Tajik-Farsi Persian Transliteration Using Statistical Machine Translation

Chris Irwin Davis
Human Language Technology Research Institute
Richardson, Texas
E-mail: cid@hlt.utdallas.edu

Abstract

Tajik Persian is a dialect of Persian spoken primarily in Tajikistan and written with a modified Cyrillic alphabet. Iranian Persian, or Farsi, as it is natively called, is the lingua franca of Iran and is written with the Persian alphabet, a modified Arabic script. Although the spoken versions of Tajik and Farsi are mutually intelligible to educated speakers of both languages, the difference between the writing systems constitutes a barrier to text compatibility between the two languages. This paper presents a system to transliterate text between these two different Persian dialects that use incompatible writing systems. The system also serves as a mechanism to facilitate sharing of computational linguistic resources between the two languages. This is relevant because of the disparity in resources for Tajik versus Farsi.

Keywords: Persian, Tajik, Farsi, transliteration, machine transliteration, machine translation

1. Introduction

Transliteration is the practice of converting a text from one writing system, or script, into that of another. It is usually employed as a means to represent foreign words or phrases that use a different writing system in the context of a native language. Here we present a system that uses statistical machine translation (SMT) techniques to transliterate between two dialects of Persian that use different writing systems.

As part of the Soviet "Russification" of Central Asia, the Cyrillic script was introduced to Tajikistan in the late 1930s. The usage of a modified Cyrillic alphabet for Tajik Persian displaced the existing Persian alphabet in that country. This created a barrier to written communication between Tajikistan and its Persian speaking neighbors, like Iran. This paper presents a system for transliterating between the two writing systems that is based upon SMT strategies.

It is conventional among native English speakers to refer to modern Iranian Persian as "Farsi". This term is actually the native word for the Persian language in Iranian Persian, akin to *Español* for Spanish or *Français* for French — though it is more common for native Iranian Persian speakers to refer to their language as "Persian" when speaking English. This is due to a variety of socio-political reasons that are beyond the scope of this paper. Since both languages are technically Persian, to avoid ambiguity, in this paper we shall refer to Iranian Persian and Tajik Persian simply as *Farsi* and *Tajik*, respectively. We also reference the languages using the ISO 639-1 standard for two letter language abbreviations.[1] This is worth noting since the standard abbreviation for Tajik (tg) may not be obvious to those unfamiliar with it since it contains a 'g' instead of a 'j'.

Among the challenges for such a transliteration system are: (1) the Cyrillic alphabet is written left-to-right with discrete letters while the traditional Persian alphabet is written right-to-left in a connected, cursive style, with the form of each letter dependent upon its position in a word and adjacent letters, (2) the Cyrillic alphabet has a full complement of vowels which cover the range of Tajik phonology, while the Persian alphabet is consonantal, like Arabic, leaving many vowels unwritten and inferred from context, (3) some morphological affixes in Tajik are represented as separate lexical units in Farsi, and (4) a single phoneme in Farsi may have up to four different letters that may represent the sound. Conversely, because the syntax and grammar of the two languages are similar, some issues are more trivial. Word order is identical for practical purposes, therefore the need for SMT distortion models is largely moot.

It is possible to simply consider this task as a traditional machine translation problem and construct a system using existing SMT system by using a sufficiently large Tajik-Farsi sentence-aligned bilingual corpus as training data. Unfortunately, parallel corpora for Tajik-Farsi are both rare and sparse. We propose to leverage the high degree of parallelism between the two languages to construct a system that instead uses transliteration based upon "translation" of the lexical representations of morphemes and phonemes.

This paper also makes use of the UniPers[2] writing system for the convenience of demonstrating Persian examples without requiring the knowledge of either the Farsi script or Cyrillic alphabet. It uses Latin-based characters and is derived from the phonology of Farsi. The most notable difference between Farsi and Tajik phonology is that Tajik retains the distinction between the voiced and unvoiced phonemes for the letters қ /q/ and ғ /ʁ/ that correspond to the Farsi letters ق and غ, respectively. In Farsi, both letters are voiced and pronounced similarly in most dialects. We

[1] www.iso.org

[2] www.unipers.com

note this here since we augment UniPers with the grapheme *ğ* to represent غ when it is necessary to disambiguate it from ق, see Table 1.

UniPers	IPA	tg	IPA	fa
q	/q/	Қ қ	/ɢ/, /ɣ/	ق
ğ	/ʁ/	Ғ ғ	/ɣ/	غ

Table 1. UniPers augmentation for *q* and *ğ*.

2. Previous Work

In this section we briefly review previous work in Statistical Machine Translation, on which our approach is based, Machine Transliteration, and finally Machine Transliteration of Persian, specifically.

2.1 Statistical Machine Translation

Koehn et al. (2003) described a phrase-based SMT system based on the noisy channel model. In phrase-based translation, the goal is to reduce the restrictions of word-based translation by translating sequences of words, or *phrases*. In this context, "phrase-based" does not imply a linguistic phrase, but an *n*-gram phrase, a statistically relevant sequence of word tokens. Phrase-based SMT comprises three primary sub-systems: translation model, language model, and decoder (Figure 1).

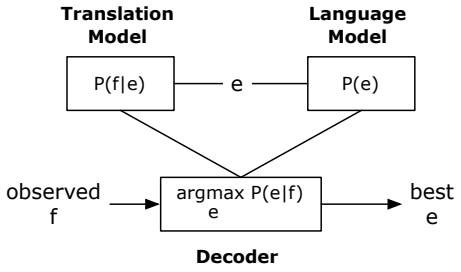

Figure 1. Noisy channel model.

By convention in SMT, the source language is denoted *f* and the target language *e*. Given an observed *f*, the noisy channel model uses a Bayesian approach to determine the most likely *e* from an *n*-best list.

$$\underset{e}{\operatorname{argmax}} P(e|f) = \underset{e}{\operatorname{argmax}} P(f|e)\ P(e)$$

Translation Model. Most contemporary SMT systems use alignment tools based upon IBM models (Brown et al., 1994; Och & Ney, 2000). These models are used to create phrase-based lookup tables that represent the probability *P*(*f*|*e*).

Language Model. A statistical Language Model assigns a probability to a sequence of *n* words P(w_1, w_2, … , w_n). It captures the how likely a sequence of words may occur in a modeled language. This is done by calculating the probability of the occurrence of an *n*-gram sequence of *n* words in a monolingual training corpus (Clarkson & Rosenfeld, 1997; Stolcke, 2002; Stolcke et al., 2011).

Decoder. Translations are executed on source text using a beam search decoder that uses the lookup tables from the Language Model to create an *n*-best list of possible translations. This *n*-best list is then re-ranked based upon how closely each candidate translation corresponds to the Language Model (Koehn 2004; Olteanu et al., 2006; Koehn et al., 2007).

2.2 Transliteration

Knight & Graehl (1997) and Knight & Graehl (1998) describe a multi-stage generative model for machine transliteration between Japanese and English. They also propose applicability of their model to Arabic and English. Some of these ideas were used by Al-Onaizan & Knight (2002) in their transliteration of Arabic names into English. Later, Haizhou et al. (2004) defined a joint source-channel model for machine transliteration.

More recently, work has been done on Machine Transliteration of Japanese-English that uses phrase-based SMT techniques (Finch & Sumita, 2007). Rama & Gali (2009) performed similar research for the Hindi-English language pair. Finch & Sumita (2010) extended their earlier work using the joint multigram model (Deligne & Bimbot, 1995) to generate the *n*-best list of transliteration hypotheses. Finch & Sumita (2010) also demonstrated language independence of their approach, achieving similar results across eight different language pairs.

Persian Transliteration. Karimi, et al. (2006) analyze grapheme-based transliteration methods on English to Persian, then introduce a new model of Persian that takes into account the practice of shortening, or even omitting, sequences of English vowels. Later, Karimi et al. (2007) proposed an algorithm for English to Persian transliteration that employed a back-transliteration method. That is, the recovery of the original source word from a reverse transliteration. The result is compared to the original word and evaluated for similarity.

3. Problem Description

Because the two languages appear so similar on the surface, it is tempting to assume the problem to be more trivial, surrendering to a simple letter substitution solution. However, several issues exist.

Farsi script letter ambiguity. Farsi uses several different groups of letters that each represent a single sound and single Tajik letter. This is an artifact of the Arabic writing system on which it is based. The Arabic alphabet contains letters for sounds that do not occur in Farsi. Although these letters have persisted in Arabic loan words, they are pronounced using Persian phonology. This causes problems of letter alignment ambiguity where a single Tajik letter may have up to four equivalent Farsi letters, Table 2.

UniPers	Tajik	Farsi
z	з	ز
		ذ
		ظ
		ض
s	с	س
		ص
		ث
t	т	ت
		ط
h	ҳ	ه
		ح

Table 2. Ambiguous Farsi Consonants.

Non-bijective alignment. Several syllables in Persian that have a single spelling in Tajik have multiple renderings in Farsi, depending upon the word. This is not just to do with the ambiguity of Farsi script letters mentioned above. There may also be silent, unpronounced letters in Farsi that are not present in Tajik. For example, the Tajik "хо-" / *xâ* / can be spelled both "خوا-" and "خا-" in Farsi, depending on the word – the Tajik words хостан (want) and хомӯш (silent) are spelled in Farsi "خواستن" and "خاموش", respectively.

Like the silent letter و above, other Farsi letters may not only be unpronounced, but used imply an unwritten vowel. For example, the common Farsi suffix, letter *heh* (ه), indicates that a word ends with the vowel phonome –*e*. This is represented by the Tajik suffix –а, effectively resulting in the mapping of a vowel to a consonant. For example, in transliterating the Tajik word "хафта" (week) to Farsi, the first "а" is omitted as an unwritten Farsi short vowel, and the second "а" is rendered as the Farsi consonant ه, resulting in the Farsi word هفته (*hafte*).

Ezâfe. The *ezâfe* is an enclitic phoneme that is added to the end of a noun to indicate that it is modified by (1) another noun, (2) an adjective, or (3) a pronoun, which follow the noun. The *ezâfe* is pronounced /-*e*/ after nouns ending in a consonant and /-*ye*/ after nouns ending in a vowel sound. This spoken phoneme is usually unwritten in Farsi, but always written in Tajik, spelled with the suffix -и. Thus, when transliterating from Tajik to Farsi, it is necessary to map the Tajik suffix -и to the null character in Farsi, and to detect when to infer the Tajik suffix -и when transliterating from Farsi to Tajik.

Direct object marker. Definite direct objects are marked with the postposition marker /*râ*/. In Farsi, this is a separate token "را" that follows the direct object after a space. In Tajik, it is connected to the noun with no space as the suffix "-ро". Thus the sentence "I read the book" would be rendered as in Table 3.

	Native Text	*UniPers*
fa	من کتاب را خواندم	*man ketâb râ xândam*
tg	ман китобро хондам	*man ketâbrâ xândam*

Table 3. Direct object postposition marker.

Case Sensitivity. The Cyrillic of Tajik utilizes case sensitivity similar to that of Russian, on which writing system it is based. Sentences begin with capital letters, as do proper nouns. By contrast, Farsi letters have no case. This presents no problem for tg-fa transliteration, but fa-tg transliteration needs to infer "lost" or hidden information, i.e., which Tajik words should be capitalized.

4. System Description

Our transliteration system is based on the concept of phrase-based statistical machine translation, described in Section 2. Instead of translating sentences using lexical *n*-grams based upon sequences of words, we translate at the word-level using grapheme-based *n*-grams.

In our system, instead of a bilingual, sentence-aligned training corpus for the Translation Model, we use training data based on aligned words, sub-word strings, and individual graphemes. Language Modeling and Decoding are likewise performed at the grapheme level.

4.1. Translation Model

We manually created a training corpus that consisted of 3503 tg-fa word pairs selected based upon their frequency in both Farsi and Tajik monolingual text. Additionally, we added tg-fa aligned consonant pairs that could be paired unambiguously (Table 4). These individual grapheme alignments facilitate the transliteration of words that are not present in the training set.

Like Karimi et al. (2007), our system avoids the observed tendency of other transliteration approaches to align consonants to consonants, and vowels to vowels, as a substitute for phonological alignment. This is because we rely on statistical alignments which are blind to alignments that may otherwise seem intuitive to human judges.

We used GIZA++ for alignment training (Och & Ney, 2000). GIZA++ assumes a whitespace as the default delimiter between *n*-gram elements during training. Therefore, we split words in the training corpus based on letter boundaries. In order to capture dependencies that rely on a character's position in a word, we also add unique initial and terminal character tokens to the beginning and end of each word prior to alignment training.

IPA	*UniPers*	Tajik	Farsi
/b/	*b*	Б б	ب
/g/	*g*	Г г	گ
/ʁ/	*gh*	Ғ ғ	غ
/d/	*d*	Д д	د
/ʒ/	*ž*	Ж ж	ژ
/k/	*k*	К к	ک
/q/	*q*	Қ қ	ق
/l/	*l*	Л л	ل
/m/	*m*	М м	م
/n/	*n*	Н н	ن
/p/	*p*	П п	پ
/ɾ/	*r*	Р р	پ
/f/	*f*	Ф ф	ف
/χ/	*x*	Х х	خ
/tʃ/	*c*	Ч ч	چ
/dʒ/	*j*	Ҷ ҷ	چ
/ʃ/	*š*	Ш ш	ش

Table 4. Unambiguous consonants alignment.

Preprocessor. As noted above, some transliterations depend on an adjacent word to disambiguate. For example, recognizing when to insert a written *ezâfe* or to concatenate the postposition marker *râ* when transliterating from Farsi to Tajik. Since the transliteration engine processes a single word at a time, we therefore deal with these cases in a preprocessing stage. When transliterating from Farsi to Tajik, all instances of *râ* are first joined to their referring noun. When transliterating from Tajik to Farsi, multi-token words like "ин ҷо" (here) and "он ҷо" (there) are likewise concatenated.

The resulting alignment model contained entries like the one graphically represented in Figure 2.

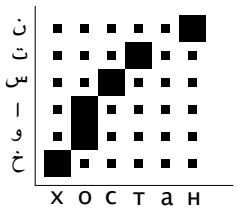

Figure 2. Example alignment for خواستن–хостан (want).

Here, one can observe both the tg-fa 1-2 alignment:

о = وا

and the null alignment:

а = ∅

4.2. Language Model

Like the Translation Model, The Language Model is also based on *n*-grams of letters, not words.

Corpora. For Language Modeling we used two different corpora, a native Farsi corpus and a native Tajik corpus, both taken from online news sources. The native Farsi corpus is the Bijankhan Corpus (Amiri et al., 2007), an existing native Farsi corpus[3]. The Bijankhan Corpus is also manually POS-tagged by native Farsi speakers. We built the native Tajik corpus from the Tajikistan-based online news site, Asia-Plus.[4] We call the Tajik Asia-Plus corpus the "TAP Corpus". The Bijankhan Corpus contains 133,614 sentences with an average sentence length of 23.97 words. The TAP Corpus contains 13,820 sentences with an average sentence length of 28.61 words (Table 5).

Language	Corpus	Sentences	Ave. Length
fa	Bijankhan	133,614	23.97
tg	TAP	13,820	28.61

Table 5. Percentage of known and unknown words.

4.3. Decoder

For the decoding stage, we used the beam search decoder Phramer (Olteanu et al., 2006). Input text was first split into words, placing one word each per line. Sentence terminal tokens were placed on a line by themselves so that sentences could be reconstructed after transliteration. Like both the Translation Model and Language Model training, each word was split using a space (U+0020) between each letter. Thus, the decoder effectively treated each word as a discrete sentence for translation.

Unlike word level SMT (at the) , There should be "no out of vocabulary" words that aren't present in the training set. The "vocabulary" of each language is the known closed set of the alphabet. In rare cases foreign words are rendered in Latin-based characters, even in Farsi. In these cases, the transliteration will pass the words unchanged in their original character set.

Unlike SMT of full sentences, we need not be concerned with a Distortion Model in the transliteration decoding stage. That is, the reordering of elements when mapping one language to another. Re-ordering of words when translating between between languages is frequently necessary. For example, English adjectives precede the noun they modify, whereas Farsi and Tajik adjectives follow their associated noun. The mapping order of both lexemes and graphemes between tg-fa is monotonic (Figure 3).

[3] http://ece.ut.ac.ir/dbrg/bijankhan

[4] http://news.tj/tj

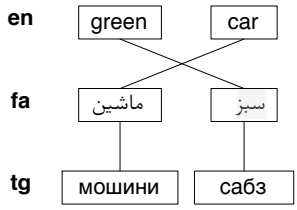

Figure 3. Distortion Model of "green car" between en-fa and fa-tg.

In phrase-based machine translation, candidate translations that consist of longer *n*-grams will usually be preferred, having a higher probability than the product of their constituent elements. This allows for the disambiguation of phonemes that may have a single spelling in one language, but multiple spellings in the other.

Table 6 shows an example of *n*-gram alignments for the constituents of two words that both contain the Tajik string "хо", namely хостан (sleep) and хоридан (itch). In Farsi, this phoneme is spelled differently in the two words. Although the Tajik bigram "хо" has ambiguous mapping to Farsi, the longer Tajik *n*-grams "хоби" and "хори" map unambiguously to Farsi *n*-grams.

English	*UniPers*	Tajik	Farsi
sleep	*xâbidan*	хобидан	خوابیدن
	xâ	хо	خوا
	bi	би	بی
	xâbi	хоби	خوابی
	dan	даи	دن
itch	*xâridan*	хоридан	خاریدن
	xâ	хо	خا
	ri	ри	ری
	xâri	хори	خاری
	dan	даи	دن

Table 6. Disambiguation of Tajik *n*-gram "хо" mapping to Farsi.

We may also be able to resolve an ambiguously mapped *n*-gram based on its position in a word. Table 7 illustrates this by indicating whether an *n*-gram is an initial or terminal string in a word by using a dash after or before the n-gram, respectively. This is accomplished in the Translation Model by placing a designated token at the beginning and end of each word prior to training, as mentioned in Section 4.1. In these examples, the ambiguity is on the Tajik side of the alignment.

In each of the three cases there is an ambiguous mapping of the Farsi letter "ی" to either the Tajik letter "и" or "ӣ". In these *n*-grams we identify a preference for aligning Farsi letter "ی" to Tajik "и" when it occurs as the second letter in a word following a consonant and to Tajik letter "ӣ" when it is in the terminal position in a word.

English	*UniPers*	Tajik	Farsi
grandmother	*bibi*	бибӣ	بیبی
	bi–	би–	بی–
	–bi	–бӣ	–بی
nose	*bini*	бинӣ	بینی
	bi–	би–	بی–
	–ni	–нӣ	–نی
need	*niâz*	ниёз	نیاز
	ni–	нӣ–	نی–
	–âz	–ёз	–از

Table 7. Disambiguation of Farsi *n*-grams mapping to Tajik based upon position.

4.4 System Transliteration Examples

Using techniques described above, we note several successful examples from our actual system output that would have been problematic with a naïve grapheme mapping strategy. These are shown in Table 8.

The Farsi word ساعت (hour) contains an explicit glottal stop ع. This is normally indicated by the letter ъ in Tajik, which would have otherwise resulted in the *incorrect* transliteration as соъат instead of соат. Our system correctly recognized the null alignment.

As noted in Table 2, the Tajik letter "з" may be mapped to any one of four different possible Farsi letters. The transliteration of ғизо (food) recognized not only the correct mapping of this letter, but also the more rare null alignment of the Tajik letter "и".

The Farsi letter "ه" is normally pronounced as *h*, however it is usually silent in the terminal position, where it indicates an implied short vowel, effectively *-ah* or *-eh*. This consonant to vowel mapping was correctly identified in the word هفته (week).

Finally, the unwritten vowel in the middle of the Farsi bigram "دل" was accurately found as "дил" and "дал" respectively in the Tajik words дил (heart) and далел (reason).

English	*UniPers*	Tajik	Farsi
hour	*sâ'at*	соат	ساعت
food	*ghazâ*	ғизо	غذا
week	*hafte*	ҳафта	هفته
heart	*dal*	дил	دل
reason	*dalil*	далел	دلیل

Table 8. Successful transliteration examples of interest.

5. Applications and Evaluation

One of the primary goals for this work is to provide Tajik access to the relatively larger set of computational linguistic resources available for Farsi. To assess the applicability of our transliteration system for the mapping of resources from Farsi to Tajik, we devised and evaluated two tasks: part of speech tagging and machine translation.

5.1 Part-of-speech tagging

For this task, we trained a Tajik POS tagger using the Farsi POS-tagged Bijankhan Corpus that we transliterated into Tajik. The training was done using a Maximum Entropy model tagger (Ratnaparkhi, 1996). The tag set we used for Tajik was a modified version of one used by the Bijankhan Corpus based on that of Aleahmad et al. (2006). The Bijankhan Corpus was built for other purposes and has very fine grained tags which are not suitable for POS tagging experiments. Bijankhan Corpus originally employed 550 different tags that were defined in a hierarchical tree structure. This number was reduced to 40 by Raja et al. (2007) using techniques from Oroumchian (2006). This still contained a high degree of granularity. More specifically, it defines six different types each of Verb, Adjective, and Adverb. For example, it further distinguishes comparatives and superlatives from other adjectives. For our purposes, we retained the 40-tag level of detail from Raja et al. (2007) to determine if a tag was correct or not, though we collapsed sub-tags into their parent category for evaluation. Thus, a comparative adjective (ADJ_CMPR) mis-tagged as a superlative adjective (ADJ_SUP) would be considered as an error, even though they are both classified as adjective. The eight top-level POS categories and their frequency in the training corpus are shown in Table 9.

POS Category	Frequency
Noun	43.42%
Preposition	12.31%
Adjective	10.77%
Verb	8.69%
Conjunction	8.09%
Pronoun	2.38%
Article	1.77%
Adverb	1.49%
Other	11.08%

Table 9. Training Corpus POS Frequencies.

For testing, we chose 1023 sentences from the TAP Corpus that we manually tagged. We then performed automatic tagging on the same sentences using our Maximum Entropy model tagger trained on the Tajik transliterated Bijankhan Corpus and compared the results against manual tagging for accuracy.

For reference, we also include the accuracy of various Farsi POS tagging approaches performed on the Bijankhan Corpus by Raja et al. (2007). Table 10 shows the percentage of both known words and unknown words in the respective corpora. Table 11 shows the accuracy of each system on its respective test corpus. Raja et al. (2007) report on three different systems trained and tested on the Bijankhan corpus all in Farsi: a Markov model tagger, a memory-based tagger (MBT), and a Maximum Likelihood Estimation tagger (MLE).

POS Tagger	Corpus	Known Words	Unknown Words
Our System	TAP	92.27%	7.73%
Raja et al (2007)	Bijankhan	97.96%	2.04%

Table 10. Known and unknown words in the test sets.

	Known words	Unknown words	Overall
Our System	94.98%	63.22%	92.52%
Markov	97.01%	77.77%	96.64%
MBT	96.86%	75.15%	96.42%
MLE	96.60%	15.00%	94.63%

Table 11. POS tagging accuracy of both in and out of vocabulary words.

5.2 Machine Translation

In this task we evaluate the applicability of using our transliteration system for Tajik machine translation. To do this, we first transliterate Tajik into Farsi and then subsequently translate the result into English using a commercial machine translation system.

We begin with the same 1023 Tajik sentences taken from the TAP corpus used for POS tag testing above. We automatically transliterated these sentences into Farsi text with our system, then translated them into English using Google Translate.[5]

For evaluation, we use BLEU method (Papineni et al., 2002), a standard for assessment of machine translation. We then compare our tg-en BLEU score with that of various fa-en machine translation systems, including that of Google Translate.

Mohaghegh (2011) described multiple related systems for Translating Farsi to English under the name PeEn-SMT. Of the various configurations used, their System 5 performed best. Within their System 5, they report results using different language models.

Table 12 shows the performance of our tg-en SMT task compared to that of the fa-en BLEU scores of PeEn-SMT System 5 and Google Translate.

[5] http://translate.google.com

MT System	BLEU
PeEn-SMT System 5 – Hamshari LM	0.2127
tg-fa → fa-en Google	***0.2349***
fa-en Google Translate	0.2611
PeEn-SMT System 5 – BBC News LM	0.2865
PeEn-SMT System 5 – IRNA LM	0.3496

Table 12. tg-en SMT via tg-fa transliteration.

We are especially interested in the performance of Google Translate, since we have a dependency on it for our fa-en stage. From this, we infer some cautious judgements about the accuracy of our tg-fa transliteration for the application of tg-en machine translation. Comparing our tg-en translation system with the Google Translate fa-en on which it depends, we report a performance that is 89.96% within the BLEU score of Google, i.e. ~90%. From this we approximate our tg-fa transliteration accuracy for the purposes of SMT at ~90%, attributing the ~10% lower score from Google's fa-en SMT to errors of tg-fa transliteration.

We also note that our system that uses an additional intermediate language performed better than the PeEn-SMT System 5 using the Hamshari language model, even though only a single source and target language were involved. From this we infer an almost 90% accuracy of transliteration from Tajik to Farsi for the machine translation task.

Lastly, we provide an example of one of the 1023 sentences from the TAP corpus used in the machine translation task in Figure 4.

Language	Sentence
tg	Имрӯз нозирони байналмиллалӣ конфронси матбуотӣ дар шаҳри Душанбе баргузор менамоянд
fa	امروز ناظرانی بین المللی کانفرانسی مطبوعاتی در شهر دوشنبه برگزار می نمایند
en	Today international observers should hold a press conference in Dushanbe

Figure 4. Example tg-en Sentence Translation via Farsi

6. Conclusion and Future Work

In this paper we presented a system to transliterate text between Tajik and Farsi. We then showed the applicability of Farsi computational resources to that of Tajik NLP tasks via Tajik-Farsi transliteration by executing two different tasks, part-of-speech tagging and machine translation. In both tasks, the results were well within accuracy ranges to render Tajik-Farsi transliteration a viable mechanism to enable these tasks. This opens the door for application of the disproportionately larger set of NLP tools for Farsi to computational language tasks in Tajik. Additionally, tools for Farsi that have been adapted to work with Tajik may also be used to create new Tajik lexical resources such as WordNets, FrameNets, and corpora.

For both tasks, we note the small size of the test set, only 1023 sentences. This carries the risk that the performance of a few sentences could non-trivially affect the overall outcome. Our results have encouraged us to continue work on this project, creating much larger test sets from the TAP corpus for evaluation.

During the assessment of performance in the two tasks, we determined two primary root causes of errors. The first is the presence of both uniquely Tajik words and Russian loan words in the Tajik corpus that don't exist in Farsi. The other is the occurrence of named entities Tajik that are uncommon in Farsi. These include geographic names, persons, and institutions that are socio-culturally Tajik.

Future Work. Based on the positive outcome of these efforts, we are working to improve both our existing resources and to develop new ones.

To increase the size of fa-tg bilingual word list used in Translation Model training, we propose to develop heuristics to correlate and align Farsi-English and Tajik-English digital dictionaries. Based on the size of available bilingual dictionaries, we estimate an increase of the size of our current fa-tg bilingual word list by six-fold.

We are developing annotation tools that are optimize for the Perso-arabic and Cyrillic character sets in order to facilitate the manual annotation of the entire TAP corpus. We plan to make this corpus available to others for research when completed.

Finally, we are working on the use of our transliteration system to bootstrap a Tajik WordNet by mapping from a n existing Farsi WordNet.

7. References

Aleahmad, A., Ramezani, Y., Oroumchian, F. (2006). *Using OWA for Persian Part of Speech Tagging*.

Al-Onaizan, Y. & Knight, K. (2002). Machine Transliteration Of Names In Arabic Text. In *Proceedings of the ACL-02 Workshop On Computational Approaches To Semitic Languages*.

Amiri, H., Hojjat, H., Oroumchian, F. (2007). Investigation On A Feasible Corpus For Persian POS Tagging. *12th International CSI Computer Conference*, Iran.

Brown, P. F., Della Pietra, S., Della Pietra, V. J., and Mercer, R. L. (1994). The Mathematics of Statistical Machine Translation: Parameter Estimation", *Computational Linguistics*, vol. 19, 1994, pp. 263-311.

Clarkson, P. & Rosenfeld, R. (1997). Statistical Language Modeling Using The CMU-Cambridge Toolkit. In *Fifth European Conference On Speech Communication And Technology*.

Deligne, S. & Bimbot, F. (1995). Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams. In *Acoustics, speech, and signal processing*, 1995. ICASSP-95. vol. 1, pp. 169-172

Finch, A. and Sumita, E. (2007). Phrase-based Machine Transliteration, *Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, pp. 13-18.

Finch, A. and Sumita, E. (2010). Transliteration Using A Phrase-based Statistical Machine Translation System To Re-score The Output Of A Joint Multigram Model, *Proceedings of the 2010 Named Entities Workshop*, pp. 48-52

Haizhou, L., Min, Z., & Jian, S. (2004). A Joint Source-channel Model For Machine Transliteration. In *Proceedings of the 42nd Annual Meeting On Association For Computational Linguistics*.

Karimi, S., Turpin, A., & Scholer, F. (2006). English To Persian Transliteration. In *String Processing And Information Retrieval*.

Karimi, S., Scholer, F., & Turpin, A. (2007). Collapsed Consonant And Vowel Models: New Approaches For English-persian Transliteration And Back-transliteration. In *Annual Meeting-Association For Computational Linguistics*.

Knight, K. & Graehl, J. (1997). Machine Transliteration. In *Proceedings Of The Eighth Conference On European Chapter Of The Association For Computational Linguistics*.

Knight, K. & Graehl, J. (1998). Machine transliteration. *Computational Linguistics*, 24(4), 599-612

Koehn, P., Och, F., & Marcu, D. (2003). Statistical Phrase-based Translation. In *NAACL/HLT 2003, Proceedings Of The Human Language Technology And North American Association For Computational Linguistics Conference*. Edmonton, Canada.

Koehn, P. (2004). Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., and Zens, R. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*.

Mohaghegh, M., Sarrafzadeh, A., & Moir, T. (2011). Improving Persian-english Statistical Machine Translation: Experiments In Domain Adaptation. In *Proceedings Of The 2nd Workshop On South And Southeast Asian Natural Language Processing (WSSANLP 2011)*.

Och, F. J. and Ney, H. (2000). Improved Statistical Alignment Models, In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 440-447.

Olteanu, M., Davis, C. I., Volosen, I., Moldovan, D. (2006) Phramer - An Open Source Statistical Phrase-based Translator. In *Proceedings of HLT-NAACL Workshop on Statistical Machine Translation*. pp. 146–149.

Oroumchian, F., Tasharofi, S., Amiri, H., Hojjat, H., & Raja, F. (2006). *Creating A Feasible Corpus For Persian POS Tagging*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A Method For Automatic Evaluation Of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics.*

Raja, F., Amiri, H., Tasharofi, S., Sarmadi, M., Hojjat, H., & Oroumchian, F. (2007). Evaluation Of Part Of Speech Tagging On Persian Text. *University of Wollongong in Dubai-Papers*, 8.

Rama, T. and Gali, K. (2009). Modeling Machine Transliteration As A Phrase Based Statistical Machine Translation Problem, *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, p. 124–127.

Ratnaparkhi, A. (1996). A Maximum Entropy Part-Of Speech Tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, May 17-18, 1996. University of Pennsylvania.

Schmid, H. (1999). Improvements In Part-of-speech Tagging With An Application To German. *Natural Language Processing Using Very Large Corpora*, 11, pp. 13-26.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Seventh international conference on spoken language processing*.

Stolcke, A., Zheng, J., Wang, W., & Abrash, V. (2011). SRILM at sixteen: Update and outlook. In *Proceedings of IEEE automatic speech recognition and understanding workshop (ASRU)*.