a standard labeling approach based on Conditional Random Fields (Lafferty *et al.*, 2001). From this step a list of 1,608 Spanish names were collected.

- A reduced list of named entities was generated by manually filtering the original list. In this process some errors derived from the first automatic step were removed, as well as any valid name entity not belonging to the two basic categories of persons and places. In this second step, the list was reduced to 948 names.

- The corresponding Chinese versions of the names were extracted from the Chinese side of the dataset. This was done automatically by aligning both corpus at the word level (Och and Ney, 2000), and using the alignment links to identify the corresponding transliteration candidates for each Spanish name in the list.

- The automatically extracted list of corresponding Chinese names was manually depurated. Because of the noisy nature of the alignment process, in several cases either more than one Chinese word was assigned to the same Spanish names or an erroneous Chinese word was selected. After this second filtering processing, the final bilingual list of 841 names was obtained.

For the preparation of the experimental dataset each side of the resulting corpus was segmented as follows: Chinese data was segmented at the character level, and Spanish data was segmented by following the three segmentation schemes described in section 2: character-based, syllable-based and sub-syllabic.

Two additional normalization processes were applied to the Spanish dataset: lowercasing and stress mark elimination. The total number of substring units and their vocabulary for each of the constructed versions of the dataset are presented in Table 2.

Dataset	Names	Substrings	Vocab.
Chinese	841	2,190	314
Spa (char)	841	4,766	24
Spa (sub)	841	3,005	108
Spa (syl)	841	2,165	491

Table 2. Names, substring units and vocabulary of substring units for each constructed dataset

As seen from the table, the tree Spanish word segmentations to be studied exhibit significantly different properties in terms of the total amount of running substrings and the vocabulary size of substring units. Indeed, the proposed sub-syllabic segmentation strategy represents an intermediate compromise in both, substrings and vocabulary, between the character-based segmentation and the syllabic-based segmentation.

In order to be able to use the generated dataset under the statistical machine translation framework described in section 3, the resulting bilingual dataset of 841 names was finally split into three subsets: train (with 691 names), development (with 50 names) and test (with 100 names).

Although a random sample strategy was used for splitting the original corpus into the three experimental subsets, special attention was paid to not include in the development and test subsets any name that would have produced out-of-vocabulary substrings.

5 Experimental Results

In this section we present and discuss the experimental results corresponding to all 72 implemented transliteration systems. All experiments were conducted over the experimental datasets described in section 4 by following the procedure described in section 3. Although we will focus our analysis on aggregated scores computed over different subsets of experiments, Tables 3a through 3f present individual system scores for all of the 72 implemented transliteration systems.

As seen from the tables, although individual results by themselves could exhibit some degree of noise due to the random variability derived from both, dataset selection and tuning processes, some clear and interesting trends can be observed form the results. For instance, notice how best scores tend to be always associated to language model of orders 3 and 4.

Similarly, it can be derived from the tables that the grow-diag-final-and alignment strategy tends to be the best alignment strategy only in those cases when the Spanish syllabic segmentation is used. Alternatively, it can be observed that in the other two cases, i.e. when Spanish character and sub-syllabic segmentations are used, the target-to-source alignment strategy is more beneficial for the Spanish-to-Chinese transliteration direction while the source-to-target alignment strategy happens to be more beneficial for the Chinese-to-Spanish direction.

In order to have a better grasp of the general trends in transliteration quality along the dimensions of each of the experimental parameters under consideration, let us now look at the aggregated results along each individual parameter variation. In this sense, Figures 3a, 3b and 3c summarize transliteration quality variations with respect to *n*-gram order, alignment strategy and Spanish segmentation, respectively.

Let us consider first Figure 3a. This figure shows the relative variations of transliteration quality with respect to *n*-gram order. These values have been computed by aggregating all system scores along the alignment strategy and Spanish segmentation dimensions for each of the two transliteration directions under consideration. Additionally, the resulting scores have been normalized with respect to the unigram case. As seen from the figure, there is a more critical incidence of the *n*-gram order on the case of Spanish-to-Chinese transliteration than in the opposite transliteration direction.

	src-2-trg	**trg-2-src**	**g-d-f-a**
1-gram	15.36	16.09	14.35
2-gram	18.98	21.87	19.43
3-gram	15.33	23.35	18.83
4-gram	18.19	24.05	19.85

Table 3a. BLEU scores for Spanish-to-Chinese systems with Spanish character segmentation

	src-2-trg	**trg-2-src**	**g-d-f-a**
1-gram	20.20	16.72	15.96
2-gram	15.58	22.85	15.37
3-gram	20.49	21.93	19.30
4-gram	21.80	21.72	19.17

Table 3b. BLEU scores for Spanish-to-Chinese systems with Spanish sub-syllabic segmentation

	src-2-trg	**trg-2-src**	**g-d-f-a**
1-gram	23.42	23.02	23.79
2-gram	25.27	24.28	31.98
3-gram	31.26	22.14	35.98
4-gram	30.83	24.41	35.48

Table 3c. BLEU scores for Spanish-to-Chinese systems with Spanish syllabic segmentation

	src-2-trg	**trg-2-src**	**g-d-f-a**
1-gram	38.38	33.96	35.58
2-gram	37.94	35.34	35.99
3-gram	35.41	39.34	37.21
4-gram	39.11	39.52	38.78

Table 3d. BLEU scores for Chinese-to-Spanish systems with Spanish character segmentation

	src-2-trg	**trg-2-src**	**g-d-f-a**
1-gram	40.17	36.53	39.94
2-gram	42.21	42.15	38.78
3-gram	39.67	43.03	40.89
4-gram	40.70	36.45	39.88

Table 3e. BLEU scores for Chinese-to-Spanish systems with Spanish sub-syllabic segmentation

	src-2-trg	**trg-2-src**	**g-d-f-a**
1-gram	37.50	30.74	37.77
2-gram	38.86	36.89	41.38
3-gram	38.66	37.20	40.83
4-gram	39.26	37.20	40.38

Table 3f. BLEU scores for Chinese-to-Spanish systems with Spanish syllabic segmentation

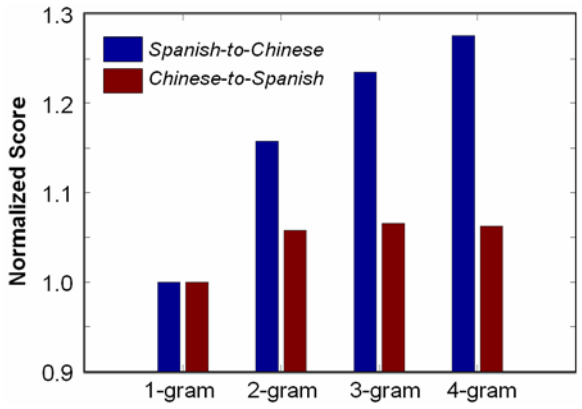

Figure 3a. Transliteration quality variations in terms of *n*-gram order

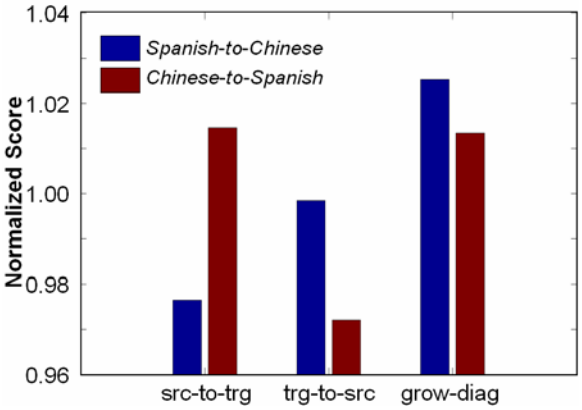

Figure 3b. Transliteration quality variations in terms of alignment strategy

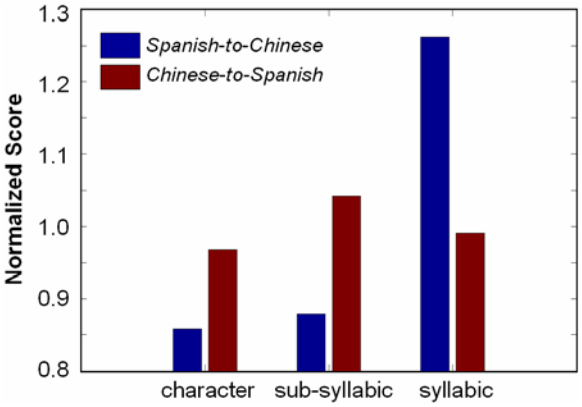

Figure 3c. Transliteration quality variations in terms of Spanish segmentation method

It is evident, from Figure 3a, that the transliteration tasks does not benefits from *n*-gram orders larger than 2 in the Chinese-to-Spanish direction, while it certainly does in the Spanish-to-Chinese case. This result can be explained by the larger character vocabulary size of Chinese when compared to Spanish segmentations.

In the case of Figure 3b, aggregation has been conducted along the *n*-gram orders and Spanish segmentations. In this case, the resulting scores have been normalized with respect to the average score value for each transliteration direction. While grow-diag-final-and is the best alignment strategy for the Spanish-to-Chinese case, source-to-target alignments also happen to be a good strategy in the Chinese-to-Spanish case. Notice, however, that relative variation of scores in Figure 3b is actually very low (about 2%), which suggests that the alignment strategy has a low incidence on transliteration quality for the tasks under consideration.

Finally, let us consider Figure 3c, where the relative variations of transliteration quality with respect to the selected Spanish segmentation method are depicted. In this cases system scores have been aggregated along both the *n*-gram order and the alignment strategy dimensions, and normalized with respect to average scores at each transliteration direction. Notice from the figure how syllabic segmentation is clearly the best option in the Spanish-to-Chinese transliteration direction, while the proposed sub-syllabic segmentation constitutes the best alternative in the Chinese-to-Spanish direction.

This latter interesting result can be explained in terms of the mapping functions required to map the corresponding substring units from one language into the other, as the larger the source vocabulary the better the mapping function is. So, in the case of the Spanish-to-Chinese task, the syllabic segmentation must provide a better mapping as it allows for a vocabulary reduction mapping, as can be verified from the vocabulary column in Table 2. On the other hand, in the Chinese-to-Spanish task the proposed method for sub-syllabic segmentation is the one providing a vocabulary reduction (as can be verified from the vocabulary column in Table 2) that allows for a better mapping function.

6 Conclusions and Future Research

In this work, we have presented a comparative evaluation among three different Spanish segmentation strategies for Spanish-Chinese transliteration, as well as two other important parameters of the transliteration system implementation: target language model order and alignment strategy for bilingual unit extraction. The transliteration task was implemented by means of Statistical Machine Translation, using Chinese characters and Spanish sub-word segments as the tex-

tual units to be translated. The three different Spanish segmentation strategies evaluated were: character-based, syllabic-based and a proposed sub-syllabic segmentation scheme. Experimental results shown that syllabic-based segmentation, along with a language model of order 4 and the grow-diag-final-and alignment method, constitutes the most effective strategy for Spanish-to-Chinese transliteration, while the proposed sub-syllabic segmentation, along with a language model of order 2 and the source-to-target alignment method, constitutes the most effective strategy for Chinese-to-Spanish transliteration.

As an additional contribution, and due to the lack of dataset for Chinese-Spanish transliteration research, we have constructed an experimental parallel corpus containing a total of 841 named entities in both Chinese and Spanish.

As future research work, we intend to expand the experimental dataset, as well as to continue evaluating the specific peculiarities of both Chinese-to-Spanish and Spanish-to-Chinese transliteration tasks. A comprehensive manual evaluation on the experimental results described here should be conducted in order to identify both, possible improvements to the proposed Spanish sub-syllabic segmentation method and some additional strategies for improving the performance of transliteration quality between Chinese and Spanish.

Acknowledgments

The author would like to thank the Institute for Infocomm Research (I^2R) and the Agency for Science, Technology And Research (A*STAR) for their support and permission to publish this work.

References

Connie R. Adsett, Yannick Marchand, and Vlado Keselj, 2009, Syllabification rules versus data-driven methods in a language with low syllabic complexity: The case of Italian, *Computer Speech and Languages*, 23(4): 444-463.

Yaser Al-Onaizan and Kevin Knight, 2002, Machine Transliteration of names in Arabic text,In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Philadelphia, PA.

Mansur Arbabi, Scott M. Fischthal, Vincent C. Cheng, and Elizabeth Bart, 1994, Algorithms for Arabic name transliteration, *IBM Journal of Research and Development*, 38(2):183-193.

Manoj Kumar Chinnakotla, and Om P. Damani, 2009, Experiences with English-Hindi, English-Tamil and English-Kannada Transliteration Tasks at NEWS 2009, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), pages 44-47, Singapore.

Marta R. Costa-jussa, Carlos A. Henriquez, and Rafael E. Banchs, 2011, Evaluating Indirect Strategies for Chinese-Spanish statistical machine translation with English as pivot language, In *Proceedings of the 27th Conference of the Spanish Society for Natural Language Processing*, Huelva, Spain.

Heriberto Cuayahuitl, 2004, A Syllabification Algorithm for Spanish, in A. Gelbukh (Ed.): CICLing 2004, LNCS 2945, pages 412-415, Springer.

Amitava Das, Asif Ekbal, Tapabrata Mondal, and Sivaji Bandyopadhyay, 2009, English to Hindi Machine Transliteration System at NEWS 2009, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), pages 80-83, Singapore.

Michel Divay and Anthony J. Vitale, 1997, Algorithms for grapheme-phoneme translation for English and French: Applications, *Computational Linguistics*, 23(4):495-524.

Rejwanul Haque, Sandipan Dandapat, Ankit Kumar Srivastava, Sudip Kumar Naskar, and Andy Way, 2009, English-Hindi Transliteration Using Context-Informed PB-SMT: the DCU System for NEWS 2009, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), pages 104-107, Singapore.

Gumwon Hong, Min-Jeong Kim, Do-Gil Lee, and Hae-Chang Rim, 2009, A Hybrid Approach to English-Korean Name Transliteration, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), pages 108-111, Singapore.

Martin Jansche and Richard Sproat, 2009, Named Entity Transcription with Pair n-Gram Models, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), pages 32-35, Singapore.

Yuxiang Jia, Danqing Zhu, Shiwen Yu, 2009, A Noisy Channel Model for Grapheme-based Machine Transliteration, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), pages 88-91, Singapore.

G. A. Kiraz and B. Mobius, 1998, Multilingual syllabification using weighted finite-state transducers, In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia.

Keving Knight and Jonathan Graehl, 1998, Machine Transliteration, *Computational Linguistics*, 24(4): 599-612.

Kevin Knight, 2009, Automata for Transliteration and Machine Translation, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), page 27, Singapore.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, 2007, MOSES: Open source toolkit for statistical machine translation, In *Proceedings of the 45th ACL Annual Meeting*, pages 177-180, Prague, Czech Republic.

Oi Y. Kwong, 2009, Graphemic Approximation of Phonological Context for English-Chinese Transliteration, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), pages 186-193, Singapore.

J. Lafferty, A. McCallum, and F. Pereira, 2001, Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data, In *Proceedings of the International Conference on Machine Learning*, pages 282-289.

Haizhou Li, Min Zhang, and Jian Su, 2004, A joint source-channel model for machine transliteration, In *Proceedings of the 42nd ACL Annual Meeting*, pages 159-166, Barcelona, Spain.

Hua Lin and Qian Wang, 2007, Mandarin Rhythm: An Acoustic Study, *Journal of Chinese Language and Computing*, 17(3): 127-140.

Sara Noeman, 2009, Language Independent Transliteration system using phrase based SMT approach on substrings, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), pages 112-115, Singapore.

Franz J. Och and Hermann Ney, 2000, A comparison of alignment models for statistical machine translation, In *Proceedings of the 18th Conference on Computational Linguistics*, pages 1086-1090, Morristown, NJ.

Franz J. Och, 2003, Minimum error rate training in statistical machine translation, *In Proceedings of the 41st ACL Annual Meeting*, pages 160-167, Sapporo, Japan.

Antonio Pamies Bertran, 1999, Prosodic Typology: On the Dichotomy between Stress-Timed and Syllable-Timed Languages, *Language Design*, 2: 103-130.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, 2001, BLEU: a method for automatic evaluation of machine translation, *IBM Research Report RC-22176*.

Kenneth L. Pike, 1945, Step-by-step procedure for marking limited intonation with its related features of pause, stress and rhythm, in Charles C. Fries (Ed.), *Teaching and Learning English as a Foreign Language*, pages 62-74, Publication of the English Language Institute, University of Michigan, Ann Arbor.

Feiliang Ren, Muhua Zhu, Huizhen Wang, and Jingbo Zhu, 2009, Chinese-English Organization Name Translation Based on Correlative Expansion, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), pages 143-151, Singapore.

Tao Tao, Su-Youn Yoon, Andrew Fister, Richard Sproat and Cheng-Xiang Zhai, 2006, Unsupervised Named Entity Transliteration Using Temporal and Phonetic Correlation, In *Proceedings of Empirical Methods in Natural Language Processing*, pages 22-23, Sydney, Australia.

Kommaluri Vijayanand, 2009, Testing and Performance Evaluation of Machine Transliteration System for Tamil Language, In *Proceedings of ACL-IJCNLP 2009 Named Entities Workshop* (NEWS 2009), pages 48-51, Singapore.

Su-Youn Yoon, Kyoung-Young Kim, and Richard Sproat, 2007, Multilingual Transliteration Using Feature based Phonetic Method, In *Proceedings of the 45th ACL Annual Meeting*, pages 112-119, Prague, Czech Republic.

Min Zhang, A. Kumaran, and Haizhou Li, 2011, Whitepaper of NEWS 2011 Shared Task on Machine Transliteration, In *Proceedings of IJCNLP 2011 Named Entities Workshop* (NEWS 2011), retrieved on June 15, 2011, from http://translit.i2r.a-star.edu.sg/news2011/news2011whitepaper.pdf