Weak Translation Problems – a case study of Scriptural Translation

M. G. Abbas Malik[1] Christian Boitet[1] Pushpak Bhattacharyya[2] Laurent Besacier[1]

(1) GETALP-LIG, Université de Grenoble (UJF), France
(2) Indian Institute of Technology Bombay (IITB), India
Abbas.Malik@imag.fr, Christian.Boitet@imag.fr, pb@iitb.ac.in, Laurent.Besacier@imag.fr

Résumé La TA généraliste de haute qualité et totalement automatique est considérée comme impossible. Nous nous intéressons aux *problèmes de traduction scripturale,* qui sont des sous-problèmes faibles du problème général de la traduction. Nous présentons les caractéristiques des problèmes faibles de traduction et les problèmes de traduction scripturale, décrivons différentes approches computationnelles (à états finis, statistiques, et hybrides) et présentons nos résultats sur différentes combinaisons de langues et systèmes d'écriture Indo-Pak.

Abstract General purpose, high quality and fully automatic MT is believed to be impossible. We are interested in *scriptural translation problems,* which are weak sub-problems of the general problem of translation. We introduce the characteristics of the weak problems of translation and of the scriptural translation problems, describe different computational approaches (finite-state, statistical and hybrid) to solve these problems, and report our results on several combinations of Indo-Pak languages and writing systems.

Mots-clés : problèmes faibles de traduction, traduction scripturale, traduction interdialectal, transcriptions, translittérations

Keywords: weak problems of translation, scriptural translation, interdialectal translation, transcription, transliteration

1 Introduction

We say that a problem of translation is *weak* if the sentence correspondence defined by that problem is such that there is always a very small number of target sentences corresponding to a given source sentence, if it is presented in isolation, and almost always only 1 target sentence in a given context. On the contrary, a general translation problem is *strong* because the number of target sentences corresponding to a source sentence of length n can be exponential (k^n for some k), even in the context of a connected text.

Such weak translation problems include (1) transliteration or/and transcription between 2 scripts, which we call "scriptural translation", (2) phonetic transcription, and (3) interdialectal translation, that implies lexical changes. Table 1 gives a list of some generic and specific sub-problems of *weak translation problems* and their instances in increasing complexity and difficulty order.

In this paper, we concentrate only on *scriptural translation problems*, because they are important in practice, especially for many Indo-Pak languages (that are spoken both in India and Pakistan), but also for other languages like Malay, Mongolian and many languages of Central Asia. Scriptural translation problems are significant because they create a *scriptural divide* that often reinforces political divisions and mutual misunderstandings, even if communities using different scripts perfectly understand each other

orally. In the case of the Indo-Pak languages (Hindi, Urdu, Punjabi, Seraiki, Sindhi and Kashmiri), these languages have more than 1178 million speakers around the world, more than the total population of Europe.

Generic sub-problems	Specific sub-problems	Instances	Constraints
Language localization	Word for word translation Intralingual translation	Québecois – French	SL = TL SW = TW
		Malay – Indonesian	
Scriptural translation	Transliteration Transcription Phonetic transcription	Punjabi/Shahmukhi – Punjabi/Gurmukhi	SL = TL SW ≠ TW
		Malay/Roman – Malay/Jawi	
		French/Roman – French/IPA	
	Transliteration Transcription	Hindi – Urdu	SL ≠ TL SW ≠ TW
		Bengali – Assamese	
		Hindi – Marathi	
Interdialectal translation	Word for word translation Intralingual translation Scriptural translation	Québecois – French	SL = TL SW = TW
		English (UK) – English (USA)	
		Punjabi/Shahmukhi – Punjabi/Gurmukhi	SL = TL SW ≠ TW
		Malay/Jawi – Indonesian/Latin	
Bilingual translation	Word for word translation Scriptural translation Bilingual translation between	Bengali – Assamese	SL ≠ TL SW ≠ TW
		Hindi – Marathi	
SL = source language, TL = target language, SW = source writing system, TW = target writing system			

Table 1. Sub-problems of weak translation (by order of increasing complexity)

2 Scriptural Translation

Scriptural translation is the process of transcribing a word written in the source language script into the target language script by preserving its articulation in the original language in such a way that the native speaker of the target language can produce the original pronunciation. It is a weak translation problem for Indo-Pak languages, but it is a strong translation problem Japanese-English and French-Chinese scriptural translation. An example of Hindi-Urdu scriptural translation is shown in Figure 1.

دنیا کو امن کی ضرورت ہے۔

दुनिय को अमन की ज़रूरत है।

[d̪unɪjɑ ko əmən ki zərurət̪ hæ.]

World needs peace.

Figure 1. Hindi–Urdu Example Sentence

It differs from general transliteration in various aspects. Usually, transliteration handles only Out-Of-Vocabulary (OOV) words, Named Entities (NEs), *etc*. On the other hand, *scriptural translation* must handle all kinds of words irrespective of their type and it provides basis for *Cross-Scriptural Machine Translation* (CSMT), *Cross-Scriptural Information Retrieval* (CSIR), *Cross-Scriptural Application Development* (CSAD), *Inter-dialectal Machine Translation* (IMT), *Cross-Dialectal Information Retrieval* (CDIR) and for solving the *weak translation problems*.

Scriptural translation problems are quite interesting theoretically because they often cannot be solved by a simple character-to-character replacement method. Even in their weak instances (such as Hindi-Urdu scriptural translation), they may be complex (need more than usual Finite-State Transducers (FST) to reach a quality high enough for real usage) and difficult (require human work to build resources, and expert knowledge). They offer the advantage that measures of accuracy based on references are valid, as there is in general only one "good" translation, which can serve as a reference.

2.1 Challenges for scriptural translation

2.1.1 Scriptural divide

There exists a written communication gap between people who can understand each other verbally but cannot read each other. They are virtually divided and become *scriptural aliens*. Examples are the Hindi & Urdu communities, the Punjabi/Shahmukhi & Punjabi/Gurmukhi communities, *etc.* An example of scriptural divide is shown in Figure 1.

2.1.2 Under-resourcedness

Under-resourcedness of a source or target language is a big challenge for *scriptural translation*. The lack of standards in writing practices or code page for a language makes scriptural translation very hard. The existence of different writing systems for a language results in a large number of variants and it becomes difficult and complex to handle them. For example, Punjabi is the largest language of Pakistan (more than 70 million), but there existed only two magazines in 1992 (Rahman 1997). In the words of (Rahman 2004), "… *there is little development in Punjabi, Pashto, Balochi and other languages*…". (Malik 2005) reports the first effort towards establishing a standard code page for Punjabi/Shahmukhi, and there is none yet for Shahmukhi. The Kashmiri and Seraiki languages are similarly under-resourced.

2.1.3 Missing necessary information

In some cases, the necessary and vital information for scriptural translation is missing in the source text. For example, the first word of the example sentence of Figure 1 دنیا [ɖ̤ʊnɪjɑ] (world) misses essential information that is obligatory to do Urdu to Hindi transliteration. Diacritical marks are part of the writing system but are sparingly used in writings (Zia 1999). Figure 2(a) and 2 (b) show the correct and the wrong Hindi transliteration of the example Urdu word respectively.

(a) with necessary information	(b) without necessary information
د ُ نِ ْ یا – د [ɖ̤] ُ [ʊ] ن [n] ِ [ɪ] ی [j] ا [ɑ]	دنیا – د [ɖ̤] ن [n] ی [j] ا [ɑ]
दु [ɖ̤ʊ] न [n] ि [ɪ] य [j] ा [ɑ] – दुनिया	द [ɖ̤] न [n] य [j] ा [ɑ] – दनया

Figure 2. Example of absence of information

2.1.4 Dissimilar spelling conventions

Different spelling conventions exist across different scripts used for the same language or for different languages because users of a script have learned to write certain words in a traditional way. For example, the words یہ [je] (this) = ی [j] + ہ [h] and وہ [vo] (that) = و [v] + ہ [h] are used in Urdu and Punjabi/Shahmukhi. The character ہ [h] produces the vowel sounds [e] and [o] in the example words respectively. On the other hand, the example words are written as ये [je] & वो [vo] and ਯੇ [je] & ਵੇ [vo] in Devanagari and Gurmukhi, respectively. There exist a large number of such different conventions between Punjabi/Shahmukhi–Punjabi Gurmukhi, Hindi–Urdu, *etc.*

Different spelling conventions are also driven by different religious influences on different communities. In the Indian sub-continent, Hindi is a part of the Hindu identity, while Urdu is a part of the Muslim identity[1] (Rahman 1997; Rai 2000). Hindi derives its vocabulary from Sanskrit, while Urdu borrows its

[1] The Hindi movement of the late 19th century played a central role in the ideologization of Hindi. The movement started in reaction to the British Act 29 of 1837 by which Persian was replaced by Hindustani/Urdu, written in Persian script, as the

literary and scientific vocabulary from Persian and Arabic. Hindi and Urdu not only borrow from Sanskrit and Persian/Arabic, but also adopt the original spellings of the borrowed word due the sacredness of the original language. These differences make scriptural translation across scripts, dialects or languages more challenging and complex.

2.1.5 Transliteration and Transcription Ambiguities

Transliteration or transcription at character level across different scripts is ambiguous. For example, a Sindhi word انسان[ɪɲsɑn] (human being) can be transliterated into Devanagari either as इंसान [ɪɲsɑn] or इसान * [ɪnsɑn] (* means wrong spelling). The transliteration process of the word from Sindhi to Devanagari is shown in Figure 3(a). The transliteration of the third character from the left, Noon (ن) [n] is ambiguous because in the middle of a word, Noon may represent a consonant [n] or the nasalization [ɲ] of a vowel.

In the reverse direction, the Sindhi Devanagari word इंसान [ɪɲsɑn] can be transliterated into a set of possible transliterations [انسان, اِنصان*, اِنثان*]; all of these possible transliterations have the same pronunciation [ɪɲsɑn] but have different spellings in the Perso-Arabic script, as shown in Figure 3(b). Similar kinds of ambiguities also arise for other pairs of scripts or languages. Thus transliteration ambiguities increase the complexity and hardness of the *scriptural translation* across scripts of the same language or of different languages.

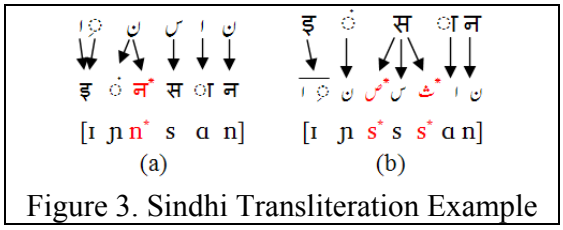
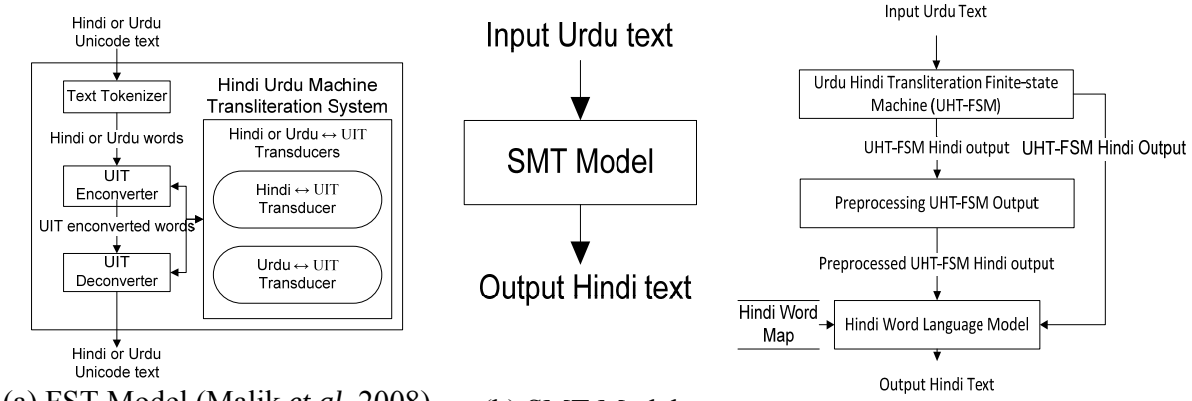


Figure 3. Sindhi Transliteration Example

3 Computational Models

Figure 4 shows different computational models employed for scriptural translation, discussed later.



(a) FST Model (Malik *et al.* 2008) (b) SMT Model (c) Hybrid Model (Malik *et al.* 2009)

Figure 4. System architectures of different models for scriptural translation

3.1 Finite-state scriptural translation system using UIT

Universal Intermediate Transcription (UIT) is a multipurpose pivot. It is used as a *phonetico-morphotactic* intermediate representation for the *surface morphotactic translation* or scriptural translation. It is a deterministic and an unambiguous scheme of transcription for Indo-Pak languages based on Speech

official vernacular of the courts of law in North India. It is the moment in history, when Hindi and Urdu started to emerge as Hindu and Muslim identities.

Assessment Methods Phonetic Alphabet (SAMPA) in ASCII range 32–126, since a text in this range is portable across computers and operating systems (Hieronymus 1993; Wells 1995).

In our finite-state model, UIT serves the purpose of a phonetico-scriptural pivot. The source text is converted into its UIT representation and this UIT representation is then transformed into the target text. Therefore, scriptural translation is a "double translation". We have used FST to perform these two scriptural translations. Our FST model is shown in Figure 4(a) and fully discussed in (Malik *et al.* 2008). We have developed FSTs for this double translation for Hindi-Urdu, Punjabi and Seraiki.

3.2 SMT systems

SMT systems are direct translation systems for the Hindi-Urdu pair. They have been developed from a parallel lexicon containing 50,000 Hindi-Urdu words. Each parallel entry contains Hindi and Urdu scripturally parallel words. A space has been introduced after each character in each Hindi and Urdu parallel word to develop Hindi-Urdu parallel training data at characters-level. In character-level SMT systems, a character is a translation unit, exactly like a word in an SMT system for the English-French pair, developed from English-French parallel corpora.

We have also developed an expert clustering algorithm to build cluster-level parallel training data. A cluster is a set of characters that must be considered as one translation unit. For example, in Punjabi, an aspirated consonant is written as a combination of the consonant to be aspirated and HEH-DOACHASHMI (ہ), and these two characters must be considered as one translation unit for scriptural translation. Therefore, we have two types of SMT systems, (1) character-level and (2) cluster-level. To check the effect of different parameters on our SMT systems, like *reordering* and *tuning*, we have varied these parameters during the development phase. In short, we have developed 12 SMT systems for Hindi to Urdu scriptural translation, and another set of 12 SMT systems for Urdu to Hindi scriptural translation.

3.3 Hybrid model

In the case of Indo-Pak languages, Urdu, Seraiki and Punjabi are written in Pakistan in a derivation of the Persio-Arabic script. Like for Arabic, people do not usually use the diacritical marks, which contain very crucial information to perform scriptural translation from that script to the other script used (usually a variant of Devanagari). Our FST model is a kind of expert system, and missing information badly affects its performance. On the other hand, SMT is a data-driven model, but the results are still not satisfactory as far as usability in real context is considered.

To compensate the absence of necessary information and produce sufficient results, we have proposed a *hybrid model* (Malik *et al.* 2009). This model is a multilevel process. It firsts performs the scriptural translation of the source text in the target language/script and relates each source word with a set of target words. The cardinality of this relation is 1 in most cases, but can reach at maximum of 4. The target language model is used as a *statistical filter* that uses the target language knowledge to filter out incorrect solutions and produce the correct one. The concept of our hybrid model is shown in Figure 5.

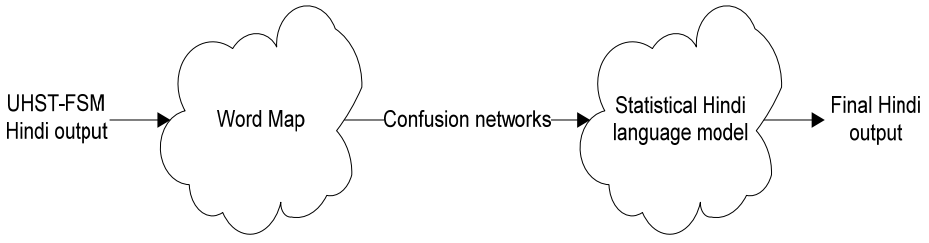


Figure 5. Statistical disambiguation process

4 Results

For automatic testing purposes, we manually developed a Hindi-Urdu test corpus, containing 200 sentences (4,250 words). We randomly selected 200 sentences from our Hindi corpus and then translated them into Urdu. The results of our different computational models are given in Table 2. For SMT, we have only reported the best results.

Model	Urdu source with diacritics		Urdu source without diacritics	
	Word Accuracy	**Sentence Accuracy**	**Word Accuracy**	**Sentence Accuracy**
Finite-state system	83.9%	10%	53%	1%
Statistical system	72.2%	5.5%	77%	5%
Hybrid system	85.8%	14%	79.1%	7%

Table 2. Urdu to Hindi scriptural translation results

5 Conclusion

The *weak translation problems* are sub-problems of translation that only admit a very small set of correct solutions, in general only one, for a given source sentence in a given context. *Scriptural translation*, a weak sub-problem of translation, is almost always a weak problem. Three approaches have been employed to solve scriptural translation problems and results have been reported for Urdu to Hindi scriptural translation. We found that the hybrid approach gives better results than the FST and SMT approaches. The analysis of *scriptural translation* leads to the conclusion that the *weak translation problems*, which seemed to be simple on the first hand, can be complex and hard problems.

References

HIERONYMUS; (1993). Ascii Phonetic Symbols for the World's Languages: Worldbet, AT&T Bell Laboratories.

MALIK; (2005). Towards a Unicode Compatible Punjabi Character Set. *27th Internationalization and Unicode Conference*, Berlin.

MALIK, BESACIER, BOITET and BHATTACHARYYA; (2009). A Hybrid Model for Urdu Hindi Transliteration. *Joint conference of the 47th Annual Meeting of the Association of Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of NLP ACL/IJCNLP Workshop on Named Entities (NEWS-09)*, Singapore.

MALIK, BOITET and BHATTACHARYYA; (2008). Hindi Urdu Machine Transliteration Using Finite-State Transducers. *22nd International Conference on Computational Linguistics (COLING)*, Manchester, ICCL.

RAHMAN; (1997). *Language and Politics in Pakistan*. Lahore, Oxford University Press.

RAHMAN; (2004). Language Policy and Localization in Pakistan: Proposal for a Paradigmatic Shift. *Crossing the Digital Divide, SCALLA Conference on Computational Linguistics*, Katmandu.

RAI; (2000). *Hindi Nationalism*. New Delhi, Orient Longman Private Limited.

WELLS; (1995) Computer-Coding the Ipa: A Proposed Extension of Sampa. www.phon.ucl.ac.uk/home/sampa.

ZIA; (1999). Standard Code Table for Urdu. *4th Symposium on Multilingual Information Processing (MLIT-4)*, Yangon, CICC.