Review of Machine Transliteration Techniques

Kamaljeet Kaur
Department of Computer Science and Engineering
Guru Nanak Dev Engineering College,
Ludhiana (Punjab) India

Parminder Singh, Ph.D.
Associate Professor,
Department of Computer Science and Engineering
Guru Nanak Dev Engineering College,
Ludhiana (Punjab) India

ABSTRACT

Transliteration is the conversion of a text from one script to another, and thus representing words from one language using the approximate phonetic or spelling equivalents of another language. Machine Transliteration has come out to be an emerging and a very important research area in the field of machine translation. Transliteration systems are very beneficial for removing the language and scriptural barriers. It has gained prime importance as a supporting tool for machine translation and cross-language information retrieval, especially when proper names and technical terms are involved. Various techniques are available for transliteration process. This paper is intended to give a brief overview of commonly used machine transliteration techniques.

Keywords

Transliteration, transliteration techniques, statistical machine translation, natural language processing.

1. INTRODUCTION

Machine transliteration has come out to be an emerging and a very important research area in the field of machine translation. Transliteration basically aims to preserve the phonological structure of words. Proper transliteration of name entities plays a very significant role in improving the quality of machine translation. The performance of machine translation and cross-language information retrieval depends extremely on accurate transliteration of named entities. Transliteration is the conversion of a text from one script to another. It involves representing words from one language using the approximate phonetic or spelling equivalents of another language.

From an information-theoretical point of view, systematic transliteration is a mapping from one system of writing into another, word by word, or ideally letter by letter. Transliterating a word from the language of its origin to a foreign language is called Forward Transliteration. For example English to Punjabi transliteration. On the other hand, transliterating a loan-word (a word borrowed from other language and incorporated) written in a foreign language back to the language of its origin is called Backward Transliteration. Machine transliteration can play an important role in various natural language applications: information retrieval and machine translation, cross-language applications, data mining and information retrieval system. This paper represents the approaches toward the process of machine transliteration.

2. MACHINE TRANSLITERATION TECHNIQUES

Various techniques for transliteration are being used, each having its own advantages and disadvantages. These techniques are categorized as shown in Figure 1.

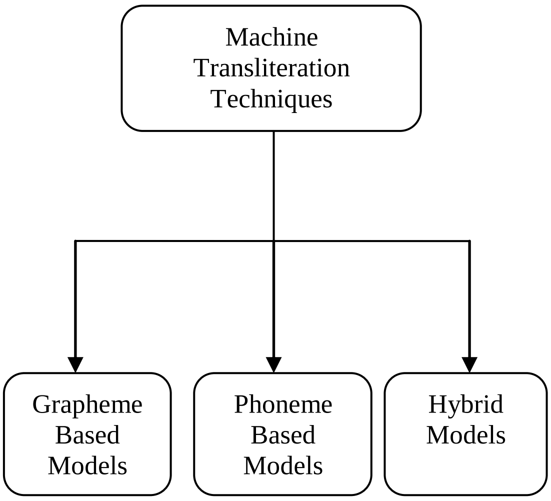

Fig 1: Machine Transliteration Techniques

This section gives a brief description of various approaches towards machine transliteration, used by various transliteration attempts. These approaches for machine transliteration are described briefly in the following sub sections:

2.1 Grapheme based Models

Grapheme refers to the basic unit of a written language that has its own meaning or grammatical importance. In Grapheme based approaches, transliteration is viewed as a process of mapping a grapheme sequence from a source language to a target language ignoring the phoneme level processes. Grapheme based models work by directly transforming source language graphemes into target language graphemes. So, they are also called direct methods. Grapheme based models are classified into the Statistical Machine Transliteration (SMT) based model, Rule based models, Hidden Markov Model (HMM), Finite State Transducer (FST) based model etc.

2.1.1 Rule based Approach

A rule based machine transliteration system consists of collection of rules called grammar rules, lexicon and software programs to process the rules. Rule based approach is the first strategy ever developed in the field of machine transliteration. RBMT (Rule Based Machine Transliteration) has much to do with morphological, syntactic and semantic information about the source and target language. Linguistic rules are built over this information. Rules play major role in various stages of translation: syntactic processing, semantic interpretation, and contextual processing of language. Rule based transliteration is based on linguistic information about source and target languages.

The main approach of RBMT systems is based on linking the structure of the given input sentence with the structure of the demanded output sentence, necessarily preserving their unique meaning. In the rule based approach, human experts specify a set of rules, aiming at describing the translation process. It then applies rules that map the grammatical segments of the source sentence to a representation in the target language. There are different types of rule based machine translation systems: Direct Systems (Dictionary Based Machine Translation) map input to output with basic rules. RBMT systems employ morphological and syntactical analysis. Interlingual RBMT systems use an abstract meaning. Transliteration in rule based system is done by pattern matching of the rules. The success lies in avoiding the pattern matching of unfruitful rules. General world knowledge is required for solving interpretation problems such as disambiguation [12].

Deep and Goyal (2011) have proposed Punjabi to English transliteration system using rule based approach [2]. Goyal and Lehal (2009) have developed Hindi to Punjabi transliteration system using rule based approach. They have implemented various rules for Hindi to Punjabi transliteration [4].

2.1.2 SMT Approach

Statistical based transliteration approaches tend to be computationally easier in language transliteration than trying to parse and evaluate grammatical rules. Statistical approaches employ various mathematical techniques. Statistical MT models take the view that every sentence in the target language is a translation of the source language sentence with some probability [1]. The best translation, of course, is the sentence that has the highest probability. The key problems in statistical MT are: estimating the probability of a translation, and efficiently finding the sentence with the highest probability. It works by finding most probable English sentence given a foreign language sentences, automatically align words and phrases within sentence pairs in a parallel corpus and then probabilities are determined automatically by training a statistical model using the parallel corpus. Based on the probabilities sentence get transliterated. Statistical approaches have a number of advantages over these non-statistical techniques. The primary advantage is that they have been shown to produce better transliteration. It has a way of dealing with lexical ambiguity.

Kaur and Josan (2011) have proposed English to Punjabi transliteration system using statistical machine transliteration approach. They have also developed statistical model that is used for transliterating Punjabi text into Hindi text [6]. Kumar and Kumar (2013) have proposed statistical machine transliteration system that is used to transliterate proper nouns written in Punjabi language into its equivalent English language [9].

2.1.3 FST Approach

Finite State Transducers are models that are being used in different areas of pattern recognition and computational linguistics. In the area of machine transliteration the transducer based approaches that are based on building models automatically from training examples are becoming more and more attractive. A transducer has the intrinsic power of transducing or transliterating. Whenever the transducer shifts from one state to another, it will print the output word, if any. So, as a result, not only will it accept the sentence of one language, but it will print the transliteration in another language. Alternatively, a transducer can be seen as a bilingual generator. A FST is an automaton that transforms one string into another. It can be seen as a network of states with transitions between them which are labeled with input and output symbols. Starting at some state and walking through the automaton to some end state, the FST can transform an input string by matching the input labels to an output string by printing corresponding output labels [8]. Figure 2 shows an example of an FST where each arc is labeled by an input and output strings separated by a colon while the nodes represent states.

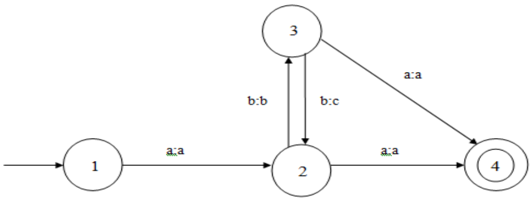

Fig 2: Example of Finite State Transducer

Knight and Graehl (1998) have developed a phoneme based statistical model using finite state transducer to do back-transliteration from Japanese to English [8]. Stall and Knight (1998) have built a model to transliterate names from Arabic into English. Their proposed system's implementation is also based on finite state transducer [14].

2.1.4 HMM Approach

A frequently used statistical model is the Hidden Markov Model (HMM). In Hidden Markov Models, given some sentence x = (x1, x2,…., xn) in the language you want to transliterate from, you want to predict what the most likely sentence y = (y1, y2,…, ym) is going to be in the language you want to transliterate to. The HMM is a finite set of states, each of which has probability distribution. Transitions between the states are kept in control by a set of probabilities called transition probabilities. In a particular state translation can be obtained according to the associated probability distribution. The HMM works by looking at all possible combinations of a sequence of one language words in another language which computes probabilities of co-occurrence of words based on a given tagged corpus and then tags texts using these probabilities. The strengths of HMM is its mathematical framework and its implementation structure.

Nabende (2009) has presented a transliteration system that is based on pair-HMM training on English-Russian data set [11].

2.2 Phoneme based Models

Phonemes are the smallest significant unit of sound or the smallest contrastive units of spoken language. Grapheme based models work by directly transforming source language graphemes into target language graphemes without explicitly utilizing phonology in the bilingual mapping. Phoneme based models, on the other hand, do not utilize orthographic information in the transliteration process. Phoneme based method is also known as Pivot method. Figure 3 shows phoneme based process.

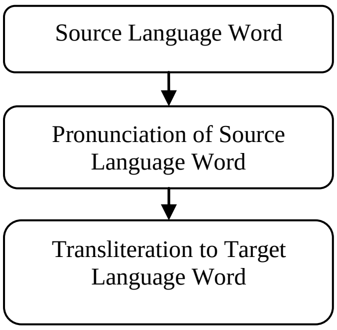
Source Language Word
Pronunciation of Source Language Word
Transliteration to Target Language Word

Fig 3: Phonemic Approach

Name transliteration occurs on the basis of pronunciation. That is, the written word of source language is mapped to written word of target language via the spoken form associated with the word. Phoneme based models are generally implemented in two steps:

i. Source Grapheme-to-Source Phoneme Transformation: It involves mapping of source language word (grapheme) to phonemic representation.

ii. Source Phoneme-to-Target Grapheme Transformation: It involves mapping of each source phoneme, composing the word, to a corresponding target language word.

Thus phoneme based models first obtain the source language pronunciation and then convert that representation into the target language graphemes. In phoneme based approaches, the transliteration key is pronunciation of the source phoneme rather than spelling or the source grapheme.

The phoneme based approach has also received remarkable attention in various works. This approach is basically source grapheme-to-source phoneme transformation and source phoneme-to-target grapheme transformation. Based on phonology, the source text can usually be transliterated into its target counterpart in terms of pronunciation similarities between them. The syllables are mapped to phonemes, based on some transcription rules. Transliteration rules are mapping templates between phonemes of source and target language. Although the phoneme based transliteration produces recognizable results but there is need to perform more formal analysis of the correspondences between source and target language phonemes [15].

Dhore et al. (2012) have used a phonetic approach for Hindi to English and Marathi to English transliteration [3]. To improve Punjabi to Hindi transliteration, Josan and Lehal (2010) have employed phonetic matching technique. They have performed character mapping based on phonetic sounds [5].

2.3 Hybrid Models

However, transliteration is a complex process, which does not rely on either source grapheme or phoneme. There is need to combine two or more transliteration approaches for better result. Hybrid approaches simply combines the grapheme based transliteration probability and the phoneme based transliteration probability using linear interpolation. Hybrid machine transliteration approach strength the statistical and rule based transliteration methodologies.

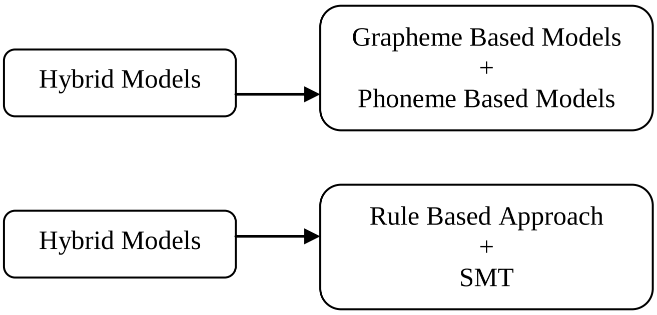
Hybrid Models
Grapheme Based Models + Phoneme Based Models
Hybrid Models
Rule Based Approach + SMT

Fig 4: Hybrid Approaches

Several researchers have used a variety of approaches to machine transliteration that involve either modeling a direct mapping between two orthographies or considering the phonetic representation for transforming strings into each other or a combination of both. Nowadays there are many transliteration systems available, where some are rule based and some are based on statistical approach.

Hybrid systems are combination of two or more transliteration approaches and leads to an appropriate transliteration. This reduces the rate of error in transliteration system to a great extent. It can be combination of grapheme based model and phoneme based model or can be combination of any grapheme based models. For example, statistical machine transliteration with rule based approach. It is possible to perform transliterations using a rules based approach. Statistics can be then used in an attempt to adjust/correct the output from the rules engine.

Hybrid models have the limited power for producing the correct transliterations because it just combines grapheme models and phoneme models. It does not consider correspondence between source grapheme and phoneme during the transliteration process. The correspondence plays important roles in machine transliteration. If the correspondence between source grapheme and phoneme in given context is known, then one can more easily infer the correct transliteration [13].

Lee and Choi (1998) have proposed English to Korean transliteration system. They have proposed a hybrid method that is more effective for transliteration. They have used statistical machine translation as a base system for both direct and pivot method [10]. Khantonthong et al. (2000) have proposed hybrid model to do automatic backward transliteration from Thai into English. The proposed hybrid approach is the combination of a statistical model and a set of context sensitive rules [7].

3. CONCLUSION

A review of the different machine transliteration techniques is presented in this paper. Various techniques for machine transliteration being used have been described. Brief outline about the existing approaches those have been used to develop machine transliteration systems is given here. Although grapheme based approaches are simple to implement as these involve direct mapping of words, but without phonetic information it may be difficult to obtain more relevant result. Hybrid approaches reduce the rate of error in transliteration system to a great extent, but it does not consider correspondence between source grapheme and phoneme during the transliteration process. Almost all existing Indian languages as well as foreign languages machine transliteration systems are based on statistical and hybrid approaches of transliteration.

4. REFERENCES

[1] AbdulJaleel, N. and Larkey, L. (2003), “Statistical transliteration for English-Arabic cross language information retrieval”, in proceedings of the 12th International Conference on Information and Knowledge Management, New York, USA, pp. 139-146.

[2] Deep, K. and Goyal, V. (2011), “Development of a Punjabi to English Transliteration System”, International Journal of Computer Science and Communication, Vol. 2, No. 2, pp. 521-526.

[3] Dhore, M., Dixit, S. and Dhore, R. (2012), “Hindi and Marathi to English NE Transliteration Tool using Phonology and Stress Analysis”, in proceedings of 24th International Conference on Computational Linguistic, Mumbai, India, pp. 111-118.

[4] Goyal, V. and Lehal, G. (2009), “Hindi-Punjabi Machine Transliteration System (For Machine Translation

System)", George Ronchi Foundation Journal, Italy, Vol. 64, No. 1, pp. 1-7.

[5] Josan, G. and Lehal, G. (2010), "A Punjabi to Hindi Machine Transliteration System", Computational Linguistics and Chinese Language Processing, Vol. 15, No. 2, pp. 77-102.

[6] Kaur, J. and Josan, G. (2011), "Statistical Approach to Transliteration from English to Punjabi", International Journal on Computer Science and Engineering, Vol. 3, No. 4, pp. 1518-1527.

[7] Khantonthon, N., Kawtraku, A. and Poovarawan, Y. (2000), "An Enhancement of Thai Text Retrieval Efficiency by Automatic Backward Transliteration", in proceedings of 7th International Workshop of Academic Information Networks on Systems, Bangkok, Thailand, pp. 73-84.

[8] Knight, K. and Graehl, J. (1998), "Machine transliteration", in proceedings of the 35th annual meetings of the Association for Computational Linguistics, Madrin, Spain, pp. 128-135.

[9] Kumar, P. and Kumar, V. (2013), "Statistical Machine Translation Based Punjabi to English Transliteration System for Proper Nouns", International Journal of Application or Innovation in Engineering & Management, Vol. 2, No. 8, pp. 318-321.

[10] Lee, J. and Choi, K. (1998), "English to Korean Statistical transliteration for information retrieval", Computer Processing of Oriental Languages, Vol. 12, pp. 17-37.

[11] Nabende, P. (2009), "Transliteration system using pair HMM with weighted FSTs", in proceedings of Named Entities Workshop on Shared Task on Transliteration, pp. 100-103.

[12] Oh, J. and Choi, K. (2002), "An English-Korean Transliteration Model Using Pronunciation and Contextual Rules", in proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan, pp. 758-764.

[13] Oh, J. and Choi, K. (2005), "An Ensemble of Grapheme and Phoneme for Machine Transliteration", in proceedings of 2nd International Joint Conference on NLP, Jeju Island, pp. 450-461.

[14] Stalls, B. and Knight, K. (1998), "Translating Names and Technical Terms in Arabic Text", in proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages, Montreal, Canada, pp. 34-41.

[15] Wan, S. and Verspoor, C. (1998), "Automatic English-Chinese name transliteration for development of multilingual resources", in proceedings of the 17th International Conference on Computational Linguistics, Montreal, Canada, pp. 1352-1356.